Informatische Werkzeuge in den Geistes- und Sozialwissenschaften 1/2

Prof. Dr. Michael Kohlhase

Professur für Wissensrepräsentation und -verarbeitung
Informatik, FAU Erlangen-Nürnberg
`Michael.Kohlhase@FAU.de`

2025-02-06

Contents

0.1 Preface

0.1.1 Course Concept

Objective: The course aims at giving students an overview over the variety of digital tools and methods at the disposal of practitioners of the humanities and social sciences, explaining their intuitions on how/why they work (the way they do). The main goal of the course is to empower students for their for the emerging discipline of "digital humanities and social sciences". In contrast to a classical course in computer science which lays the mathematical and computational foundations which will become useful in the long run, we want to introduce methods and tools that can become *useful in the short term* and thus generate immediate success and gratification, thus alleviating the "programming shock" (the brain stops working when in contact with computer science tools or computer scientists) common in the humanities and social sciences.
Original Context: The course "Informatische Werkzeuge in den Geistes- und Sozialwissenschaften" is a first-year, two-semester course in the bachelor program "Digitale Geistes- und Sozialwissenschaften" (Digital Humanities and Social Sciences: DigiHumS) at FAU Erlangen-Nürnberg.
Open to External Students: Other bachelor programs are increasingly co-opting the course as specialization option or a key skill. There is no inherent restriction to DHSS students in this course.
Prerequisites: There are no formal prerequisites – after all it starts in the first semester for DigiHumS – but a good deal of motivation, openness towards exploring the weird and wonderful world of digital methods and tools, and a certain perseverance in the face of not understanding directly help tremendously and helps having fun in this course.

We do assume that students have a personal laptop, or access to a computer where they have admin rights, i.e. can install software. This is necessary for solving the homework. In particular, smartphones and most tablet computers will not suffice.

0.1.2 Course Contents

The course comprises two parts that are given as two-hour/week lectures.
IWGS 1 (the first semester): begins with an introduction to programming in Python which we will use as the main computational tool in the course; see **??** and **??**. In particular we will cover

- systematics and culture of programming
- program and control structures
- basic data structures like numbers and strings, in particular character encodings, Unicode, and regular expressions.

Building on this, we will cover

1. digital documents and document processing, in particular; text files, markupsystems, HTML, and XML; see **??**.
2. basic concepts of the World Wide Web; see **??**
3. Web technologies for interactive documents and their applications; in particular internet infrastructure, web browsers and servers, PHP, dynamic HTML, JavaScript, and CSS; see **??**.

IWGS 2 (the second semester): covers selected topics and exemplary tools that will become useful in the DH. We will cover

1. Databases; in particular entity relationship diagrams, CRUD operations, and querying; see **??**.
2. Image processing tools, see **??**

3. Using the ontologies and the semantic web for Cultural Heritage; see **??**

4. The WissKI System: A Virtual Research Environment for Cultural Heritage; see **??**

5. Copyright and Data Privacy as legal foundations of DH tools; see **??**

Idea: The first semester lays the foundations by introducing programming in Python and work our way towards web applications, which form the base of most modern tools in the DH. In **??**, we pull all parts together to build a first, simple web application with persistent storage that manages a set of books.

After an excursion into project management systems, we introduce methods and tools for their management. Here, we extend our web application to deal with image fragments; actually building a simple replacement for a prominent DH web application.

Finally, after another excursion – this time into the legal foundations of intellectual property and data privacy the course culminates in an introduction of the WissKI system, a virtual research environment for documenting cultural heritage artifacts. Indeed the WissKI system combines all topics in the course so far.

0.1.3 Programming Exercises and JuptyterLab as a Web IDE

Programming Exercises: Most of the computer tools introduced in this course require programming e.g. for configuration, extension, or input preprocessing or work much better when the user understands the basic underlying concepts at the program level. Therefore we accompany the course with a set of (programming) exercises (given as homework to the IWGS students) that allow practicing that.

Web IDEs: In the IWGS course at FAU, which is adressed to students from the humanities and social sciences, we do not have access to a pool of standardized hardware. Students have to use their own computing devices for the programming exercises. In any group with diverse hardware, installing software, standardizing software versions, ...becomes a serious problem, even if the group only has 50 members; in IWGS, we need the Python interpreter, a text editor or integrated development environment (IDE), and various Python libraries. In IWGS we solve this by using a web IDE, which only presupposes a web browser on student hardware.

Jupyterlab: After experimenting with commercial web IDEs we settled on jupyterLab, even though it does not focus on IDE features. Jupyter notebooks allow to mix documentation, code snippets, and exercise text of programming exercises and package them into learning objects that can be downloaded, interacted with, and submitted easily. jupyterLab acts as the user interface for managing and editing jupyter notebooks and supplies standardized shell and Python REPLs for students. The jupyterLab server runs as a virtual machine on the instructor's hardware. Resource consumption is minimal in our experience (except in the week before the exam). See [JKI] for a documentation of how to set up a server for a small course like IWGS.

Limitations of JupyterLab: Of course, students who want to engage in more serious software development will eventually have to "graduate" to a regular IDE when programs become larger and more long-lived. But this – and the necessary software engineering skills – is emphatically not the focus of the IWGS course.

Exercise Notebooks: The exercise notebooks (in notebook format and PDF – unfortunately only in German) can be found at `https://kwarc.info/teaching/IWGS/NB`. They comprise

- outright programming exercises that introduce the Python language or allow to play with the respective concepts in Python

- code reading/debugging exercises where the character of Beatrice Beispiel almost solves interesting problems, and

- development steps towards larger applications, which often involve completing Python skeletons using the concepts taught in the lectures.

In all cases, the necessary increments to be supplied by the students are designed to not let the Python skills become a barrier, but give students the opportunity to develop the necessary programming skills in passing.

We have themed the exercises with DigiHumS topics to keep them interesting for our students.

0.1.4 This Document

Format: The document mixes the slides presented in class with comments of the instructor to give students a more complete background reference.

Caveat: This document is primarily made available for the students of the IWGS course only. After two iterations of this course it is reasonably feature-complete, but will evolve and be polished in coming academic years.

Licensing: This document is licensed under a Creative Commons license that requires attribution, allows commercial use, and allows derivative works as long as these are licensed under the same license. **Knowledge Representation Experiment:** This document is also an experiment in knowledge representation. Under the hood, it uses the sTeX package [Koh08; sTeX], a TeX/LaTeX extension for semantic markup, which allows to export the contents into active documents that adapt to the reader and can be instrumented with services based on the explicitly represented meaning of the documents.

Other Resources: The lecture notes will be complemented by a selection of problems (with and without solutions) that can be used for self-study; see `http://kwarc.info/teaching/IWGS`.

0.1.5 Acknowledgments

Materials: The materials in this course are partially based on various lectures the author has given at Jacobs University Bremen in the years 2010-2016, these in turn have been partially based on materials and courses by Dr. Heinrich Stamerjohanns, PD Dr. Florian Rabe, and Prof. Dr. Peter Baumann. **??** have been provided by Philipp Kurth and Dr. Frank Bauer.

All course materials have been restructured and semantically annotated in the sTeX format, so that we can base additional semantic services on them.

Teaching Assistants: The organization and material choice in the IWGS has significantly been influenced by Jonas Betzendahl and Philipp Kurth, who have been very active and dedicated teaching assistants and have given feedback on all aspects of the course. They have also provided almost all of the IWGS exercises – see **??**.

DigiHumS Administrators: Jacqueline Klusik-Eckert and Philipp Kurth who used to administrate the DigiHumS major at FAU together have been helpful in navigating the administrative waters of an unfamiliar faculty.

WissKI Specialists and Colleagues: **??** has profited from discussions with Peggy Große and Juliane Hamisch, then two WissKI specialists at FAU. My colleagues Prof. Peter Bell has provided the idea and data for the "Kirmes Pictures Project" that grounds some of the second semester.

JupyterLab: The JupyterLab server at `https://juptyter.kwarc.info` (see **??**) has been developed, operated, and maintained by Jonas Betzendahl. For details see [JKI].

IWGS Students: The following students have submitted corrections and suggestions to this and earlier versions of the notes: Paul Moritz Wegener, Michael Gräwe.

0.2 Recorded Syllabus

In this section, we record the progress in the form of a "recorded syllabus", i.e. a syllabus that is created after the fact rather than before. For the topics planned for this course, see **??**.

Syllabus – Winter 2024/25: The recorded syllabus for this semester is in the course page in the ALeA system at `https://courses.voll-ki.fau.de/course-home/iwgs-1`. The table of contents in the IWGS notes at `https://courses.voll-ki.fau.de` indicates the material covered to date in yellow.

The recorded syllabus of IWGS-2 can be found at `https://courses.voll-ki.fau.de/course-home/iwgs-2`

Chapter 1

Preliminaries

1.1 Administrativa

We will now go through the ground rules for the course. This is a kind of a social contract between the instructor and the students. Both have to keep their side of the deal to make learning as efficient and painless as possible.

Prerequisites

- ▷ **General Prerequisites:** Motivation, interest, curiosity, hard work.
  nothing else! We will teach you all you need to know
- ▷ You can do this course if you want! (we will help)

Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 1 2025-02-06

Now we come to a topic that is always interesting to the students: the grading scheme: The short story is that things are complicated. We have to strike a good balance between what is didactically useful and what is allowed by Bavarian law and the FAU rules.

Assessment, Grades

- ▷ **Grading Background/Theory:** Only modules are graded! (by the law)
  - ▷ Module "DH-Einführung" (DHE) $\widehat{=}$ courses IWGS1/2, DH-Einführung. (7.5 ECTS)
  - ▷ DHE module grade $\rightsquigarrow$ pass/fail determined by "portfolio" $\widehat{=}$ collection of contributions/assessments.
  - ▷ Module "DH-Einführung mit Übungen" (DHÜ) $\widehat{=}$ courses IWGS1/2, (10 ECTS)
  - ▷ DHÜ module grade $\rightsquigarrow$ 1-5 50% exam, 50% homework assignments, 10% bonus points from prepquizzes.
- ▷ **Assessment Practice:** The IWGS assessments in the "portfolio" consist of
  - ▷ weekly homework assignments, (practice IWGS concepts and tools)
  - ▷ 60 minutes exam directly after lectures end: $\sim$ Feb. 10. 2025.

▷ **Retake Exam:** 60 min exam at the end of the exam break. (∼ May 4. 2025)

▷ **To help you succeed:** We offer you

- ▷ **External motivation**: informal points for homeworks and a grade for exam, (even though only pass/fail relevant in the end)
- ▷ weekly online prepquizzes that help you prepare for the course. (check understanding/preparation)

Homework assignments, quizzes, and end-semester exam may seem like a lot of work – and indeed they are – but you will need practice (getting your hands dirty) to master the IWGS concepts. We will go into the details next.

Preparedness Quizzes

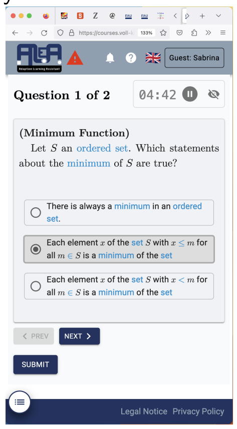
▷ **PrepQuizzes:** Before every lecture we offer a 10 min online quiz – the PrepQuiz – about the material from the previous week. (16:00-16:10; starts in week 2)

▷ **Motivations:** We do this to

- ▷ keep you prepared and working continuously. (primary)
- ▷ bonus points for the DHÜ students if you passed the exam with $\geq 50\%$ points. (potential part of your grade)
- ▷ update the ALeA learner model. (fringe benefit)

▷ The prepquiz will be given in the ALeA system

- ▷ `https://courses.voll-ki.fau.de/quiz-dash/iwgs-1`
- ▷ You have to be logged into ALeA! (via FAU IDM)
- ▷ You can take the prepquiz on your laptop or phone, . . .
- ▷ . . . in the lecture or at home . . .
- ▷ . . . via WLAN or 4G Network. (do not overload)
- ▷ Prepquizzes will only be available 16:00-16:10!



Question 1 of 2 04:42

(Minimum Function)
Let S an ordered set. Which statements about the minimum of S are true?

- ○ There is always a minimum in an ordered set.
- ◉ Each element x of the set S with $x \leq m$ for all $m \in S$ is a minimum of the set
- ○ Each element x of the set S with $x < m$ for all $m \in S$ is a minimum of the set

1.2 Getting Most out of IWGS

In this section we will discuss a couple of measures that students may want to consider to get most out of the IWGS course.

None of the things discussed in this section – homeworks, tutorials, study groups, and attendance – are mandatory (we cannot force you to do them; we offer them to you as learning opportunities), but most of them are very clearly correlated with success (i.e. passing the exam and getting a good grade), so taking advantage of them may be in your own interest.

IWGS Homework Assignments

- ▷ **Goal:** Homework assignments reinforce what was taught in lectures.
- ▷ **Homework Assignments:** Small individual problem/programming/proof task
  - ▷ but take time to solve (at least read them directly ⇝ questions)
- ▷ **Didactic Intuition:** Homework assignments give you material to test your understanding and show you how to apply it.
- ▷ ⚠ Homeworks give no points, but without trying you are unlikely to pass the exam.
- ▷ **Homework Workflow:** in ALeA (see below)
  - ▷ Homework assignments will be published on thursdays: see `https://courses.voll-ki.fau.de/hw/iwgs-1`
  - ▷ Go to the Tutorials to discuss them.
  - ▷ Submission of solutions via the StudOn system in the week after
  - ▷ graded by the TA.
- ▷ **Homework/Tutorial Discipline:**
  - ▷ Start early! (many assignments need more than one evening's work)
  - ▷ Don't start by sitting at a blank screen (talking & study groups help)
  - ▷ Humans will be trying to understand the text/code/math when grading it.
  - ▷ Go to the tutorials, discuss with your TA! (they are there for you!)

It is very well-established experience that without doing the homework assignments (or something similar) on your own, you will not master the concepts, you will not even be able to ask sensible questions, and take very little home from the course. Just sitting in the course and nodding is not enough! If you have questions please make sure you discuss them with the instructor, the teaching assistants, or your fellow students. There are three sensible venues for such discussions: online in the lectures, in the tutorials, which we discuss now, or in the course forum – see below. Finally, it is always a very good idea to form study groups with your friends.

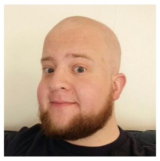
IWGS Tutorials

- ▷ Weekly tutorials and homework assignments (first one in week two)

▷ **Tutor:** (Doctoral Student in CS)

▹ Jonas Betzendahl: `jonas.betzendahl@fau.de`

They know what they are doing and really want to help you learn! (dedicated to DH)



▷ Jonas will also grade the homework assignments for the DFÜ students. (grade-relevant)

▷ **Goal 1:** Reinforce what was taught in class (important pillar of the IWGS concept)

▷ **Goal 2:** Let you experiment with Python (think of them as Programming Labs)

▷ **Life-saving Advice:** go to your tutorial, and prepare it by having looked at the lecture notes and the homework assignments

▷ **Inverted Classroom:** the latest craze in didactics (works well if done right)
in IWGS: lecture + homework assignments + tutorials $\widehat{=}$ inverted classroom

Collaboration

▷ **Definition 1.2.1.** Collaboration (or cooperation) is the process of groups of agents acting together for common, mutual benefit, as opposed to acting in competition for selfish benefit. In a collaboration, every agent contributes to the common goal and benefits from the contributions of others.

▷ In learning situations, the benefit is "better learning".

▷ **Observation:** In collaborative learning, the overall result can be significantly better than in competitive learning.

▷ **Good Practice:** Form study groups. (long- or short-term)

1. ⚠ those learners who work most, learn most!
2. ⚠ freeloaders – individuals who only watch – learn very little!

▷ It is OK to collaborate on homework assignments in IWGS! (no bonus points)

▷ Choose your study group well! (We will (eventually) help via ALeA)

As we said above, almost all of the components of the IWGS course are optional. That even applies to attendance. But make no mistake, attendance is important to most of you. Let me explain, . . .

Do I need to attend the IWGS Lectures

▷ Attendance is not mandatory for the IWGS course. (official version)

- **Note:** There are two ways of learning: (both are OK, your mileage may vary)
  - Approach B: Read a book/papers (here: lecture notes)
  - Approach I: come to the lectures, be involved, interrupt the instructor whenever you have a question.

  The only advantage of I over B is that books/papers do not answer questions
- Approach S: come to the lectures and sleep does not work!
- The closer you get to research, the more we need to discuss!

SOME RIGHTS RESERVED

1.3 Learning Resources for IWGS

But what if you are not in a lecture or tutorial and want to find out more about the IWGS topics?

Textbook, Handouts and Information, Forums, Videos

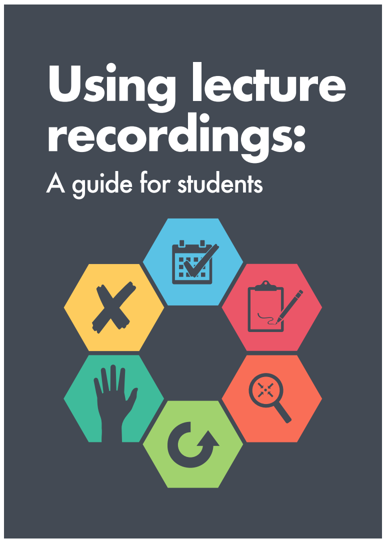
- **No Textbook:** but lots of online Python tutorials on the web.
- **Lecture notes:** will be posted at `https://kwarc.info/teaching/IWGS` (see references)
  - I mostly prepare/adapt/correct them as we go along
  - please e-mail me any errors/shortcomings you notice. (improve for the group)
- The lecture videos of are at `https://www.fau.tv/course/id/4020`
- **Do not let the videos mislead you:** Coming to class is highly correlated with passing the exam!
- Matrix chat at `#iwgs:fau.de` (via FAU IDM $\leadsto$ instructions)
- **StudOn Forum:** `https://www.studon.fau.de/crs5904688.html` for
  - announcements, homeworks (my view on the forum)
  - questions, discussion among your fellow students (your forum too, use it!)
- If you become an active discussion group, the course forum turns into a valuable resource!

SOME RIGHTS RESERVED

FAU has issued a very insightful guide on using lecture videos. It is a good idea to heed these recommendations, even if they seem annoying at first.

Practical recommendations on Lecture Videos

- **Excellent Guide:** [Nor+18a] (German version at [Nor+18b])



- Attend lectures.
- Take notes.
- Be specific.
- Catch up.
- Ask for help.
- Don't cut corners.

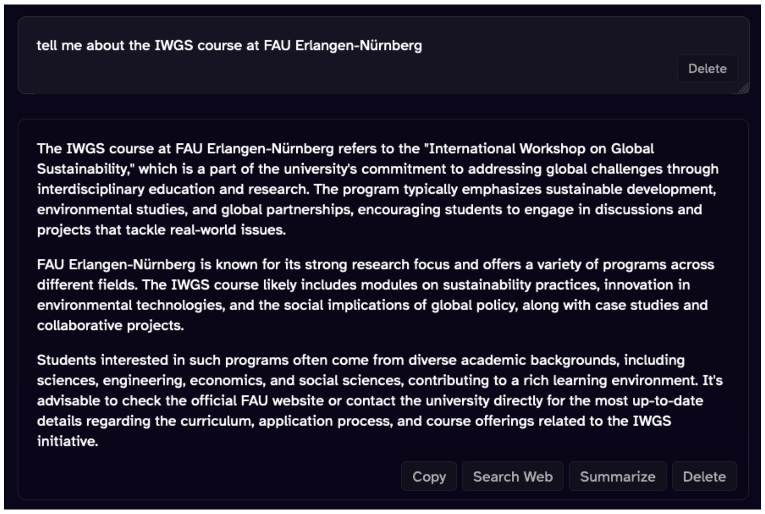
NOT a Resource for : LLMs – AI-tools like ChatGPT

▷ **Definition 1.3.1.** A large language model (LLM) is a computational model capable of language generation or other natural language processing tasks.

▷ **Example 1.3.2.** OpenAI's GPT, Google's Bard, and Meta's Llama.

▷ **Definition 1.3.3.** A chatbot is a software application or web interface that is designed to mimic human conversation through text or voice interactions. Modern chatbots are usually based on LLMs.

▷ **Example 1.3.4 (ChatGPT talks about IWGS).** (Aha, where does this come from?)



tell me about the IWGS course at FAU Erlangen-Nürnberg

The IWGS course at FAU Erlangen-Nürnberg refers to the "International Workshop on Global Sustainability," which is a part of the university's commitment to addressing global challenges through interdisciplinary education and research. The program typically emphasizes sustainable development, environmental studies, and global partnerships, encouraging students to engage in discussions and projects that tackle real-world issues.

FAU Erlangen-Nürnberg is known for its strong research focus and offers a variety of programs across different fields. The IWGS course likely includes modules on sustainability practices, innovation in environmental technologies, and the social implications of global policy, along with case studies and collaborative projects.

Students interested in such programs often come from diverse academic backgrounds, including sciences, engineering, economics, and social sciences, contributing to a rich learning environment. It's advisable to check the official FAU website or contact the university directly for the most up-to-date details regarding the curriculum, application process, and course offerings related to the IWGS initiative.

▷ **Example 1.3.5 (In the IWGS exam).** ChatGPT scores almost perfectly (master solution quality)

- ▹ ChatGPT can pass the exam . . . (We could award it a Master's degree)
- ▹ But can you? (the IWGS exams will be in person on paper)

You will only pass the exam, if you can do IWGS yourself!

▷ **Intuition:** AI tools like GhatGPT, CoPilot, etc. (see also [She24])

- ▹ can help you solve problems, (valuable tools in production situations)
- ▹ hinders learning if used for homeworks/quizzes, etc. (like driving instead of jogging)

▷ **What (not) to do:** (to get most of the brave new AI-supported world)

- ▹ try out these tools to get a first-hand intuition what they can/cannot do
- ▹ challenge yourself while learning so that you can also do it (mind over matter!)

1.4 Goals, Culture, & Outline of the Course

Goals of "IWGS"

▷ **Goal:** giving students an overview over the variety of digital tools and methods

▷ **Goal:** explaining their intuitions on how/why they work (the way they do).

▷ **Goal:** empower students for their for the emerging field "digital humanities and social sciences".

▷ **NON-Goal:** Laying the mathematical and computational foundations which will become useful in the long run.

▷ **Method:** Introduce methods/tools that can become *useful in the short term*

- ▹ generate immediate success and gratification, (important for motivation)
- ▹ alleviate the "programming shock" (the brain stops working when in contact with computer science tools or computer scientists) common in the humanities and social sciences.

One of the most important tasks in an inter/trans-disciplinary enterprise – and that what "digital humanities" is, fundamentally – is to understand the disciplinary language, intuitions and foundational assumptions of the respective other side. Assuming that most students are more versed in the "humanities and social sciences" side we want to try to give an overview of the "computer science culture".

Academic Culture in Computer Science

▷ **Definition 1.4.1.** The academic culture is the overall style of working, research, and discussion in an academic field.

- ▷ **Observation 1.4.2.** *There are significant differences in the academic culture between computer science, the humanities and the social sciences.*
- ▷ Computer science is an engineering discipline (we build things)
  - ▹ given a problem we look for a (mathematical) model, we can think with
  - ▹ once we have one, we try to re-express it with fewer "primitives" (concepts)
  - ▹ once we have, we generalize it (make it more widely applicable)
  - ▹ only then do we implement it in a program (ideally)

  Design of versatile, usable, and elegant tools is an important concern
- ▷ Almost all technical literature is in English. (technical vocabulary too)
- ▷ CSlings love shallow hierarchies. (no personality cult; alle per Du)

Please keep in mind that – self-awareness is always difficult – the list above may be incomplete and clouded by mirror-gazing. We now come to the concrete topics we want to cover in IWGS. The guiding intuition for the selection is to concentrate on techniques that may become useful in day-to-day DH work – not CS completeness or teaching efficiency.

Outline of IWGS 1:

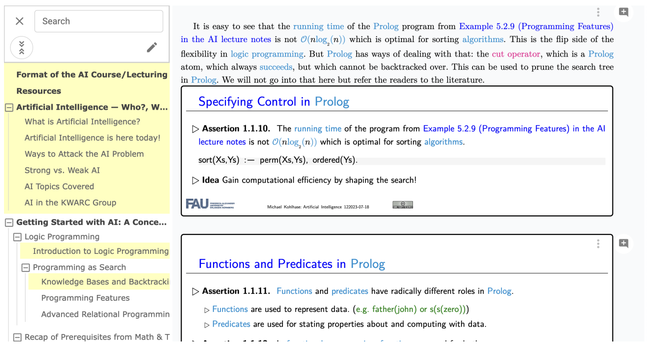
- ▷ Programming in Python: (main tool in IWGS)
  - ▹ Systematics and culture of programming
  - ▹ Program and control structures
  - ▹ Basic data structures like numbers and wordsstring, character encodings, unicode, and regular expressions
- ▷ Electronic documents and document processing:
  - ▹ text files
  - ▹ markup systems, HTML, and CSS
  - ▹ XML: Documents are trees.
- ▷ Web technologies for interactive documents and web applications
  - ▹ internet infrastructure: web browsers and server
  - ▹ serverside computation: bottle routing and
  - ▹ clientside interaction: dynamic HTML, JavaScript, HTML forms
- ▷ Web application project (fill in the blanks to obtain a working web app)

1.5 AI-Supported Learning for IWGS

In this section we introduce the ALeA (Adaptive Learning Assistant) system, a learning support system we will use to support students in IWGS.

ALeA: Adaptive Learning Assistant

- ▷ **Idea:** Use AI methods to help teach/learn AI (AI4AI)
- ▷ **Concretely:** Provide HTML versions of the IWGS slides/lecture notes and embed learning support services into them. (for pre/postparation of lectures)
- ▷ **Definition 1.5.1.** Call a document active, iff it is interactive and adapts to specific information needs of the readers. (lecture notes on steroids)
- ▷ **Intuition:** ALeA serves active course materials. (PDF mostly inactive)
- ▷ **Goal:** Make ALeA more like a instructor + study group than like a book!
- ▷ **Example 1.5.2 (Course Notes).** $\widehat{=}$ Slides + Comments


Search
Format of the AI Course/Lecturing
Resources
Artificial Intelligence — Who?, W...
What is Artificial Intelligence?
Artificial Intelligence is here today!
Ways to Attack the AI Problem
Strong vs. Weak AI
AI Topics Covered
AI in the KWARC Group
Getting Started with AI: A Conce...
Logic Programming
Introduction to Logic Programming
Programming as Search
Knowledge Bases and Backtracki
Programming Features
Advanced Relational Programmin
Recap of Prerequisites from Math & T

It is easy to see that the running time of the Prolog program from Example 5.2.9 (Programming Features) in the AI lecture notes is not $\mathcal{O}(n\log_2(n))$ which is optimal for sorting algorithms. This is the flip side of the flexibility in logic programming. But Prolog has ways of dealing with that: the cut operator, which is a Prolog atom, which always succeeds, but which cannot be backtracked over. This can be used to prune the search tree in Prolog. We will not go into that here but refer the readers to the literature.

Specifying Control in Prolog

▷ **Assertion 1.1.10.** The running time of the program from Example 5.2.9 (Programming Features) in the AI lecture notes is not $\mathcal{O}(n\log_2(n))$ which is optimal for sorting algorithms.

```
sort(Xs,Ys) :— perm(Xs,Ys), ordered(Ys).
```

▷ **Idea** Gain computational efficiency by shaping the search!

FAU Michael Kohlhase: Artificial Intelligence 122023-07-18

Functions and Predicates in Prolog

▷ **Assertion 1.1.11.** Functions and predicates have radically different roles in Prolog.
- ▷ Functions are used to represent data. (e.g. father(john) or s(s(zero)))
- ▷ Predicates are used for stating properties about and computing with data.

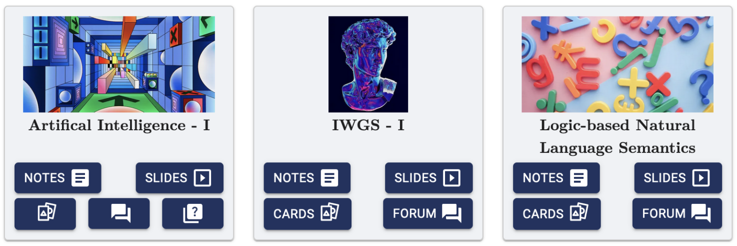
⇝ yellow parts in table of contents (left) already covered in lectures.

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 14 2025-02-06

The central idea in the AI4AI approach – using AI to support learning AI – and thus the ALeA system is that we want to make course materials – i.e. what we give to students for preparing and postparing lectures – more like teachers and study groups (only available 24/7) than like static books.

VoLL-KI Portal at `https://courses.voll-ki.fau.de`

- ▷ **Portal for ALeA Courses:** `https://courses.voll-ki.fau.de`


Artifical Intelligence - I
NOTES
SLIDES
IWGS - I
NOTES
SLIDES
CARDS
FORUM
Logic-based Natural Language Semantics
NOTES
SLIDES
CARDS
FORUM

- ▷ **IWGS in ALeA:** `https://courses.voll-ki.fau.de/course-home/iwgs-1`

- ▷ All details for the course.
- ▷ recorded syllabus (keep track of material covered in course)
- ▷ syllabus of the last semesters (for over/preview)

▷ **ALeA Status:** The ALeA system is deployed at FAU for over 1000 students taking eight courses

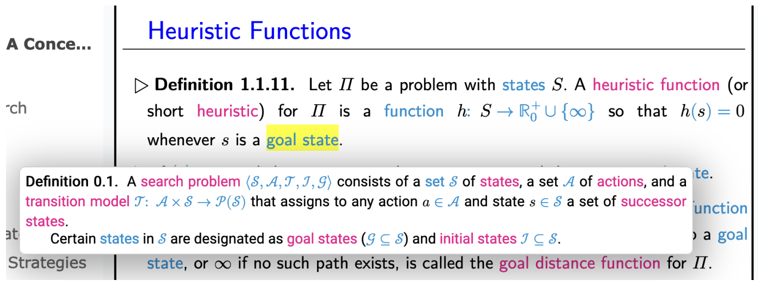
- ▷ (some) students use the system actively (our logs tell us)
- ▷ reviews are mostly positive/enthusiastic (error reports pour in)

The ALeA IWGS page is the central entry point for working with the ALeA system. You can get to all the components of the system, including two presentations of the course contents (notes- and slides-centric ones), the flashcards, the localized forum, and the quiz dashboard.
We now come to the heart of the ALeA system: its learning support services, which we will now briefly introduce. Note that this presentation is not really sufficient to undertstand what you may be getting out of them, you will have to try them, and interact with them sufficiently that the learner model can get a good estimate of your competencies to adapt the results to you.

Learning Support Services in ALeA

▷ **Idea:** Embed learning support services into active course materials.

▷ **Example 1.5.3 (Definition on Hover).** Hovering on a (cyan) term reference reminds us of its definition. (even works recursively)



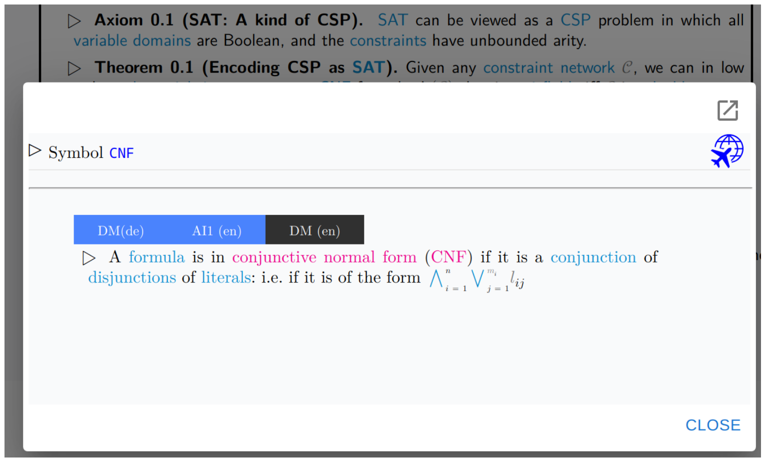
Heuristic Functions

▷ **Definition 1.1.11.** Let Π be a problem with states S. A heuristic function (or short heuristic) for Π is a function $h\colon S \to \mathbb{R}^+_0 \cup \{\infty\}$ so that $h(s) = 0$ whenever s is a goal state.

Definition 0.1. A search problem $\langle \mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$ consists of a set $\mathcal{S}$ of states, a set $\mathcal{A}$ of actions, and a transition model $\mathcal{T}\colon \mathcal{A}\times\mathcal{S}\to\mathcal{P}(\mathcal{S})$ that assigns to any action $a\in\mathcal{A}$ and state $s\in\mathcal{S}$ a set of successor states.
Certain states in $\mathcal{S}$ are designated as goal states ($\mathcal{G}\subseteq\mathcal{S}$) and initial states $\mathcal{I}\subseteq\mathcal{S}$.

te.
unction
a goal
state, or ∞ if no such path exists, is called the goal distance function for Π.

▷ **Example 1.5.4 (More Definitions on Click).** Clicking on a (cyan) term reference shows us more definitions from other contexts.


▷ **Axiom 0.1 (SAT: A kind of CSP).** SAT can be viewed as a CSP problem in which all variable domains are Boolean, and the constraints have unbounded arity.

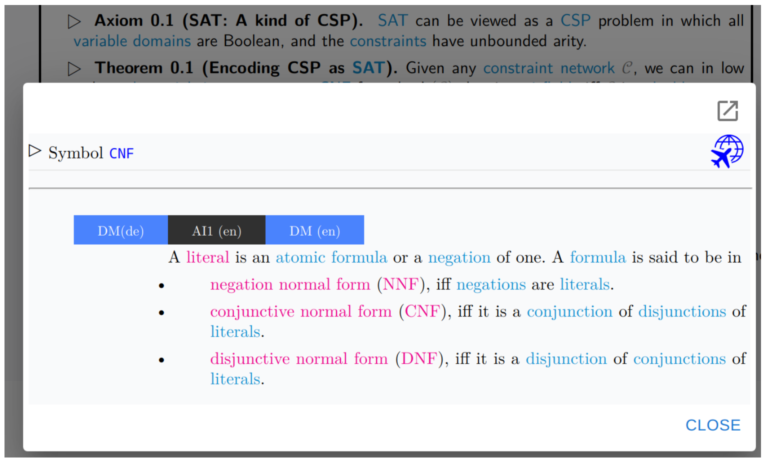
▷ **Theorem 0.1 (Encoding CSP as SAT).** Given any constraint network $\mathcal{C}$, we can in low

▷ Symbol CNF

DM(de) AI1 (en) DM (en)

▷ A formula is in conjunctive normal form (CNF) if it is a conjunction of disjunctions of literals: i.e. if it is of the form $\bigwedge_{i=1}^{n} \bigvee_{j=1}^{m_i} l_{ij}$

CLOSE


▷ **Axiom 0.1 (SAT: A kind of CSP).** SAT can be viewed as a CSP problem in which all variable domains are Boolean, and the constraints have unbounded arity.

▷ **Theorem 0.1 (Encoding CSP as SAT).** Given any constraint network $\mathcal{C}$, we can in low

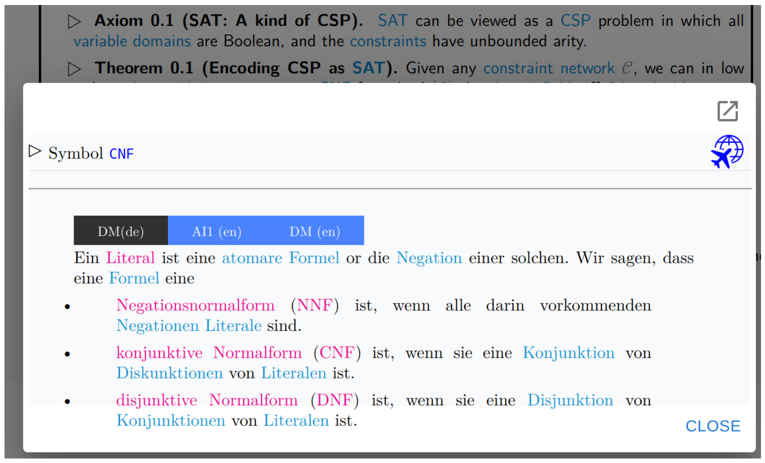
▷ Symbol CNF

DM(de) AI1 (en) DM (en)

A literal is an atomic formula or a negation of one. A formula is said to be in

- negation normal form (NNF), iff negations are literals.
- conjunctive normal form (CNF), iff it is a conjunction of disjunctions of literals.
- disjunctive normal form (DNF), iff it is a disjunction of conjunctions of literals.

CLOSE


▷ **Axiom 0.1 (SAT: A kind of CSP).** SAT can be viewed as a CSP problem in which all variable domains are Boolean, and the constraints have unbounded arity.

▷ **Theorem 0.1 (Encoding CSP as SAT).** Given any constraint network $\mathcal{C}$, we can in low

▷ Symbol CNF

DM(de) AI1 (en) DM (en)

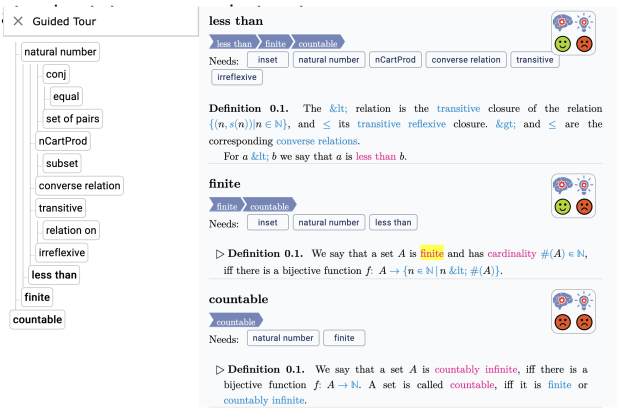
Ein Literal ist eine atomare Formel or die Negation einer solchen. Wir sagen, dass eine Formel eine

- Negationsnormalform (NNF) ist, wenn alle darin vorkommenden Negationen Literale sind.
- konjunktive Normalform (CNF) ist, wenn sie eine Konjunktion von Diskunktionen von Literalen ist.
- disjunktive Normalform (DNF) ist, wenn sie eine Disjunktion von Konjunktionen von Literalen ist.

CLOSE

▷ **Example 1.5.5 (Guided Tour).** A guided tour for a concept c assembles definitions/etc. into a self-cont

c = countable ⇝



less than

less than › finite › countable

Needs: inset | natural number | nCartProd | converse relation | transitive | irreflexive

Definition 0.1. The &lt; relation is the transitive closure of the relation $\{(n,s(n))|n\in\mathbb{N}\}$, and $\leq$ its transitive reflexive closure. &gt; and $\leq$ are the corresponding converse relations.

For a &lt; b we say that a is less than b.

finite

finite › countable

Needs: inset | natural number | less than

▷ **Definition 0.1.** We say that a set A is finite and has cardinality $\#(A)\in\mathbb{N}$, iff there is a bijective function $f\colon A\to\{n\in\mathbb{N}\,|\,n$ &lt; $\#(A)\}$.

countable

countable

Needs: natural number | finite

▷ **Definition 0.1.** We say that a set A is countably infinite, iff there is a bijective function $f\colon A\to\mathbb{N}$. A set is called countable, iff it is finite or countably infinite.

▷ . . . your idea here . . . (the sky is the limit)

Note that this is only an initial collection of learning support services, we are constantly working on additional ones. Look out for feature notifications (
) on the upper right hand of the ALeA screen.

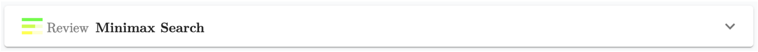
(Practice/Remedial) Problems Everywhere

▷ **Problem:** Learning requires a mix of understanding and test-driven practice.

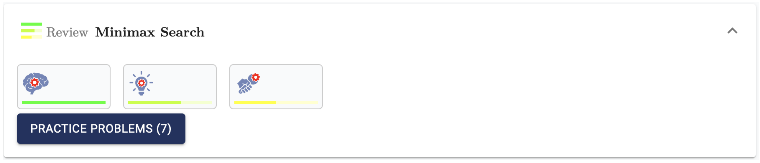
▷ **Idea:** ALeA supplies targeted practice problems everywhere.

▷ **Concretely:** Revision markers at the end of sections.

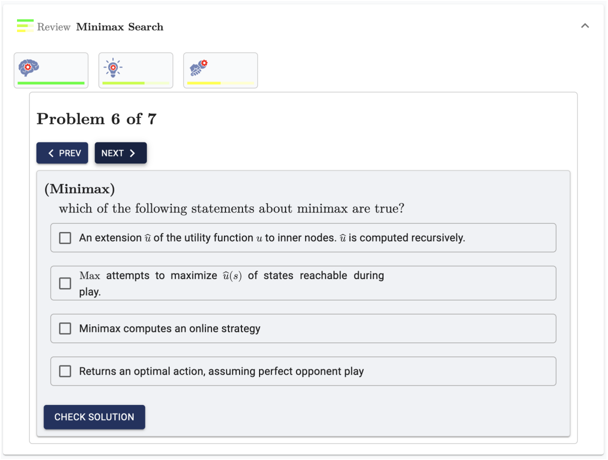
- ▷ A relatively non-intrusive overview over competency



- ▷ Click to extend it for details.



- ▷ Practice problems as usual. (targeted to your specific competency)


Review **Minimax Search**

Problem 6 of 7

PREV NEXT

(Minimax)
which of the following statements about minimax are true?

- [ ] An extension $\hat{u}$ of the utility function u to inner nodes. $\hat{u}$ is computed recursively.
- [ ] Max attempts to maximize $\hat{u}(s)$ of states reachable during play.
- [ ] Minimax computes an online strategy
- [ ] Returns an optimal action, assuming perfect opponent play

CHECK SOLUTION

While the learning support services up to now have been adressed to individual learners, we now turn to services addressed to communities of learners, ranging from study groups with three learners, to whole courses, and even – eventually – all the alumni of a course, if they have not de-registered from ALeA.

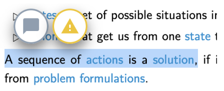
Currently, the community aspect of ALeA only consists in localized interactions with the course materials.
The ALeA system uses the semantic structure of the course materials to localize some interactions that are otherwise often from separate applications. Here we see two:

1. one for reporting content errors – and thus making the material better for all learners – and‘’

2. a localized course forum, where forum threads can be attached to learning objects.

Localized Interactions with the Community

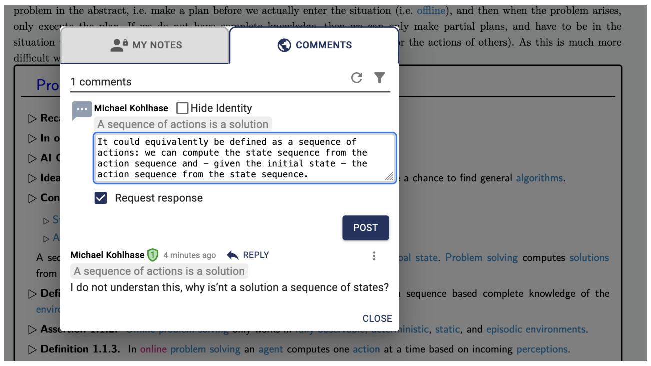
▷ Selecting text brings up localized – i.e. anchored on the selection – interactions:


A sequence of actions is a solution, if i
from problem formulations.

▷ post a (public) comment or take (private) note

▷ report an error to the course authors/instructors

▷ Localized comments induce a thread in the ALeA forum (like the StudOn Forum, but targeted towards specific learning objects.)



1 comments

Michael Kohlhase Hide Identity

A sequence of actions is a solution

```
It could equivalently be defined as a sequence of
actions: we can compute the state sequence from the
action sequence and – given the initial state – the
action sequence from the state sequence.
```

☑ Request response

POST

Michael Kohlhase 4 minutes ago REPLY

A sequence of actions is a solution

I do not understan this, why is'nt a solution a sequence of states?

CLOSE

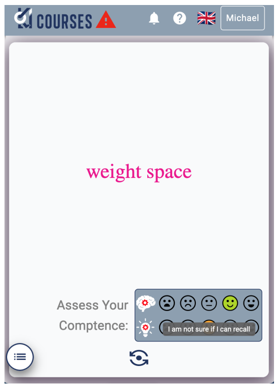
▷ Answering questions gives karma $\hat{=}$ a public measure of user helpfulness.

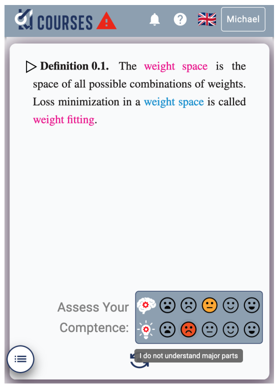
▷ Notes can be anonymous (⇝ generate no karma)

Let us briefly look into how the learning support services introduced above might work, focusing on where the necessary information might come from.

New Feature: Drilling with Flashcards

▷ Flashcards challenge you with a task (term/problem) on the front. . .




▷ **Definition 0.1.** The weight space is the space of all possible combinations of weights. Loss minimization in a weight space is called weight fitting.

. . . and the definition/answer is on the back.

▷ Self-assessment updates the learner model (before/after)

▷ **Idea:** Challenge yourself to a card stack, keep drilling/assessing flashcards until the learner model eliminates all.

▷ **Bonus:** Flashcards can be generated from existing semantic markup (educational equivalent to free beer)

We have already seen above how the learner model can drive the drilling with flashcards. It can also be used for the configuration of card stacks by configuring a domain e.g. a section in the course materials and a competency threshold. We now come to a very important issue that we always face when we do AI systems that interface with humans. Most web technology companies that take one the approach "the user pays for the services with their personal data, which is sold on" or integrate advertising for renumeration. Both are not acceptable in university setting.

But abstaining from monetizing personal data still leaves the problem how to protect it from intentional or accidental misuse. Even though the GDPR has quite extensive exceptions for research, the ALeA system – a research prototype – adheres to the principles and mandates of the GDPR. In particular it makes sure that personal data of the learners is only used in learning support services directly or indirectly initiated by the learners themselves.

Learner Data and Privacy in ALeA

▷ **Observation:** Learning support services in ALeA use the learner model; they

- ▷ need the learner model data to adapt to the invidivual learner!
- ▷ collect learner interaction data (to update the learner model)

▷ **Consequence:** You need to be logged in (via your FAU IDM credentials) for useful learning support services!

▷ **Problem:** Learner model data is highly sensitive personal data!

▷ **ALeA Promise:** The ALEA team does the utmost to keep your personal data safe. (SSO via FAU IDM/eduGAIN, ALEA trust zone)

▷ **ALeA Privacy Axioms:**

1. ALEA only collects learner models data about logged in users.
2. Personally identifiable learner model data is only accessible to its subject (delegation possible)
3. Learners can always query the learner model about its data.
4. All learner model data can be purged without negative consequences (except usability deterioration)
5. Logging into ALEA is completely optional.

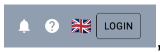
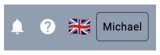
▷ **Observation:** Authentication for bonus quizzes are somewhat less optional, but you can always purge the learner model later.

So, now that you have an overview over what the ALEA system can do for you, let us see what you have to concretely do to be able to use it.

Concrete Todos for ALeA

▷ **Recall:** You will use ALeA for the prepquizzes (or lose bonus points)
All other use is optional. (but AI-supported pre/postparation can be helpful)

▷ To use the ALeA system, you will have to log in via SSO: (do it now)

- ▷ go to `https://courses.voll-ki.fau.de/course-home/iwgs-1`,
- ▷ in the upper right hand corner you see 
,
- ▷ log in via your FAU IDM credentials. (you should have them by now)
- ▷ You get access to your personal ALeA profile via 

(plus feature notifications, manual, and language chooser)

▷ **Problem:** Most ALeA services depend on the learner model. (to adapt to you)

▷ **Solution:** Initialize your learner model with your educational history!

- ▷ **Concretely**: enter taken CS courses (FAU equivalents) and grades.
- ▷ ALeA uses that to estimate your CS/AI competencies. (for your benefit)
- ▷ then ALeA knows about you; I don't! (ALeA trust zone)

Even if you did not understand some of the AI jargon or the underlying methods (yet), you should be good to go for using the ALeA system in your day-to-day work.

Part I

IWGS-1: Programming, Documents, Web Applications

Chapter 2

Introduction to Programming

2.1 What is Programming?

Programming is an important and distinctive part of "Informatische Werkzeuge in den Geistes- und Sozialwissenschaften" – the topic of this course. Before we delve into learning Python, we will review some of the basics of computing to situate the discussion.
To understand programming, it is important to realize that computers are universal machines. Unlike a conventional tool e.g a spade – which has a limited number of purposes/behaviors – digging holes in case of a spade, maybe hitting someone over the head, a computer can be given arbitrary[1] purposes/behaviors by specifying them in form of a program.
This notion of a program as a behavior specification for an universal machine is so powerful, that the field of computer science is centered around studying it – and what we can do with programs, this includes

i) storing and manipulating data about the world,

ii) encoding, generating, and interpreting image, audio, and video,

iii) transporting information for communication,

iv) representing knowledge and reasoning,

v) transforming, optimizing, and verifying other programs,

vi) learning patterns in data and predicting the future from the past.

Computer Hardware/Software & Programming

▷ **Definition 2.1.1.** Computers consist of hardware and software.

▷ **Definition 2.1.2.** Hardware consists of

[1]as long as they are "computable", not all are.

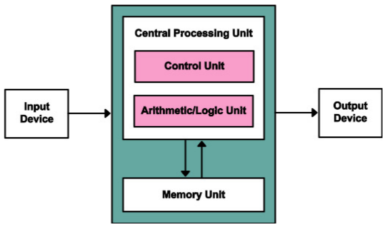
▷ a central processing unit (CPU)

▷ memory: e.g. RAM, ROM, . . .

▷ storage devices: e.g. Disks, SSD, tape, . . .

▷ input: e.g. keyboard, mouse, touchscreen, . . .

▷ output: e.g. screen, earphone, printer, . . .



▷ **Definition 2.1.3.** Software consists of

▷ data that represents objects and their relationships in the world

▷ programs that inputs, manipulates, outputs data

$$\begin{tikzcd} \text{data} \arrow[rr, leftrightarrow] \arrow[dr, leftrightarrow] & & \text{program} \arrow[dl, leftrightarrow] \\ & \text{hardware} & \end{tikzcd}$$

▷ **Remark:** Hardware stores data and runs programs.

A universal machine has to have – so experience in computer science shows certain distinctive parts.

- A CPU that consists of a
  - control unit that interprets the program and controls the flow of instructions and
  - a arithmetic/logic unit (ALU) that does the actual computations internally.
- Memory that allows the system to store data during runtime (volatile storage; usually RAM) and between runs of the system (persistant storage; usually hard disks, solid state disks, magnetic tapes, or optical media).
- I/O devices for the communication with the user and other computers.

With these components we can build various kinds of universal machines; these range from thought experiments like Turing machines, to today's general purpose computers like your laptop with various embedded systems (wristwatches, Internet routers, airbag controllers, . . .) in-between. Note that – given enough fantasy – the human brain has the same components. Indeed the human mind is a universal machine – we can think whatever we want, react to the environment, and are not limited to particular behaviors. There is a sub-field of computer science that studies this: Artificial Intelligence (AI). In this analogy, the brain is the "hardware" –sometimes called "wetware" because it is not made of hard silicon or "meat machine"[2]. It is instructional to think about what the program and the data might be in this analogy.

Programming Languages

▷ Programming $\hat{=}$ writing programs (Telling the computer what to do)

▷ *Remark 2.1.4.* The computer does exactly as told

▷ extremely fast extremely reliable

▷ completely stupid: will not do what you mean unless you tell it exactly

[2]Marvin Minsky; one of the founding fathers of the field of Artificial Intelligence

▷ Programming can be extremely fun/frustrating/addictive (try it)

▷ **Definition 2.1.5.** A programming language is the formal language in which we write programs (express an algorithm concretely)

▹ formal, symbolic, precise meaning (a machine must understand it)

▷ There are lots of programming languages

▹ design huge effort in computer science

▹ all programming languages equally strong

▹ each is more or less appropriate for a specific task depending on the circumstances

▷ Lots of programming paradigms: imperative, functional, logic, object oriented programming.

AI studies human intelligence with the premise that the brain is a computational machine and that intelligence is a "program" running on it. In particular, the working hypothesis is that we can "program" intelligence. Even though AI has many successful applications, it has not succeeded in creating a machine that exhibits the equivalent to general human intelligence, so the jury is still out whether the AI hypothesis is true or not. In any case it is a fascinating area of scientific inquiry.

Note: This has an immediate consequence for the discussion in our course. Even though computers can execute programs very efficiently, you should not expect them to "think" like a human. In particular, they will execute programs exactly as you have written them. This has two consequences:

- the behavior of programs is – in principle – predictable
- all errors of program behavior are your own (the programmer's)

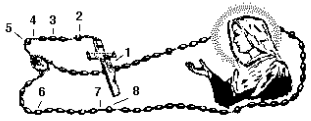
In computer science, we distinguish two levels on which we can talk about programs. The more general is the level of algorithms, which is independent of the concrete programming language. Algorithms express the general ideas and flow of computation and can be realized in various languages, but are all equivalent – in terms of the algorithms they implement.

As they are not bound to programming languages algorithms transcend them, and we can find them in our daily lives, e.g. as sequences of instructions like recipes, game instructions, and the like. This should make algorithms quite familiar; the only difference of programs is that they are written down in an unambiguous syntax that a computer can understand.

Program Execution

▷ **Definition 2.1.6.** Algorithm: informal description of what to do (good enough for humans)

▷ **Example 2.1.7.**



▷ **Example 2.1.8.** Program: computer processable version, e.g. in Python.

```
for x in range(0, 3):
    print ("we tell you",x,"time(s)")
```

▷ **Definition 2.1.9.** Interpreter: reads a program and executes it directly

- ▹ special case: interactive interpretation (lets you experiment easily)

▷ **Definition 2.1.10.** Compiler: translates a program (the source) into another program (the binary) in a much simpler programming language for optimized execution on hardware directly.

▷ *Remark 2.1.11.* Compilers are efficient, but more cumbersome for development.

We have two kinds of programming languages: one which the CPU can execute directly – these are very very difficult for humans to understand and maintain – and higher-level ones that are understandable by humans. If we want to use high-level languages – and we do, then we need to have some way bridging the language gap: this is what compilers and interpreters do.

2.2 Programming in IWGS

After the general introduction to programming in **??**, we now instantiate the situation to the IWGS course, where we use Python as the primary programming language.

Programming in IWGS: Python

▷ We will use Python as the programming language in this course

▷ We cover just enough Python, so that you

- ▹ understand the joy and principle of programming
- ▹ can play with objects we present in IWGS.

▷ After a general introduction we will introduce language features as we go along

▷ For more information on Python (homework/preparation)

RTFM ($\widehat{=}$ "read those fine manuals")

▷ **RTFM Resources:** There are also lots of good tutorials on the web,

- ▹ I like [LP; Sth; Swe13];
- ▹ but also see the language documentation [P3D].
- ▹ [Kar] is an introduction geared to the (digital) humanities

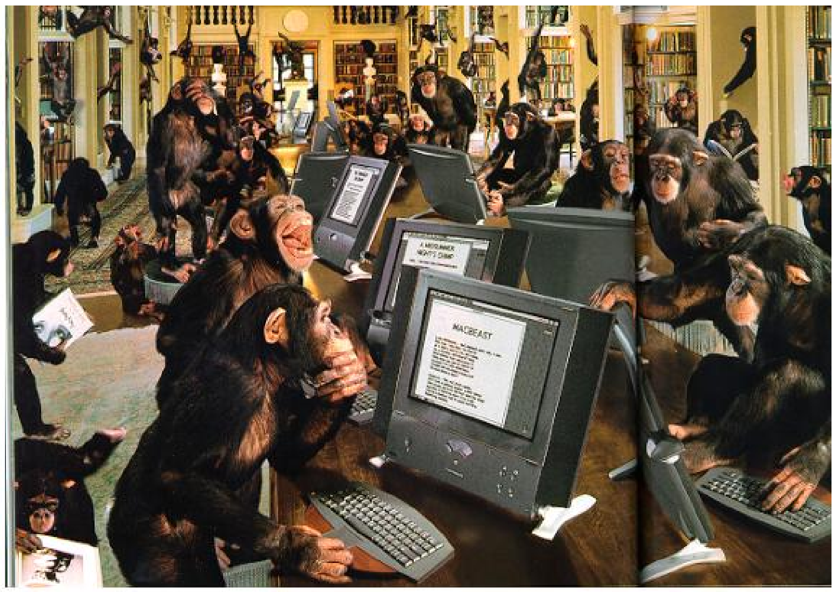
Note that IWGS is not a programming course, which concentrates on teaching a programming language in all it gory detail. Instead we want to use the IWGS lectures to introduce the necessary concepts and use the tutorials to introduce additional language features based on these.

But Seriously. . . Learning programming in IWGS

- ▷ The IWGS course teaches you
  - ▹ a general introduction to programming and Python (next)
  - ▹ various useful concepts and how they can be done in Python (in principle)
- ▷ The IWGS tutorials:
  - ▹ teach the actual skill and joy of programming (hacking $\neq$ security breach)
  - ▹ supply you with problems so you can practice that.
- ▷ **Richard Stallman (MIT) on Hacking:** "What they had in common was mainly love of excellence and programming. They wanted to make their programs that they used be as good as they could. They also wanted to make them do neat things. They wanted to be able to do something in a more exciting way than anyone believed possible and show "Look how wonderful this is. I bet you didn't believe this could be done."
- ▷ **So, ...** Let's hack

However, the result would probably be the following:

2am in the Kollegienhaus CIP Pool



If we just start hacking before we fully understand the problem, chances are very good that we will waste time going down blind alleys, and garden paths, instead of attacking problems. So the main motto of this course is:

no, let's think

- We have to fully understand the problem, our tools, and the solution space first (That is what the IWGS course is for)
  - read Richard Stallman's quote carefully ⇝ problem understanding is a crucial prerequisite for hacking.
- *The GIGO Principle: Garbage In, Garbage Out* (– ca. 1967)
- *Applets, Not Craplets*^{tm} (– ca. 1997)

2.3 Programming in Python

In this section we will introduce the basics of the Python language. Python will be used as our means to express algorithms and to explore the computational properties of the objects we introduce in IWGS.

2.3.1 Hello IWGS

Before we get into the syntax and meaning of Python, let us recap why we chose this particular language for IWGS.

Python in a Nutshell

- **Why Python?:**
  - general purpose programming language
  - imperative, interactive interpreter
  - syntax very easy to learn (spend more time on problem solving)
  - scales well:
    - easy for beginners to write simple programs,
    - but advanced software can be written with it as well.
- **Interactive mode:** The Python shell IDLE3
- **For the eager (optional):**
  Establish a Python interpreter (version 3.7) (not 2.?.?, that has different syntax)
  - install Python from `http://python.org` (for offline use)
  - make sure (tick box) that the python executable is added to the path. (makes shell interaction much easier)

Installing Python: Python can be installed from `http://python.org` ⇝ "Downloads", as a `MSWindows` installer or a `macOS` disk image. For `linux` it is best installed via the package manager, e.g. using

```
sudo apt−get update
sudo apt−get install python3.7
```

The download will install the Python interpreter and the Python shell IDLE3 that can be used for interacting with the interpreter directly.

It is important that you make sure (tick the box in the Windows installer) that the python executable is added to the path. In the shell[1], you can then use the command

```
python ⟪filename⟫
```

to run the python file ⟪filename⟫. This is better than using the windows-specific

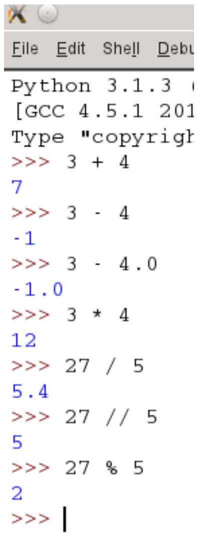
```
py ⟪filename⟫
```

which does not need the python interpreter on the path as we will see later.

Arithmetic Expressions in Python

▷ Expressions are "programs" that compute values (here: numbers)

- ▷ Integers (numbers without a decimal point)
  - ▷ operators: addition (+), subtraction (), multiplication (∗), division (/), integer division (//), remainder/modulo (%), . . .
  - ▷ Division yields a float
- ▷ Floats (numbers with a decimal point)
  - ▷ Operators: integer below (floor), integer above (ceil), exponential (exp), square root (sqrt), . . .
- ▷ Numbers are values, i.e. data objects that can be computed with. (reference the last computed one with _)
- ▷ **Definition 2.3.1.** Expressions are created from values (and other expressions) via operators.
- ▷ **Observation:** The Python interpreter simplifies expressions to values by computation.



```
Python 3.1.3
[GCC 4.5.1 201
Type "copyrigh
>>> 3 + 4
7
>>> 3 - 4
-1
>>> 3 - 4.0
-1.0
>>> 3 * 4
12
>>> 27 / 5
5.4
>>> 27 // 5
5
>>> 27 % 5
2
>>>
```

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 30 2025-02-06

Before we go on to learn more basic Python operators and instructions, we address an important general topic: comments in program code.

Comments in Python

▷ **Generally:** It is highly advisable to insert comments into your programs,

- ▷ especially, if others are going to read your code, (TAs/graders)
- ▷ you may very well be one of the "others" yourself, (in a year's time)
- ▷ writing comments first helps you organize your thoughts.

[1] EdNote: fully introduce the concept of a shell in the next round

▷ Comments are ignored by the Python interpreter but are useful information for the programmer.

▷ **In Python:** there are two kinds of comments

- ▷ Single line comments start with a #
- ▷ Multiline comments start and end with three quotes (single or double: `"""` or `'''`)

▷ **Idea:** Use comments to

- ▷ specify what the intended input/output behavior of the program or fragment
- ▷ give the idea of the algorithm achieves this behavior.
- ▷ specify any assumptions about the context (do we need some file to exist)
- ▷ document whether the program changes the context.
- ▷ document any known limitations or errors in your code.

2.3.2 JupyterLab, a Python Web IDE for IWGS

In IWGS, we want to use the jupyterLab cloud service. This runs the Python interpreter on a cloud server and gives you a browser window with a web IDE, which you can use for interacting with the interpreter. You will have to make an account there; details to follow.

jupyterLab A Cloud IDE for Python

▷ **For helping you** it would be good if the TAs could access to your code

▷ **Idea:** Use a web IDE (a web based integrated development environment): jupyterLab, which you can use for interacting with the interpreter.

▷ We will use jupyterLab for IWGS. (but you can also use Python locally)

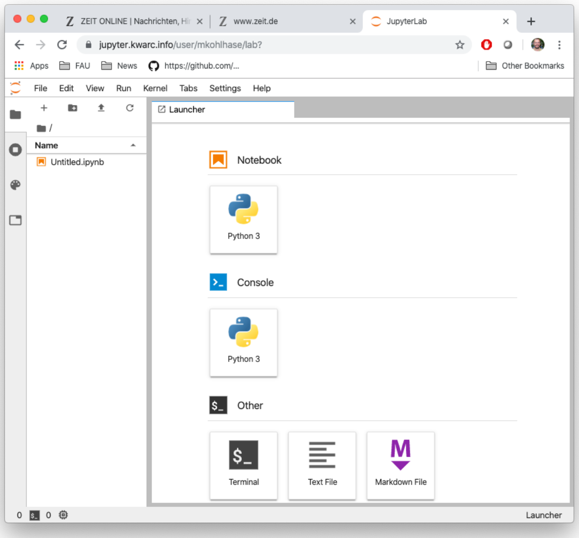
▷ **Homework:** Set up jupyterLab

- ▷ make an account at `http://jupyter.kwarc.info`

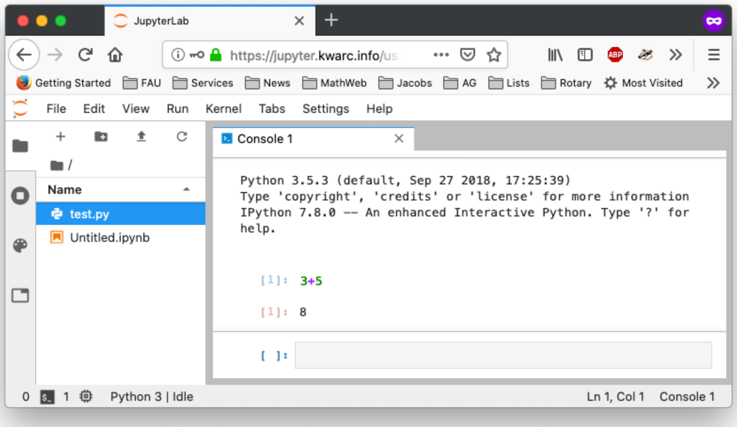
The advantage of a cloud IDE like jupyterLab for a course like IWGS is that you do not need any installation, cannot lose your files, and your teachers (the course instructor and the teaching assistants) can see (and even directly interact with) the your run time environment. This gives us a much more controlled setting and we can help you better.
Both IDLE3 as well as jupyterLab come with an integrated editor for writing Python programs. These editors gives you Python syntax highlighting, and interpreter and debugger integration. In short, IDLE3 and jupyterLab are integrated development environments for Python. Let us now go through the interface of the jupyterLab IDE.

jupyterLab Components

▷ **Definition 2.3.2.** The jupyterLab dashboard gives you access to all components.

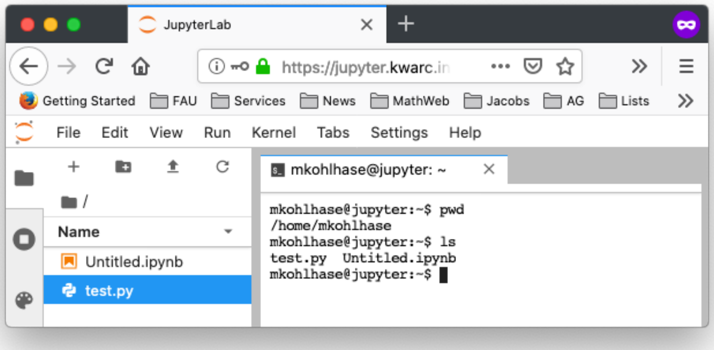


▷ **Definition 2.3.3.** The jupyterLab python console, i.e. a Python interpreter in your browser. (use this for Python interaction and testing.)



```
Python 3.5.3 (default, Sep 27 2018, 17:25:39)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.8.0 -- An enhanced Interactive Python. Type '?' for
help.

[1]: 3+5

[1]: 8

[ ]:
```

▷ **Definition 2.3.4.** The jupyterLab terminal, i.e. a `UNIX` shell in your browser. (use this for managing files)


JupyterLab
https://jupyter.kwarc.in
Getting Started FAU Services News MathWeb Jacobs AG Lists
File Edit View Run Kernel Tabs Settings Help
Name
Untitled.ipynb
test.py
mkohlhase@jupyter: ~

```
mkohlhase@jupyter:~$ pwd
/home/mkohlhase
mkohlhase@jupyter:~$ ls
test.py  Untitled.ipynb
mkohlhase@jupyter:~$
```

▷ **Definition 2.3.5.** A shell is a command line interface for accessing the services of a computer's operating system.

There are multiple shell implementations: sh, csh, bash, zsh; they differ in advanced features.

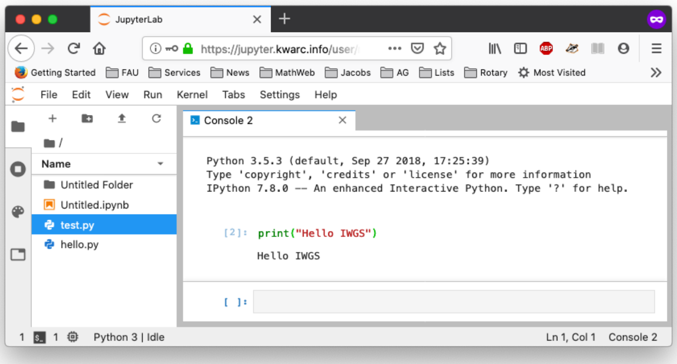
▷ **Useful shell commands:** See e.g. [All18] for a basic tutorial

- ▷ ls: "list" the files in this directory
- ▷ mkdir: "make" folder (called "directory")
- ▷ pwd: "print working directory" (where am I)
- ▷ cd ⟪dirname⟫: "change directory"
  - ▷ if ⟪dirname⟫ = ..: one up in the directory tree
  - ▷ empty ⟪dirname⟫: go to your home directory.
- ▷ rm ⟪name⟫: remove file/directory
- ▷ cp/mv ⟪filename⟫ ⟪newname⟫: copy to or rename
- ▷ cp/mv ⟪filename⟫ ⟪dirname⟫: copy or move to
- ▷ . . . see [All18] for more . . .

Now that we understand our tools, we can wrote our first program: Traditionally, this is a "hello-world program" (see [HWC] for a description and a list of hello world programs in hundreds of languages) which just prints the string "Hello World" to the console. For Python, this is very simple as we can see below. We use this program to explain the concept of a program as a (text) file, which can be started from the console.

A first program in Python

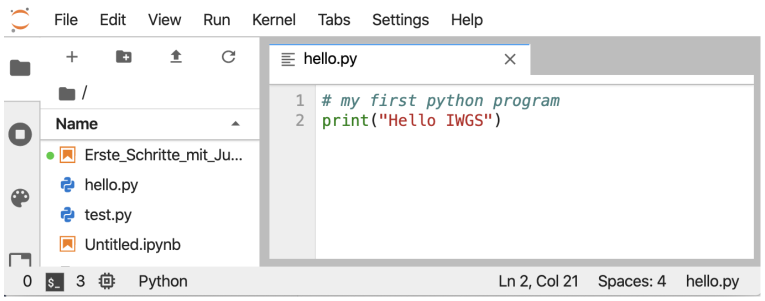
▷ **A classic "Hello World" program:** start your python console, type **print**("Hello IWGS"). (print a string)



```
Python 3.5.3 (default, Sep 27 2018, 17:25:39)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.8.0 -- An enhanced Interactive Python. Type '?' for help.

[2]: print("Hello IWGS")

     Hello IWGS

[ ]:
```

▷ **Alternatively:**

1. got to the jupyterLab dashboard select “Text File”,
2. Type your program,



```
# my first python program
print("Hello IWGS")
```

3. Save the file as hello.py
4. Go to your terminal and type python3 hello.py

3’ **Alternatively**: go to your python console and type (in the same directory)

```
import hello
```

We have seen that we can just call a program from the terminal, if we stored it in a file. In fact, we can do better: we can make our program behave like a native shell instruction.

1. The file extension .py is only used by convention, we can leave it out and simply call the file hello.

2. Then we can add a special Python comment in the first line

```
python ⟨⟨filename⟩⟩
```

   which the terminal interprets as “call the program python3 on me”.

3. Finally, we make the file hello executable, i.e. tell the terminal the file should behave like a shell command by issuing

```
chmod u+x booksapp
```

   in the directory where the file hello is stored.

4. We add the line

```
export PATH="./:${PATH}"
```

to the file .bashrc. This tells the terminal where to look for programs (here the respective current directory called .)

With this simple recipe we could in principle extend the repertoire of instructions of the terminal and automate repetitive tasks.
We now come to the signature component of jupyterLab: jupyter notebooks. They take the important practice of documenting code to a whole new level. Instead of just allowing comments in program files, they provide rich text cells, in which we can write elaborate text.

Jupyter Notebooks

▷ **Definition 2.3.6.** Jupyter notebooks are documents that combine live runnable code with rich, narrative text (for comments and explanations).

▷ **Definition 2.3.7.** Jupyter notebooks consist of cells which come in three forms:

- ▷ a raw cell shows text as is,
- ▷ a markdown cell interprets the contents as markdown text, (later more)
- ▷ a code cell interprets the contents as (e.g. Python) code.

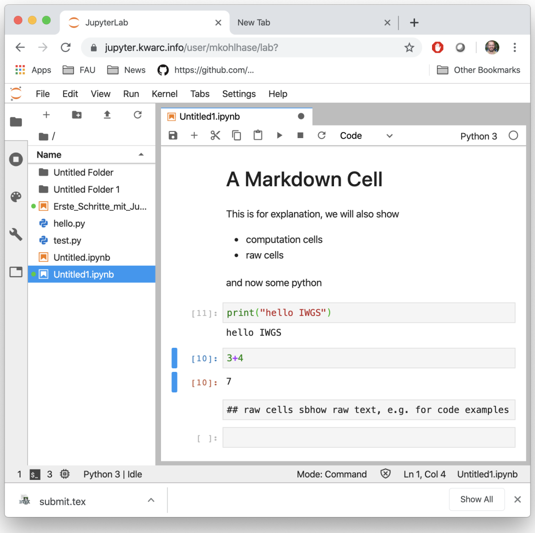
▷ Cells can be executed by pressing "shift enter". (Just "enter" gives a new line)

▷ **Idea:** Jupyter notebooks act as a REPL, just as IDLE3, but allows

- ▷ documentation in raw and markdown cells and
- ▷ changing and re-executing existing cells.

Jupyter Notebooks

▷ **Example 2.3.8 (Showing off Cells in a Notebook).**



A Markdown Cell

This is for explanation, we will also show

- computation cells
- raw cells

and now some python

```
[11]: print("hello IWGS")
hello IWGS
[10]: 3+4
[10]: 7
## raw cells sbhow raw text, e.g. for code examples
[ ]:
```

Markdown a simple Markup Format Generating HTML

.

- ▷ **Idea:** We can translate between markup formats.
- ▷ **Definition 2.3.9.** Markdown is a family of markup formats whose control words are unobtrusive and easy to write in a text editor. It is intended to be converted to HTML and other formats for display.
- ▷ **Example 2.3.10.** Markdown is used in applications that want to make user input easy and efficient, e.g. wikis and issue tracking systems.
- ▷ **Workflow:** Users write markdown, which is formatted to HTML and then served for display.
- ▷ A good cheet-sheet for markdown control words can be found at `https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet`.

2.3.3 Variables and Types

And we start with a general feature of programming languages: we can give names to values and use them multiple times. Conceptually, we are introducing shortcuts, and in reality, we are giving ourselves a way of storing values in memory so that we can reference them later.

Variables in Python

- ▷ **Idea:** Values (of expressions) can be given a name for later reference.
- ▷ **Definition 2.3.11.** A variable is an identifier (the variable name) that references

a memory location which contains a value. Storing the value v into the memory of a variable x is called assigning v to x.

A variable must be initialized (assigned an initial value) before it can be (usefully) referenced.

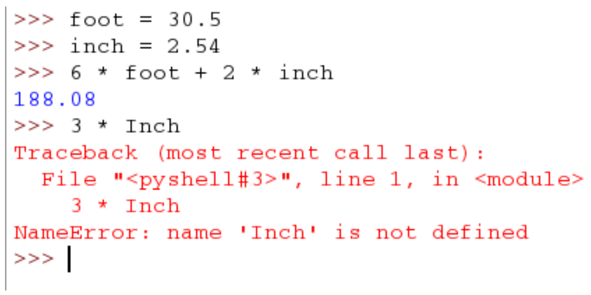
▷ **Note:** In Python a variable name

- ▷ must start with letter or _,
- ▷ cannot be a Python keyword
- ▷ is case-sensitive (foobar, FooBar, and fooBar are different variables)

▷ A variable name can be used in expressions everywhere its value could be.

▷ **Definition 2.3.12 (in Python).** A variable assignment $\langle\langle \mathrm{var}\rangle\rangle{=}\langle\langle \mathrm{val}\rangle\rangle$ assigns a new value to a variable.

▷ **Example 2.3.13 (Playing with Python Variables).**



```
>>> foot = 30.5
>>> inch = 2.54
>>> 6 * foot + 2 * inch
188.08
>>> 3 * Inch
Traceback (most recent call last):
  File "<pyshell#3>", line 1, in <module>
    3 * Inch
NameError: name 'Inch' is not defined
>>> |
```

Let us fortify our intuition about variables with some examples. The first shows that we sometimes need variables to store objects out of the way and the second one that we can use variables to assemble intermeditate results.

Variables in Python: Extended Example

▷ **Example 2.3.14 (Swapping Variables).** To exchange the values of two variables, we have to cache the first in an auxiliary variable.

```
a = 45
b= 0
print("a =", a, "b =", b)
print("Swap the contents of a and b")
swap = a
a= b
b = swap
print("a =", a, "b =", b)
```

Here we see the first example of a Python script, i.e. a series of Python commands, that jointly perform an action (and communicates it to the user).

▷ **Example 2.3.15 (Variables for Storing Intermediate Variables).**

```
>>> x = "OhGott"
>>> y = x+x+x
>>> z = y+y+y
>>> z
'OhGottOhGottOhGottOhGottOhGottOhGottOhGottOhGottOhGott'
```

If we use variables to assemble intermediate results, we can use telling names to document what these intermediate objects are – something we did not do well in ??; but admittely, the meaning of the objects in this contrived example is questionable.
The next phenomenon in Python is also common to many (but not all) programming languages: expressions are classified by the kind of objects their values are. Objects can be simple (i.e. of a basic type; Python has five of these) or complex, i.e. composed of other objects; we will go into that below.

Data Types in Python

▷ **Recall:** Python programs process data (values), which can be combined by operators and variable into expressions.

▷ Data types group data and tell the interpreter what to expect

- ▷ 1, 2, 3, etc. are data of type "integer"
- ▷ "hello" is data of type "string"

▷ Data types determine which operators can be applied

▷ In Python, every values has a type, variables can have any type, but can only be assigned values of their type.

▷ **Definition 2.3.16.** Python has the following five basic types

Data type	Keyword	contains	Examples
integers	int	bounded integers	1, −5, 0, ...
floats	float	floating point numbers	1.2, .125, −1.0, ...
strings	str	strings	"Hello", 'Hello', "123", 'a', ...
Booleans	bool	truth values	True, False
complexes	complex	complex numbers	2+3j,...

▷ We will ecounter more types later.

We will now see what we can – and cannot – do with data types, this becomes most noticable in variable assignments which establishes a type for the variable (this cannot be change any more) and in the application of operators to arguments (which have to be of the correct type).

Data Types in Python (continued)

▷ The type of a variable is automatically determined in the first variable assignment (before that the variable is unbound)

```
>>> firstVariable = 23 # integer
>>> type(firstVariable)
<class 'int'>
weight = 3.45 # float
```

```
first = 'Hello' # str
```

▷ **Hint:** The Python function type to computes the type (don't worry about the **class** bit)

Data Types in Python (continued)

▷ **Observation 2.3.17.** *Python is strongly typed, i.e. types have to match*

▷ Use data type conversion functions int(), float(), complex(), bool(), and str() to adjust types

▷ **Example 2.3.18 (Type Errors and Type Coersion).**

```
>>> 3+"hello"
Traceback (most recent call last):
  File "<pyshell#1>", line 1, in <module>
    3+"hello"
TypeError: unsupported operand type(s) for +: 'int' and 'str'
>>> str(4)+"hello"
'4Hello'
```

2.3.4 Python Control Structures

So far, we only know how to make programs that are a simple sequence of instructions no repetitions, no alternative pathways. ?? is a perfect example. We will now change that by introducing control structures, i.e complex program instructions that change the control flow of the program.

Conditionals and Loops

▷ **Problem:** Up to now programs seem to execute all the instructions in sequence, from the first to the last. (a linear program)

▷ **Definition 2.3.19.** The control flow of a program is the sequence of execution of the program instructions. It is specified via special program instructions called control structures.

▷ **Definition 2.3.20.** Conditional execution (also called branching) allows to execute (or not to execute) certain parts of a program (the branches) depending on a condition. We call a code block that enables conditional execution a conditional statement or conditional.

▷ **Definition 2.3.21.** A condition is a Boolean expression in a control structure.

▷ **Definition 2.3.22.** A loop is a control structure that allows to execute certain parts of a program (the body) multiple times depending on the value of its conditions. A break instruction terminate the loop irrespective of the value of the condition.

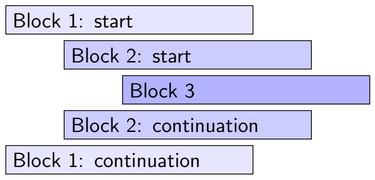
▷ **Example 2.3.23.** In Python, conditions are constructed by applying a Boolean operator to arguments, e.g. 3>5, x==3, x!=3, . . .
or by combining simpler conditions by Boolean connectives or, and, and not (using brakets if necessary), e.g. x>5 or x<3

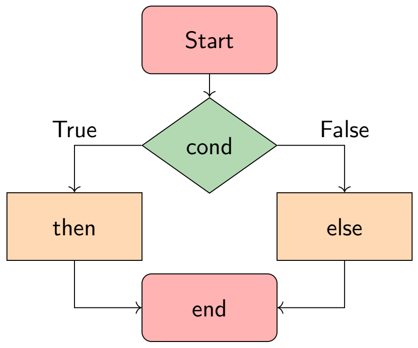
After this general introduction – conditional execution and loops are supported by all programming languages in some form – we will see how this is realized in Python

Conditionals in Python

▷ **Definition 2.3.24.** Conditional execution via **if**/**else** statements

```
if ⟨⟨condition⟩⟩ :
    ⟨⟨then − part⟩⟩
else :
    ⟨⟨else − part⟩⟩
⟨⟨morecode⟩⟩
```





▷ ⟨⟨then − part⟩⟩ and ⟨⟨else − part⟩⟩ have to be indented equally. (e.g. 4 blanks)

▷ If control structures are nested they need to be further indented consistently.

Python uses indenting to signify nesting of body parts in control structures – and other structures as we will see later. This is a very un-typical syntactic choice in programming languages, which typically use brackets, braces, or other paired delimiters to indicate nesting and give the freedom of choice in indenting to programmers. This freedom is so ingrained in programming practice, that we emphasize the difference here. The following example shows conditional execution in action.

Conditional Execution Example

▷ **Example 2.3.25 (Empathy in Python).**

```
answer = input("Are you happy? ")
if answer == 'No' or answer == 'no':
    print("Have a chocolate!")
else:
    print("Good!")
print("Can I help you with something else?")
```

Note the indenting of the body parts.

▷ **BTW:** input is an operator that prints its argument string, waits for user input, and returns that.

But conditional execution in Python has one more trick up its sleeve: what we can do with two branches, we can do with more as well.

Variant: Multiple Branches

▷ Making multiple branches is similar

```
if ⟪condition⟫ :
    ⟪then − part⟫
elif ⟪condition⟫ :
    ⟪otherthen − part⟫
else :
    ⟪else − part⟫
```

▷ The there can be more than one **elif** clause.

▷ The ⟪condition⟫s are evaluated from top to bottom and the ⟪then − part⟫ of the first one that comes out true is executed. Then the whole control structure is exited.

▷ multiple branches could achieved by nested **if**/**else** structures.

▷ **Example 2.3.26 (Better Empathy in Python).** In ?? we print Good! even if the input is e.g. I feel terrible, so extend **if**/**else** by

```
elif answer == 'Yes' or answer == 'yes' :
    print("Good!")
else :
    print("I do not understand your answer")
```

Note that the **elif** is just "syntactic sugar" that does not add anything new to the language: we could have expressed the same functionality as two nested if/else statements

```
if ⟪condition⟫ :
    ⟪then − part⟫
    if ⟪condition⟫ :
        ⟪otherthen − part⟫
    else :
        ⟪else − part⟫
```

But this would have introduced an additional layer of nesting (per **elif** clause in the original). The nested syntax also obscures the fact that all branches are essentially equal.
Now let us see the syntax for loops in Python.

Loops in Python

▷ **Definition 2.3.27.** Python makes loops via **while** blocks

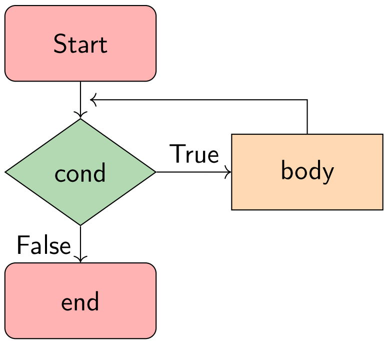
▷ syntax of the **while** loop

```
while ⟨⟨condition⟩⟩ :
    ⟨⟨body⟩⟩
⟨⟨morecode⟩⟩
```

▷ breaking out of loops with **break**

▷ skipping the current body with **continue**

▷ ⟨⟨body⟩⟩ must be indented!



As always we will fortify our intuition with a couple of small examples.

Examples of Loops

▷ **Example 2.3.28 (Counting in python).**

```
# Prints out 0,1,2,3,4
count = 0
while count < 5:
    print(count)
    count += 1 # This is the same as count = count + 1
```

This is the standard pattern for using **while**: using a loop variable (here count) and incrementing it in every pass through the loop.

▷ **Example 2.3.29 (Breaking an unbounded Loop).**

```
# Prints out 0,1,2,3,4 but uses break
count = 0
while True:
    print(count)
    count += 1
    if count >= 5:
        break
```

?? and **??** do the same thing: counting from zero to four, but using different mechanisms. This is normal in programming there is not “one correct solution”. But the first solution is the “standard one”, and is preferred, sind it is shorter and more readable. The **break** functionality shown off in the second one is still very useful. Take for instance the problem of computing the product of the numbers -10 to 1.000.000. The naive implementation of this is on the left below which does a lot of unnecessary work, because as soon was we passed 0, then the whole product must be zero. A more efficient implementation is on the right which breaks after seeing the first zero.

Direct Implementation

```
count = −10
prod = 1
while count < 1000000:
    prod ∗= count
    count += 1
```

More Efficient

```
count = −10
prod = 1
while count <= 1000000:
    prod ∗= count
    if count = 0 :
        break
    count += 1
```

Examples of Loops

▷ **Example 2.3.30 (Exceptions in the Loop).**

```
# Prints out only odd numbers — 1,3,5,7,9
count = 0
while count < 10
    count += 1
    # Check if x is even
    if count % 2 == 0:
        continue
    print(count)
```

2.4 Some Thoughts about Computers and Programs

Finally, we want to go over a couple of general issues pertaining to programs and (universal) machines. We will just go over them to get the intuitions – which are central for understanding computer science and let the lecture "Theoretical Computer Science" fill in the details and justifications.

Computers as Universal Machines (a taste of theoretical CS)

▷ **Observation:** Computers are universal tools: their behavior is determined by a program; they can do anything, the program specifies.

▷ **Context:** Tools in most other disciplines are specific to particular tasks. (except in e.g. ribosomes in cell biology)

▷ *Remark 2.4.1 (Deep Fundamental Result).* There are things no computer can compute.

▷ **Example 2.4.2.** There cannot be a program that decides whether another program will terminate in finite time.

▷ *Remark 2.4.3 (Church-Turing Hypothesis).* There are two classes of languages

- ▷ Turing complete (or computationally universal) ones that can compute what is theoretically possible.
- ▷ data languages that cannot. (but describe data sets)

- **Observation 2.4.4 (Turing Equivalence).** *All programming languages are (made to be) universal, so they can compute exactly the same. (compilers/interpreters exist)*
- **. . . in particular . . . :** Everybody who tells you that one programming languages is the best has no idea what they're talking about (though differences in efficiency, convenience, and beauty exist)

Artificial Intelligence

- **Another Universal Tool:** The human mind. (We can understand/learn anything.)
- **Strong Artificial Intelligence:** claims that the brain is just another computer.
- **If that is true** then
  - the human mind underlies the same restrictions as computational machines
  - we may be able to find the "mind-program".

We now come to one of the most important, but maybe least acknowledged principles of programming languages: The principle of compositionality. To fully understand it, we need to fix some fundamental vocabulary.

Top Principle of Programming: Compositionality

- **Observation 2.4.5.** *Modern programming languages compose various primitives and give them a pleasing, concise, and uniform syntax.*
- **Question:** What does all of this even mean?
- **Definition 2.4.6.** In a programming language, a primitive is a "basic unit of processing", i.e. the simplest element that can be given a procedural meaning (its semantics) of its own.
- **Definition 2.4.7 (Compositionality).** All programming languages provide composition principles that allow to compose smaller program fragments into larger ones in such a way, that the semantics of the larger is determined by the semantics of the smaller ones and that of the composition principle employed.
- **Observation 2.4.8.** *The semantics of a programming language, is determined by the meaning of its primitives and composition principles.*
- **Definition 2.4.9.** Programming language syntax describes the surface form of the program: the admissible character sequences. It is also a composition of the syntax for the primitives.

All of this is very abstract – it has to be as we have not fixed a programming language yet and you will only understand the true impact of the compositionality principle over time and with programming experience. Let us now see what this means concretely for our course.

Consequences of Compositionality

▷ **Observation 2.4.10.** *To understand a programming language, we (only) have to understand its primitives, composition principles, and their syntax.*

▷ **Definition 2.4.11.** The "art of programming" consists of composing the primitives of a programming language.

▷ **Observation 2.4.12.** *We only need very few – about half a dozen – primitives to obtain a Turing complete programming language.*

▷ **Observation 2.4.13.** *The space of program behaviors we can achieve by programming is infinites large nonetheless.*

▷ *Remark 2.4.14.* More primitives make programming more convenient.

▷ *Remark 2.4.15.* Primitives in one language can be composed in others.

A note on Programming: Little vs. Large Languages

▷ **Observation 2.4.16.** *Most such concepts can be studied in isolations, and some can be given a syntax on their own. (standardization)*

▷ **Consequence:** If we understand the concepts and syntax of the sublanguages, then learning another programming language is relatively easy.

2.5 More about Python

After we have had some general thoughts about prorgramming in general, we can get back to concrete Python facilities and idoms. We will concentrate on those – there are lots and lots more – that are useful in IWGS.

2.5.1 Sequences and Iteration

We now come to a commonly used class of objects in Python: sequences, such as lists, sets, tuples, and ranges as well as dictionaries.

They are used for storing, accumulating, and accessing objects in various ways in programs. They all have in common, that they can be used for iteration, thus creating a uniform interface to similar functionality.

Lists in Python

▷ **Definition 2.5.1.** A list is a finite sequence of objects, its elements.

▷ In programming languages, lists are used for locally storing and passing around collections of objects.

▷ In Python lists can be written as a sequence of comma separated expressions between square brackets.

▷ **Definition 2.5.2.** We call [⟨⟨seq⟩⟩] the list constructor.

▷ **Example 2.5.3 (Three lists).** Elements can be of different types in Python

```
list1 = ['physics', 'chemistry', 1997, 2000];
list2 = [1, 2, 3, 4, 5 ];
list3 = ["a", "b", "c", "d"];
```

▷ **Example 2.5.4.** List elements can be accessed by specifying ranges

```
>>> list1[0]        >>> list1[−2]        >>> list2[1:4]
'physics'           1997                 [2, 3, 4]
```

▷ **Definition 2.5.5.** Selecting sublists by specifying start and/or end is called list slicing.

As ?? shows, Python treats counting in list accessors somewhat peculiarly. It starts counting with zero when counting from the front and with one when counting from the back.
But lists are not the only things in Python that can be accessed in the way shown in ??. Python introduces a special class of types the sequence types.

Sequences in Python

▷ **Definition 2.5.6.** Python has more types that behave just like lists, they are called sequence types.

▷ The most important sequence types for IWGS are lists, strings and ranges.

▷ **Definition 2.5.7.** A range is a finite sequence of numbers it can conveniently be constructed by the range function: range(⟨⟨start⟩⟩,⟨⟨stop⟩⟩,⟨⟨step⟩⟩) constructs a range from ⟨⟨start⟩⟩ (inclusive) to ⟨⟨stop⟩⟩ (exclusive) with step size ⟨⟨step⟩⟩.

▷ **Example 2.5.8.** Lists can be constructed from ranges:

```
>>> list(range(1,6,2))
[1,3,5]
```

range(1,6,2) makes a "range" from 1 to 6 with step 2, list makes it a list.

Ranges are useful, because they are easily and flexibly constructed for iteration (up next).
You may ask yourselves, why Python has a special data structure for ranges. The main reason is that we can treat them more efficiently than lists. Consider the range constructed by range(1,1000000000), i.e. the numbers between 1 and a billion. If we were to represent this as a list, then this would probably take most of the memory available on your laptop, even if we do not do anything with it. But in the range, Python only needs to actually create those parts of the range that it actually needs. Say we want to access the 1000th element, then the Python interpreter can just compute that as 1+1000 when it needs to (and free the memory when that

is no longer needed); in particular, the interpreter does even have to create all the intermediate elements.

Iterating over Sequences in Python

▷ **Definition 2.5.9.** A for loop iterates a program fragment over a sequence; we call the process iteration. Python uses the following general syntax:

```
for ⟪var⟫ in ⟪range⟫:
    ⟪body⟫
⟪othercode⟫
```

▷ **Example 2.5.10.** A range function makes an sequence over which we can iterate.

```
for x in range(0, 3):
    print ("we tell you",x,"time(s)")
```

▷ **Example 2.5.11.** Lists and strings can also act as sequences. (try it)

```
print("Let me reverse something for you!")
x = input("please type somegthing!")
for i in reversed(list(x)):
    print(i)
```

But lists are not the only data structure for collections of objects. Python provides others that are organized slightly differently for different applications. We give a particularly useful example here: dictionaries.

Python Dictionaries

▷ **Definition 2.5.12.** A dictionary is an unordered collection of ordered pairs (k,v), where we call k the key and v the value.

▷ In Python dictionaries are written with curly brackets, pairs are separated by commata, and the value is separated from the key by a colon.

▷ **Example 2.5.13.** Dictionaries can be used for various purposes,

```
painting = {                      dict_de_en = {          enum = {
    "artist": "Rembrandt",          "Maus": "mouse",        1: "copy",
    "title": "The Night Watch",     "Ast": "branch",        2: "paste",
    "year": 1642                    "Klavier": "piano"      3: "adapt"
}                                 }                       }
```

▷ Dictionaries and sequences can be nested, e.g. for a list of paintings.

Dictionaries give "keyed access" to collections of data: we can access a value via its key. In particular, we do not have to remember the position of a value in the collection.

Interacting with Dictionaries

▷ **Example 2.5.14 (Dictionary operations).**

- ▷ painting["title"] returns the value for the key "title" in the dictionary painting.
- ▷ painting["title"]="De Nachtwacht" changes the value for the key "title" to its original Dutch (or adds item "title": "De Nachtwacht")

▷ **Example 2.5.15 (Printing Keys and Values).**

keys	values	key/value pairs

`for x in thisdict.keys():`
`    print(x)`
	`for x in thisdict.values():`
	`    print(x)`
		`for x, y in thisdict.items():`
		`    print(x, y)`

▷ More dictionary commands:

- ▷ **if** ⟨⟨key⟩⟩ **in** ⟨⟨dict⟩⟩ checks whether ⟨⟨key⟩⟩ is a key in ⟨⟨dict⟩⟩.
- ▷ painting.pop("title") removes the "title" item from painting.

Note that thisdict.keys has a short form: we can just iterate over the keys using **for** x **in** thisdict:.

2.5.2 Input and Output

The next topic of our stroll through Python is one that is more practically useful than intrinsically interesting: file input/output. Together with the regular expressions this allows us to write programs that transform files.

Input/Output in Python

▷ **Recall:** The CPU communicates with the user through input devices like keyboards and output devices like the screen.

▷ Programming languages provide special instructions for this.

▷ In Python we have already seen

- ▷ input(⟨⟨prompt⟩⟩) for input from the keyboard, it returns a string.
- ▷ **print**(⟨⟨objects⟩⟩,sep=⟨⟨separator⟩⟩,end=⟨⟨endchar⟩⟩) for output to the screen.

▷ But computers also supply another object to input from and output to (up next)

We now fix some of the nomenclature surrounding files and file systems provided by most operating system. Most programming languages provide their own bindings that allow to manipulate files.

Secondary (Disk) Storage; Files, Folders, etc.

▷ **Definition 2.5.16.** A file is a resource for recording data in a storage device. File size is measured in bit.

▷ **Definition 2.5.17.** Files are identified by a file name which usually consists of a base name and an extension separated by a dot character.

Files are managed by a file system which organize them hierarchically into named folders and locate them by a path; a sequence of folder names. The file name and the path together fully identify a file.

▷ Some file systems restrict the characters allowed in the file name and/or lengths of the base name or extension.

▷ **Definition 2.5.18.** Once a file has been opened, the CPU can write to it and read from it. After use a file should be closed to protect it from accidental reads and writes.

Many operating systems use files as a primary computational metaphor, also treating other resources like files. This leads to an abstraction of files called streams, which encompass files as well as e.g. keyboards, printers, and the screen, which are seen as objects that can be read from (keyboards) and written to (e.g. screens). This practice allows flexible use of programs, e.g. redirecting a the (screen) output of a program to a file by simply changing the output stream.
Now we can come to the Python bindings for the file input/output operations. They are rather straightforward.

Disk Input/Output in Python

▷ **Definition 2.5.19.** Python uses file objects to encapsulate all file input/output functionality.

▷ In Python we have special instructions for dealing with files:

- ▹ open(⟪path⟫,⟪iospec⟫) returns a file object f; ⟪iospec⟫ is one of r (read only; the default), a (append $\widehat{=}$ write to the end), and r+ (read/write).
- ▹ f.read() reads the file represented by file object f into a string.
- ▹ f.readline() reads a single line from the file (including the newline character \n) otherwise returns the empty string ''.
- ▹ f.write(⟪str⟫) appends the string ⟪str⟫ to the end of f, returns the number of characters written.
- ▹ f.close() closes f to protect it from accidental reads and writes.

▷ **Example 2.5.20 (Duplicating the contents of a file).**

```
f = open('workfile','r+')
filecontents = f.read()
f.write(filecontents)
```

The only interesting thing is that we have to declare our intentions when we opening a file. This allows the file system to protect the files against unintended actions and also optimize the data transfer to the storage devices involved.
Let us now look at some examples to fortify our intuition about what we can do with files in practice.

Disk Input/Output in Python (continued)

▷ **Example 2.5.21 (Reading a file linewise).**

```
>>> f.readline()
'This is the first line of the file.\n'
>>> f.readline()
'Second line of the file\n'
>>> f.readline()
''
```

```
>>> for line in f:
... print(line, end='')
...
This is the first line of the file.
Second line of the file
```

▷ If you want to read all the lines of a file in a list you can also use list(f) or f.readlines().

▷ For reading a Python file we use the **import**(⟪basename⟫) instruction

- ▹ it searches for the file ⟪basename⟫.py, loads it, interprets it as Python code, and directly executes it.
- ▹ primarily used for loading Python libraries (additional functionality)
- ▹ also useful for loading Python-encoded data (e.g. dictionaries)

The code snippet on the right of ?? show that files can be iterated over using a for loop: the file object is implicitly converted into a sequences of strings via the readline method.

2.5.3 Functions and Libraries in Python

We now come to a general device for organizing and modularizing code provided by most programming languages, including Python. Like variables, functions give names to Python objects – here fragments of code – and thus make them reusable in other contexts.

Functions in Python (Introduction)

▷ **Observation:** Sometimes programming tasks are repetitive

```
print("Hello Peter, how are you today? How about some IWGS?")
print("Hello Roxana, how are you today? How about some IWGS?")
print("Hello Frodo, how are you today? How about some IWGS?)
...
```

▷ **Idea:** We can automate the repetitive part by functions.

▷ **Example 2.5.22.**We encapsultate the greeting functionality in a function:

```
def greet (who):
    print("Hello ",who," how are you today? How about some IWGS?")
greet("Peter")
greet("Roxana")
greet("Frodo")
greet(input ("Who are you?"))
...
```

and use it repeatedly.

▷ Functions can be a very powerful tool for structuring and documenting programs (if used correctly)

Functions in Python (Example)

▷ **Example 2.5.23 (Multilingual Greeting).** Given a value for lang

```
def greet (who):
    if lang == 'en' :
        print("Hello ",who," how are you today? How about some IWGS?")
    elif lang == 'de' :
        print("Sehr geehrter ",who,", wie geht's heute? Wie waere es mit IWGS?")
```

we can even localize (i.e. adapt to the language specified in lang) the greeting.

We can now make the intuitions above formal and give the exact Python syntax of functions.

Functions in Python (Definition)

▷ **Definition 2.5.24.** A Python function is defined by a code snippet of the form

```
def f (p_1,...,p_n):
    """docstring, what does this function do on parameters
        :param p_i: document arguments}
    """
    ⟨⟨body⟩⟩ # it can contain p_1, ..., p_n, and even f
    return ⟨⟨value⟩⟩ # value of the function call (e.g text or number)
⟨⟨morecode⟩⟩
```

- ▷ the indented part is called the body of f, (⚠ : whitespace matters in Python)
- ▷ the p_i are called parameters, and n the arity of f.

A function f can be called on arguments $a_1, \ldots, a_n$ by writing the expression $f(a_1, \ldots, a_n)$. This executes the body of f where the (formal) parameters p_i are replaced by the arguments a_i.

We now come to a peculiarity of an object-oriented language like Python: it treats types as first-class entities, which can be defined by the user – they are called classes then. We will not go into that here, since we will not need classes in IWGS, but have have to briefly talk about methods, which are essentially functions, but have a special notation.
Python provides two kinds of function-like facilities: regular functions as discussed above and methods, which come with Python classes. We will not attempt a presentation of object oriented programming and its particular implementation in Python this would be beyond the mandate of the IWGS course – but give a brief introduction that is sufficient to use methods.

Functions vs. Methods in Python

▷ There is another mechanism that is similar to functions in Python. (we briefly introduce it here to delineate)

- **Background:** Actually, the types from ?? are classes, ...
- **Definition 2.5.25.** In Python all values belong to a class, which provide special functions we call methods. Values are also called objects, to emphasise class aspects. Method application is written with dot notation: ⟨⟨obj⟩⟩.⟨⟨meth⟩⟩(⟨⟨args⟩⟩) corresponds to ⟨⟨meth⟩⟩(⟨⟨obj⟩⟩,⟨⟨args⟩⟩).
- **Example 2.5.26.** Finding the position of a substring

```
>>> s = 'This is a Python string' # s is an object of class 'str'
>>> type(s)
<class 'str'> # see, I told you so
>>> s.index('Python') # dot notation (index is a string method)
10
```

Functions vs. Methods in Python

- **Example 2.5.27 (Functions vs. Methods).**

```
>>> sorted('1376254') # no dots!
['1', '2', '3', '4', '5', '6', '7']
```

```
>>> mylist = [3, 1, 2]
>>> mylist.sort() # dot notation
>>> mylist
[1, 2, 3]
```

- **Intuition:** Only methods can change objects, functions return changed copies (of the objects they act on).

For the purposes of IWGS, it is sufficient to remember that methods are a special kind of functions that employ the dot notation. They are provided by the class of an object.
It is very natural to want to share successful and useful code with others, be it collaborators in a larger project or company, or the respective community at large. Given what we have learned so far this is easy to do: we write up the code in a (collection of) Python files, and make them available for download. Then others can simply load them via the **import** command.

Python Libraries

- **Idea:** Functions, classes, and methods are re usable, so why not package them up for others to use.
- **Definition 2.5.28.** A Python library is a Python file with a collection of functions, classes, and methods. It can be imported (i.e. loaded and interpreted as a Python program fragment) via the **import** command.
- There are ≥ 150.000 libraries for Python (≙ packages on `http://pypi.org`)
  - search for them at `http://pypi.org` (e.g. 815 packages for "music")
  - install them with pip install ⟨⟨packagename⟩⟩
  - look at how they were done (all have links to source code)

▷ maybe even contribute back (report issues, improve code, ...) (open source)

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 69 2025-02-06 SOME RIGHTS RESERVED

The Python community is an open source community, therefore many developers organize their code into libraries and license them under open source licenses.

Software repositories like PyPI (the Python Package Index) collect (references to) and make them for the package manager pip, a program that downloads Python libraries and installs them on the local machine where the **import** command can find them.

2.5.4 A Final word on Programming in IWGS

This leaves us with a final word on the way we will handle prgramming in this course: IWGS is not a programming course, and we expect you to pick up Python from the IWGS and web/book resources. So, recall:

For more information on Python

RTFM ($\widehat{=}$ "read the fine manuals")

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 70 2025-02-06 SOME RIGHTS RESERVED

Our very quick introduction to Python is intended to present the very basics of programming and get IWGS students off the ground, so that they can start using programs as tools for the humanities and social sciences.

But there is a lot more to the core functionality Python than our very quick introduction showed, and on top of that there is a wealth of specialized packages and libraries for almost all computational and practical needs.

2.6 Exercises

Problem 6.1 (Hello World)

Write an extended "Hello World Program" in a file called exthello.py. The program should print information about you. Specifically, the information should be:

```
Hello World!  I am <your name>.
This is my first exercise in IWGS.
```

Problem 6.2 (Variable Assignment and Output)

Write a program in Python that calculates the total number of seconds in a leap year, stores the result in a variable and then displays that to the user.

Problem 6.3 (Variable Reuse)

Programming often has efficiency as one of its goals. After all, why go through the trouble of telling a computer how to do something, if you could do it better and quicker yourself?

Write a program in Python that prints the string "supercalifragilisticexpialidocious" five times, but *without* typing the word five times yourself.

Problem 6.4 (Human Readable Time)

In programming, it is often the case that your program collects a lot of data from various sources. It then becomes essential to present this data in a way that the user (usually a human!)

can easily understand. For example, most humans don't know how long a longer timespan is if it is given only in seconds.

Write a program in Python that first initialises a variable seconds = 1234567. Then, the program should calculate and print how long this timespan is in days, hours, minutes and seconds instead of just seconds.

Problem 6.5 (String Presentation)

Keeping with the importance of well-presented information: You can use certain special symbols in strings to give them a better formatting when they are ultimately printed. For example, when you put "\n" into a string, instead of printing these symbols, the output switches to a *new line*.

Write a Python program that prints your favourite haiku (a poem with five syllables on the first line, seven on the second and five on the third) on three three lines, but using only *one* **print** statement.

P.S.: If you don't have a favourite haiku and can't think of one yourself, you can use this one:

My cow gives less milk,
now that it has been eaten,
by a fierce dragon.

Problem 6.6 (User Input)

One of the most important things to learn about a programming language is how to get input from the user in front of the screen. In Python, one way of doing this is the input instruction.

For example: if you write answer = input("Do you like sharks?"), this will print the message you gave ("Do you like sharks?"), wait for the user to submit a response and store it as a string in the variable answer when you run the program. You can then use it like any other value stored in a variable.

Write a simple program that prints a generic greeting message, then asks the user to input their name, stores the input in a variable and then finishes with a goodbye message that uses the name the user gave.

Problem 6.7 (Simple Branching)

The next important concept is control flow. A program that always does the same thing gets boring fast. We want to write programs that do different things under different circumstances. In Python, one way to do this are conditional statements.

Write a Python program that asks the user if they have a pet. If their answer was "yes", the program should ask what kind of pet they have. Since sloths are the cutest animals (at least for this exercise), the program should print "awww!" if the user's second answer was "sloth" and "cool!" if it was something else. If the user does not answer with "yes" the first time around, the program should quit with a goodbye message.

Problem 6.8 (Simple Looping)

Computers are very good at doing the same thing over and over again without complaining or messing up. Humans are not. In Python, we can use a loops if we want something done multiple times.

Suppose your boss wants the string "Programming **is** cool!" printed exactly 1337 times (for some reason ...). Typing up the string yourself takes about nine seconds each time, printing it in a loop takes no time.

To save time, write a Python program that prints the sentence "Programming **is** cool!" 1337 times using a loop. Your program should also keep track of (store in a variable) how much time the loop saved the programmer in total (9 seconds per iteration of the loop). Print this value after the loop finishes.

Problem 6.9 (Temperature Conversion)

Write two Python programs, named celsius2fahrenheit and fahrenheit2celsius, that given a number as input from the user convert it to the respective other temperature scale and print the result.

The conversion formulas are as follows:

$$[^\circ C] = ([^\circ F] - 32) \cdot \frac{5}{9} \qquad [^\circ F] = [^\circ C] \cdot \frac{9}{5} + 32$$

Remember that input will save the input as a string, not as a floatnumber. You can convert a string to a floatnumber using the function float.

Example: float("3.1415") will evaluate to the floatnumber 3.1415. If the text given to float does not actually represent a floatnumber (e.g. float("bad")), Python will throw an error.

Afterwards, please test your programs against another converter (easily found via your internet search engine of choice) to make sure that your functions produce the correct results.

Chapter 3

Numbers, Characters, and Strings

In our basic introduction to programming above we have convinced ourselves that we need some basic objects to compute with, e.g. Boolean values for conditionals, numbers to calculate with, and characters to form strings for input and output. In this chapter we will look at how these are represented in the computer, which in principle can only store binary digits voltage or no noltage on a wire – which we think of as 1 and 0.
In this chapter we look at the representation of the basic data structures of programming languages (numbers and characters) in the computer; Boolean values ("True" and "False") can directly be encoded as binary digits.

Documents as Digital Objects

- ▷ **Question:** how do texts get onto the computer?(after all, computers can only do 0/1)
- ▷ **Hint:** At the most basic level, texts are just sequences of characters.
- ▷ **Answer:** We have to encode characters as sequences of bits.
- ▷ **We will go into how:**
  - ▷ documents are represented as sequences of characters,
  - ▷ characters are represented as numbers,
  - ▷ numbers are represented as bits (0/1).

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 71 2025-02-06

3.1 Representing and Manipulating Numbers

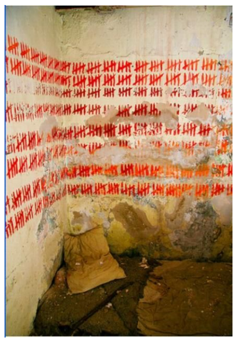
We start with the representation of numbers. There are multiple number systems, as we are interested in the principles only, we restrict ourselves to the natural numbers – all other number systems can be built on top of these. But even there we have choices about representation, which influence the space we need and how we compute with natural numbers.
The first system for number representations is very simple; so simple in fact that it has been discovered and used a long time ago.

Natural Numbers

▷ Numbers are symbolic representations of numeric quantities.

▷ There are many ways to represent numbers (more on this later)

▷ Let's take the simplest one (about 8,000 to 10,000 years old)



▷ We count by making marks on some surface.

▷ For instance //// stands for the number four (be it in 4 apples, or 4 worms)

In addition to manipulating normal objects directly linked to their daily survival, humans also invented the manipulation of place-holders or symbols. A *symbol* represents an object or a set of objects in an abstract way. The earliest examples for symbols are the cave paintings showing iconic silhouettes of animals like the famous ones of Cro-Magnon. The invention of symbols is not only an artistic, pleasurable "waste of time" for humans, but it had tremendous consequences. There is archaeological evidence that in ancient times, namely at least some 8000 to 10000 years ago, humans started to use tally bones for counting. This means that the symbol "bone with marks" was used to represent numbers. The important aspect is that this bone is a symbol that is completely detached from its original down to earth meaning, most likely of being a tool or a waste product from a meal. Instead it stands for a universal concept that can be applied to arbitrary objects.
So far so good, let us see how this would be represented on a computer:

Unary Natural Numbers on the Computer

▷ **Definition 3.1.1.** We call the representation of natural numbers by slashes on a surface the unary natural numbers.

▷ **Question:** How do we represent them on a computer? (not bones or walls)

▷ **Idea:** If we have a memory bank of n binary digits, initialize all by 0, represent each slash by a 1 from the right.

▷ **Example 3.1.2.** Memory bank with 32 binary digits, representing the number 11.

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1

▷ **Problem:** For realistic arithmetic we need better number representations than the unary natural numbers (e.g. for representing the number of EU citizens $\widehat{=}$ 100 000 pages of /)

The problem with the unary number system is that it uses enormous amounts of space, when writing down large numbers. We obviously need a better representation. The unary natural numbers are very simple and direct, but they are neither space-efficient, nor easy to manipulate. Therefore we will use different ways of representing numbers in practice.

Positional Number Systems

- **Problem:** Find a better representation system for natural numbers.
- **Idea:** Build a clever code on the unary natural numbers, use position information and addition, multiplication, and exponentiation.
- **Definition 3.1.3.** A positional number system $\mathcal{N}$ is a pair $\langle D, \varphi\rangle$ with
  - D is a finite set of b digits; $b := \#(D)$ is the base or radix of $\mathcal{N}$.
  - $\varphi\colon D \to [0,b-1]$ is bijective.

  We extend φ to a bijection between sequences $d_k, \ldots, d_0$ of digits and natural numbers by setting

  $$\varphi(d_k, \ldots, d_0) := \sum_{i=0}^{k} \varphi(d_i) \cdot b^i$$

  We say that the digit sequence $n_b := d_k, \ldots, d_0$ is the positional notation of a natural number n, iff $\varphi(d_k, \ldots, d_0) = n$.
- **Example 3.1.4.** $\langle\{a, b, c\}, \varphi\rangle$ with with $\varphi(a) := 0$, $\varphi(b):=1$, and $\varphi(c) := 2$ is a positional number system for base three. We have

  $$\varphi(c, a, b) = 2 \cdot 3^2 + 0 \cdot 3^1 + 1 \cdot 3^0 = 18 + 0 + 1 = 19$$

If we look at the unary number system from a greater distance, we see that we are not using a very important feature of strings here: position. As we only have one letter in our alphabet, we cannot, so we should use a larger alphabet. The main idea behind a positional number system $\mathcal{N} = \langle D_b, \varphi_b\rangle$ is that we encode numbers as strings of digits in D_b, such that the position matters, and to give these encodings a meaning by mapping them into the unary natural numbers via a mapping φ_b.

Commonly Used Positional Number Systems

- **Definition 3.1.5.** The following positional number systems are in common use.

name	set	base	digits	example
unary	$\mathbb{N}_1$	1	0	00000_1
binary	$\mathbb{N}_2$	2	0,1	0101000111_2
octal	$\mathbb{N}_8$	8	0,1,. . . ,7	63027_8
decimal	$\mathbb{N}_{10}$	10	0,1,. . . ,9	162098_{10} or 162098
hexadecimal	$\mathbb{N}_{16}$	16	0,1,. . . ,9,A,. . . ,F	$FF3A12_{16}$

Binary digits are also called bits, and a sequence of eight bits an octet.

▷ **Notation:** Attach the base of $\mathcal{N}$ to every number from $\mathcal{N}$. (default: decimal)

▷ **Trick:** Group triples or quadruples of binary digits into recognizable chunks (add leading zeros as needed)

▷ $110001101011100_2 = \underbrace{0110_2}_{6_{16}}\underbrace{0011_2}_{3_{16}}\underbrace{0101_2}_{5_{16}}\underbrace{1100_2}_{C_{16}} = 635C_{16}$

▷ $110001101011100_2 = \underbrace{110_2}_{6_8}\underbrace{001_2}_{1_8}\underbrace{101_2}_{5_8}\underbrace{011_2}_{3_8}\underbrace{100_2}_{4_8} = 61534_8$

▷ $F3A_{16} = \underbrace{F_{16}}_{1111_2}\underbrace{3_{16}}_{0011_2}\underbrace{A_{16}}_{1010_2} = 111100111010_2$, $4721_8 = \underbrace{4_8}_{100_2}\underbrace{7_8}_{111_2}\underbrace{2_8}_{010_2}\underbrace{1_8}_{001_2} = 100111010001_2$

We have all seen positional number systems: our decimal system is one (for the base 10). Other systems that important for us are the binary system (it is the smallest non degenerate one) and the octal (base 8) and hexadecimal (base 16) systems. These come from the fact that binary numbers are very hard for humans to scan. Therefore it became customary to group three or four digits together and introduce (compound) digits for these groups. The octal system is mostly relevant for historic reasons, the hexadecimal system is in widespread use as syntactic sugar for binary numbers, which form the basis for electronic circuits, since binary digits can be represented physically by voltage/no voltage.

Arithmetics in Positional Number Systems

▷ For arithmetic just follow the rules from elementary school (for the right base)

▷ Tom Lehrer's "New Math": `https://www.youtube.com/watch?v=DfCJgC2zezw`

▷ **Example 3.1.6.**

Addition base 4

$$\begin{array}{cccc} & 1 & 2 & 3 \\ + & 1_1 & 2_1 & 3 \\ \hline & 3 & 1 & 2 \end{array}$$

binary multiplication

$$\begin{array}{cccccc} & & 1 & 0 & 1 & 0 \\ & * & & 1 & 1 & 0 \\ \hline & & 0 & 0 & 0 & 0 \\ & 1 & 0 & 1 & 0 & \\ 1 & 0 & 1 & 0 & & \\ \hline 1 & 1 & 1 & 1 & 0 & 0 \end{array}$$

How to get back to Decimal (or any other system)

▷ **Observation:** ?? specifies how we can get from base b representations to decimal. We can always go back to the base b numbers.

▷ **Observation 3.1.7.** *To convert a decimal number n to base b, use successive integer division (division with remainder) by b:*

$i := n$; **repeat** (record $i \bmod b$, $i := i \operatorname{div} b$) **until** $i = 0$.

▷ **Example 3.1.8 (Convert 456 to base 8).** Result: 710_8

$$\begin{array}{ll} 456\,\mathrm{div}\,8 = 57 & 456\,\mathrm{mod}\,8 = 0 \\ 57\,\mathrm{div}\,8 = 7 & 57\,\mathrm{mod}\,8 = 1 \\ 7\,\mathrm{div}\,8 = 0 & 7\,\mathrm{mod}\,8 = 7 \end{array}$$

3.2 Characters and their Encodings: ASCII and UniCode

IT systems need to encode characters from our alphabets as bit strings (sequences of binary digits (bits) 0 and 1) for representation in computers. To understand the current state – the unicode standard – we will take a historical perspective. It is important to understand that encoding and decoding of characters is an activity that requires standardization in multi-device settings – be it sending a file to the printer or sending an e-mail to a friend on another continent. Concretely, the recipient wants to use the same character mapping for decoding the sequence of bits as the sender used for encoding them – otherwise the message is garbled.

We observe that we cannot just specify the encoding table in the transmitted document itself, (that information would have to be en/decoded with the other content), so we need to rely document-external external methods like standardization or encoding negotiation at the meta-level. In this section we will focus on the former.
The ASCII code we will introduce here is one of the first standardized and widely used character encodings for a complete alphabet. It is still widely used today. The code tries to strike a balance between being able to encode a large set of characters and the representational capabilities in the time of punch cards (see below).

The ASCII Character Code

▷ **Definition 3.2.1.** The American Standard Code for Information Interchange (ASCII) is a character encoding that assigns characters to numbers 0 127.

Code	⋯0	⋯1	⋯2	⋯3	⋯4	⋯5	⋯6	⋯7	⋯8	⋯9	⋯A	⋯B	⋯C	⋯D	⋯E	⋯F
0⋯	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1⋯	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2⋯		!	”	#	$	%	&	’	(	)	*	+	,	−	.	/
3⋯	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4⋯	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5⋯	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	ˆ	_
6⋯	‘	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7⋯	p	q	r	s	t	u	v	w	x	y	z	{	|	}	˜	DEL

▷ The first 32 characters are control characters for ASCII devices like printers.

▷ **Motivated by punch cards:** The character 0 (0000000_2 in binary) carries no information NUL, (used as dividers)
Character 127 ($\widehat{=}\ 1111111_2$) can be used for deleting (overwriting) last value (cannot delete holes)

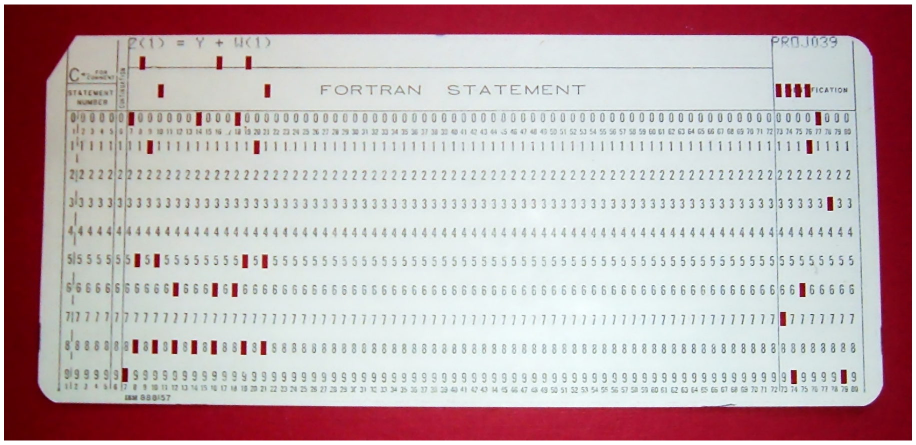
▷ The ASCII code was standardized in 1963 and is still prevalent in computers today. (but seen as US centric)

Punch cards were the preferred medium for long-term storage of programs up to the late 1970s, since they could directly be produced by card punchers and automatically read by computers.

A Punchcard

▷ **Definition 3.2.2.** A punch card is a piece of stiff paper that contains digital information represented by the presence or absence of holes in predefined positions.

▷ **Example 3.2.3.** This punch card encodes the `FORTRAN` statement Z(1) = Y + W(1)



Up to the 1970s, computers were batch machines, where the programmer delivered the program to the operator (a person behind a counter who fed the programs to the computer) and collected the printouts the next morning. Essentially, each punch card represented a single line (80 characters) of program code. Direct interaction with a computer is a relatively young mode of operation.
The ASCII code as above has a variety of problems, for instance that the control characters are mostly no longer in use, the code is lacking many characters of languages other than the English language it was developed for, and finally, it only uses seven bits, where an octet (eight bits) is the preferred unit in information technology. Therefore a whole zoo of extensions were introduced, which — due to the fact that there were so many of them — never quite solved the encoding problem.

Problems with ASCII encoding

▷ **Problem:** Many of the control characters are obsolete by now/ (e.g. `NUL`,`BEL`, or `DEL`)

▷ **Problem:** Many European characters are not represented. (e.g. è,ñ,ü,ß,. . .)

▷ **European ASCII Variants:** Exchange less-used characters for national ones.

▷ **Example 3.2.4 (German ASCII).** Remap e.g. [$\mapsto$ Ä,] $\mapsto$ Ü in German ASCII ("Apple][" comes out as "`Apple ÜÄ`")

▷ **Definition 3.2.5 (ISO-Latin (ISO/IEC 8859)).** 16 Extensions of ASCII to 8-bit (256 characters) ISO Latin 1 $\widehat{=}$ "Western European", ISO Latin 6 $\widehat{=}$ "Arabic", ISO Latin 7 $\widehat{=}$ "Greek". . .

▷ **Problem:** No cursive Arabic, Asian, African, Old Icelandic Runes, Math,. . .

▷ **Idea:** Do something totally different to include all the world's scripts: For a scalable architecture, separate

- ▷ what characters are available, and (character set)
- ▷ a mapping from bit strings to characters. (character encoding)

The goal of the UniCode standard is to cover all the worlds scripts (past, present, and future) and provide efficient encodings for them. The only scripts in regular use that are currently excluded are fictional scripts like the elvish scripts from the Lord of the Rings or Klingon scripts from the Star Trek series.
An important idea behind UniCode is to separate concerns between standardizing the character set — i.e. the set of encodable characters and the encoding itself.

Unicode and the Universal Character Set

▷ **Definition 3.2.6 (Twin Standards).** A scalable architecture for representing all the worlds writing systems:

- ▷ The universal character set (UCS) defined by the ISO/IEC 10646 International Standard, is a standard set of characters upon which many character encodings are based.
- ▷ The unicode standard defines a set of standard character encodings, rules for normalization, decomposition, collation, rendering and bidirectional display order.

▷ **Definition 3.2.7.** Each UCS character is identified by an unambiguous name and an natural number called its code point.

▷ The UCS has 1.1 million code points and nearly 100 000 characters.

▷ **Definition 3.2.8.** Most (non-Chinese) characters have code points in $[1,65536]$: the basic multilingual plane (BMP).

▷ **Definition 3.2.9 (Notation).** For code points in the (BMP), four hexadecimal digits are used, e.g. $\mathrm{U}+0058$ for the character $\mathrm{LATINCAPITALLETTERX}$;

Note that there is indeed an issue with space-efficient character encodings here. UniCode reserves space for 2^{32} (more than a million) characters to be able to handle future scripts. But just simply using 32 bits for every UniCode character would be extremely wasteful: UniCode-encoded versions of ASCII files would be four times as large.
Therefore UniCode allows multiple character encodings. UTF-32 is a simple 32-bit code that directly uses the code points in binary form. UTF-8 is optimized for western languages and coincides with the ASCII where they overlap. As a consequence, ASCII encoded texts can be decoded in UTF-8 without changes — but in the UTF-8 encoding, we can also address all other unicode characters (using multi-byte characters).

Character Encodings in Unicode

▷ **Definition 3.2.10.** A character encoding is a mapping from bit strings to UCS code points.

▷ **Idea:** Unicode supports multiple character encodings (but not character sets) for efficiency.

▷ **Definition 3.2.11 (Unicode Transformation Format).**

- ▷ UTF-8, 8-bit, variable width character encoding, which maximizes compatibility with ASCII.
- ▷ UTF-16, 16-bit, variable width character encoding (popular in Asia)
- ▷ UTF-32, a 32-bit, fixed width character encoding (as a fallback)

▷ **Definition 3.2.12.** The UTF-8 encoding follows the following schema:

Unicode	octet 1	octet 2	octet 3	octet 4
$\mathrm{U}+000000-\mathrm{U}+00007\mathrm{F}$	0xxxxxxx			
$\mathrm{U}+000080-\mathrm{U}+0007\mathrm{FF}$	110xxxxx	10xxxxxx		
$\mathrm{U}+000800-\mathrm{U}+00\mathrm{FFFF}$	1110xxxx	10xxxxxx	10xxxxxx	
$\mathrm{U}+010000-\mathrm{U}+10\mathrm{FFFF}$	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

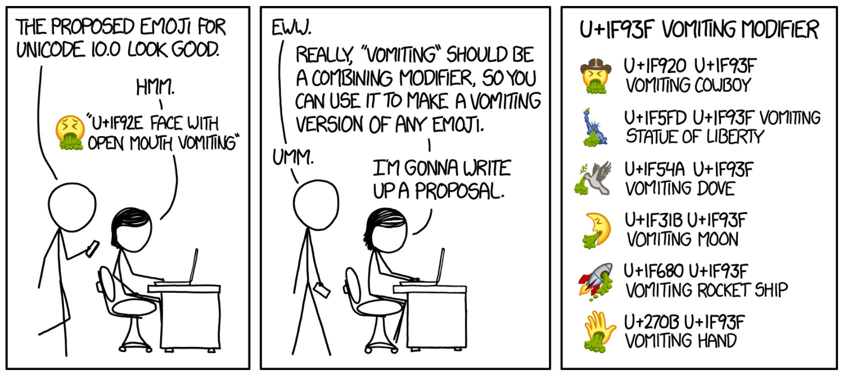
▷ **Example 3.2.13.** $\$ = \mathrm{U}+0024$ is encoded as 00100100 (1 byte)
¢ $= \mathrm{U}+00\mathrm{A}2$ is encoded as 11000010,10100010 (two bytes)
€ $= \mathrm{U}+20\mathrm{AC}$ is encoded as 11100010,10000010,10101100 (three bytes)

Note how the fixed bit prefixes in the UTF-8 encoding are engineered to determine which of the four cases apply, so that UTF-8 encoded documents can be safely decoded.

XKCD's Take on Recent Unicode Extensions

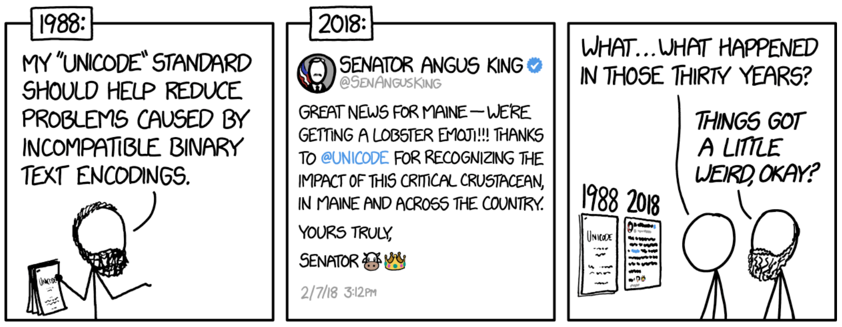
▷ UniCode 6.0 adopted hundreds of emoji characters in 2010 (2666 in July 2017)

▷ Modifying characters (`https://xkcd.com/1813/`)


THE PROPOSED EMOJI FOR UNICODE 10.0 LOOK GOOD.
HMM.
"U+1F92E FACE WITH OPEN MOUTH VOMITING"
EWW.
REALLY, "VOMITING" SHOULD BE A COMBINING MODIFIER, SO YOU CAN USE IT TO MAKE A VOMITING VERSION OF ANY EMOJI.
UMM.
I'M GONNA WRITE UP A PROPOSAL.
U+1F93F VOMITING MODIFIER
U+1F920 U+1F93F VOMITING COWBOY
U+1F5FD U+1F93F VOMITING STATUE OF LIBERTY
U+1F54A U+1F93F VOMITING DOVE
U+1F31B U+1F93F VOMITING MOON
U+1F680 U+1F93F VOMITING ROCKET SHIP
U+270B U+1F93F VOMITING HAND

XKCD's Take on Recent Unicode Extensions (cont.)

▷ Recent UniCode extensions (https://xkcd.com/1953/)


1988:
MY "UNICODE" STANDARD SHOULD HELP REDUCE PROBLEMS CAUSED BY INCOMPATIBLE BINARY TEXT ENCODINGS.
2018:
SENATOR ANGUS KING @SENANGUSKING
GREAT NEWS FOR MAINE—WE'RE GETTING A LOBSTER EMOJI!!! THANKS TO @UNICODE FOR RECOGNIZING THE IMPACT OF THIS CRITICAL CRUSTACEAN, IN MAINE AND ACROSS THE COUNTRY.
YOURS TRULY, SENATOR 🐮👑
2/7/18 3:12PM
WHAT...WHAT HAPPENED IN THOSE THIRTY YEARS?
THINGS GOT A LITTLE WEIRD, OKAY?
1988 2018

3.3 More on Computing with Strings

We now extend our repertoire on handling and formatting strings in Python: we will introduce string literals, which allow writing complex strings.

Playing with Strings and Characters in Python

▷ **Definition 3.3.1.** Python strings are sequences of UniCode characters.

▷ ⚠ In Python, characters are just strings of length 1.

▷ ord gives the UCS code point of the character, chr character for a number.

▷ **Example 3.3.2 (Playing with Characters).**

```
def lc(c) :
    return chr(ord(c) + 32)
def uc(c) :
    return chr(ord(c) − 32)
>>> uc('d')
'D'
>>> lc('D')
'd'
```

▷ Strings can be accessed by ranges $[i{:}j]$ ($[i] \mathrel{\hat{=}} [i{:}i]$)

▷ **Example 3.3.3.** Taking strings apart and re-assembling them.

```
def cap(s) :
    if s == "":
        return "" # base case
    else:
        return uc(s[0]) + cap(s[1:len(s)])

>>> cap('iwgs')
'IWGS'
```

?? may be difficult to understand at first. It is a programming technique called recursion, i.e.

functions that call themselves from within their body to solve problems by utilizing solutions to smaller instances of the same problem. Recursion can lead to very concise code, but requires some getting-used-to.

In ?? we define a function cap that given a string s returns a string that is constructed by combining the first character uppercased by the uc function with the result of calling the cap function on the rest string – s without the first character. The base case for the recursion is the empty string, where uc also returns the empty string. So let us see what happens in our test cap('iwgs'):
cap('iwgs') ⇝ uc('i')+cap('wgs') ⇝ 'I'+uc('w')+cap('gs') ⇝ 'I'+'W'+uc('g')+cap('s') ⇝ 'IW'+'G'+cap('s') ⇝ 'IWG'+uc('s')+cap('') ⇝ 'IWG'+'S'+cap('') ⇝ 'IWGS'+'' ⇝ 'IWGS'

⚠ ?? and ?? (or any other examples in this lecture) are not production code, but didactically motivated – to show you what you can do with the objects we are presenting in Python.

In particular, if we "lowercase" a character that is already lowercase – e.g. by lc('c'), then we get out of the range of the UCS code: the answer is \x83, which is the character with the hexadecimal code 83 (decimal 131), i.e. the character No Break Here.

In production code (as used e.g. in the Python lower method), we would have some range checks, etc.

String Literals in Python

▷ **Problem:** How to write strings including special characters?

▷ **Definition 3.3.4.** A literal is a notation for representing a fixed value for a data structure in source code.

▷ **Definition 3.3.5.** Python uses string literals, i.e character sequences surrounded by one, two, or three sets of matched single or double quotes for string input. The content can contain escape sequences, i.e. the escape character backslash followed by a code character for problematic characters:

Seq	Meaning	Seq	Meaning
\\	Backslash (\)	\'	Single quote (')
\"	Double quote (")	\a	Bell (BEL)
\b	Backspace (BS)	\f	Form-feed (FF)
\n	Linefeed (LF)	\r	Carriage Return (CR)
\t	Horizontal Tab (TAB)	\v	Vertical Tab (VT)

In triple-quoted string literals, unescaped newlines and quotes are honored, except that three unescaped quotes in a row terminate the literal.

Raw String Literals in Python

▷ **Definition 3.3.6.** Prefixing a string literal with a r or R turns it into a raw string literal, in which backslashes have no special meaning.

▷ **Note:** Using the backslash as an escape character forces us to escape it as well.

▷ **Example 3.3.7.** The string "a\nb\nc" has length five and three lines, but the string r"a\nb\nc" only has length seven and only one line.

Now that we understand the "theory" of encodings, let us work out how to program with them in Python:
Programming with UniCode strings is particularly simple, strings in Python are UTF-8-encoded UniCode strings and all operations on them are UniCode-based[1]. This makes the introduction to UniCode in Python very short, we only have to know how to produce non-ASCII characters, i.e. the characters that are not on regular keyboards.

If we know the code point, this is very simple: we just use UniCode escape sequences.

Unicode in Python

▷ *Remark 3.3.8.* The Python string data type is UniCode, encoded as UTF-8.

▷ **How to write UniCode characters?:** there are five ways

- ▷ write them in your editor (make sure that it uses UTF-8)
- ▷ otherwise use Python escape sequences (try it!)

```
>>> "\xa3" # Using 8−bit hex value
'\u00A3'
>>> "\u00A3" # Using a 16−bit hex value
'\u00A3'
>>> "\U000000A3" # Using a 32−bit hex value
'\u00A3'
>>> "\N{Pound Sign}" # character name
'\u00A3'
```

Note that the discussion about entry methods for unicode characters applies to the bare Python interpreter, not Python-specific text editor modes or user interfaces, which are often helpful by automatically replacing the input by the respective glyphs themselves.
String literals are convenient for creating simple string objects. For more complex ones, we usually want to build them from pieces, usually using the values of variables or the results of functions. This is what f strings are for in Python; we will cover that now.

Formatted String Literals (aka. f-strings)

▷ **Problem:** In a program we often want to build strings from pieces that we already have lying around interspersed by other strings.

▷ **Solution:** Use string concatenation:

```
>>> course="IWGS"
>>> students=6*11
>>> "The " + course + " course has " + str(students) + " students"
'The IWGS course has 66 students'
```

▷ We can do better! (mixing blanks and quotes is error-prone)

▷ **Definition 3.3.9.** Formatted string literals (aka. f strings) are string literals can contain Python expressions that will be evaluated – i.e. replaced with their values at runtime.

[1]Older programming languages have ASCII strings only, and UniCode strings are supplied by external libraries.

F strings are prefixed by f or F, the expressions are delimited by curly braces, and the characters { and } themselves are represented by {{ and }}.

▷ **Example 3.3.10 (An f-String for IWGS).**

```
>>> course="IWGS"
>>> f"The {course} course has {6*11} students"
'The IWGS course has 66 students'
```

F-String Example with a Dictionary

▷ **Example 3.3.11 (An F-String with a Dictionary).**

```
>>> course = {'name':"IWGS",'students':'66'}
>>> f"The {course['name']} course has {course['students']} students."
'The IWGS course has 66 students.'
```

Note that we alternated the quotes here to avoid the following problems:

```
>>> f'The course {course['name']} has {course['students']} students.'
  File "<stdin>", line 1
        f'The course {course['name']} has {course['students']} students.'
                             ^
SyntaxError: invalid syntax
```

3.4 More on Functions in Python

We now extend our repertoire of dealing with functions in Python.

In a sense, we now know all we have to about Python function: we can define them and apply them to arguments. But Python offers us much more: Python

- treats functions as “first-class objects”, i.e. entities that can be given to other functions as arguments, and can be returned as results.
- provides more ways of passing arguments to a function than the rather rigid way we have seen above. This can be very convenient and make code more readable.

We will cover these features now. The main motivation for this is that they are widely used in programming and being able to read them is important for collaborating with experienced programmers and reading existing code.
We digress to the internals of functions that make them even more powerful. It turns out that we do not have to give a function a name at all.

Anonymous Functions (lambda)

▷ **Observation 3.4.1.** *A Python function definition combines making a function object with giving it a name.*

▷ **Definition 3.4.2.** Python also allows to make anonymous functions via the function literal lambda for function objects:

lambda p_1,. . .,p_n: ⟪expr⟫

▷ **Example 3.4.3.** The following two Python fragments are equivalent:

```
def cube (x):          cube = lambda x: x*x*x
    x*x*x
```

The right one is just a variable assignment that assigns a function object to the variable cube. (In fact Python uses the right one internally)

▷ **Question:** Why use anonymous functions?

▷ **Answer:** We may not want to invent (i.e. waste) a name if the function is only used once. (examples on the next slide)

Anonymous functions do not seem like a big deal at first, but having a way to construct a function that can be used in any expression, is very powerful as we will see now.

Higher-Order Functions in Python

▷ **Definition 3.4.4.** We call a function a higher order function, iff it takes a function as argument.

▷ **Definition 3.4.5.** map and filter are built-in higher order functions in Python. They take a function and a list as arguments.

- ▷ map(f,L) returns the list of f-values of the elements of L.
- ▷ filter(p,L) returns the sub-list L' of those l in L, such that $p(l)$=True.

▷ **Example 3.4.6.** Mapping over and filtering a list

```
>>> li = [5, 7, 22, 97, 54, 62, 77, 23, 73, 61]
>>> list(map(lambda x: x*2 , li))
[10, 14, 44, 194, 108, 124, 154, 46, 146, 122]
>>> list(filter(lambda x: (x%2 != 0) , li))
[5, 7, 97, 77, 23, 73, 61]
```

Admittedly, in our example, we could also have defined a named function twice and then mapped that over li:

```
def twice (x):
    x*x
map twice li
```

But the code from **??** is more compact. Once we get used to the programming idiom and understand it, it becomes quite readable.

Another important feature of Python functions is flexible argument passing. This allows to define functions that supply complex behaviors – for which we need to set many parameters but simple calling patterns – which is good to hide complexity from the programmer.

The first argument passing feature we want to discuss is the use of keyword arguments, which gets around the problem of having to remember the position of an argument of a multi-parameter function.

Argument Passing in Python: Keyword Arguments

▷ **Definition 3.4.7.** The last $k \leq n$ of n parameters of a function can be keyword arguments of the form p_i=$\langle\langle val\rangle\rangle_i$: If no argument a_i is given in the function call, the default value $\langle\langle val\rangle\rangle_i$ is taken.

▷ **Example 3.4.8.** The head of the open function is

```
def open(file, mode='r', buffering=−1, encoding=None, errors=None,
                newline=None, closefd=True, opener=None)
```

Even if we only call it with open("foo"), we can use parameters like mode or opener in the body; they have the corresponding default value.

We can also give more arguments via keywords, even out of order

```
open("foo", buffering=1, mode="+a")
```

BTW: The opener argument of open is a function, and often an anonymous function is used if it is specified.

The next feature is dual to the last: instead of letting the caller leave out some arguments, we allow the caller more, which is then bound to a list parameter.

Argument Passing in Python: Flexible Arity

▷ **Definition 3.4.9.** Python functions can take a variable number of arguments: def f $(p_1,\ldots,p_k,*r)$ allows $n \geq k$ arguments, e. g. $f(a_1,\ldots,a_k,a_{k+1},\ldots,a_n)$ and binds the parameter r the rest argument to the list $[a_{k+1},\ldots,a_n]$.

▷ **Example 3.4.10.** A somewhat construed function that reports the number of extra arguments

```
def flexary (a,b,*c):
    return len(c)
>>> flexary (1,2,3,4,5)
>>> 3
```

▷ **Definition 3.4.11.** The star operator unpacks a list into an argument sequence.

▷ **Example 3.4.12 (Passing a starred list).**

```
def test(arg1, arg2, arg3):
     ...
args = ["two", 3]
test(1, *args)
```

Actually the star operator can be used in other situations as well, consider for instance

```
>>> numbers = [2, 1, 3, 4, 7]
>>> more_numbers = [*numbers, 11, 18]
>>> print(*more_numbers, sep=', ')
2, 1, 3, 4, 7, 11, 18
```

Here we have used the star operator twice: First to pass the list numbers as arguments to the list constructor and a second time to pass the extended list more_numbers to the print function.
Finally, we can combine the ideas from the last two to make keyword arguments flexary.

Argument Passing in Python: Flexible Keyword Arguments

▷ **Definition 3.4.13.** Python functions can take keyword arguments:
if k is a sequence of key/value pairs then def $f(p_1,\ldots,p_n,**k)$ binds the keys to values in the body of f.

▷ **Example 3.4.14.**

```
def kw_args(farg, **kwargs):
    print (f"formal arg: {farg}")
    for key in kwargs :
        print (f"another keyword arg: {key}: {kwargs[key]}")
>>> kw_args(1, myarg2="two", myarg3=3)
formal arg: 1
another keyword arg: myarg2 : two
another keyword arg: myarg3 : 3
```

Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 95 2025-02-06

Just as for the flexible arity case above, we have an operator that unpacks argument structures, here a dictionary.

Argument Passing in Python: Flexible Keyword Arguments (cont.)

▷ **Definition 3.4.15.**3 The double star operator unpacks a dictionary into a sequence of keyword arguments.

▷ **Example 3.4.16 (Passing around dates as dictionaries).**

```
date_info = {'day': "01", 'month': "01", 'year': "2020"}
def filename (year='2019',month=1,day=1)
    f"{year}—{month}—{day}.txt"
>>> filename(**date_info)
'2020—01—01.txt'
```

▷ **Example 3.4.17 (Mixing formal and keyword arguments).**

```
def pdict(a1, a2, a3):
    print('a1: ',a1,', a2: ',a2,', a3: ',a3)
dict = {"a3": 3, "a2": "two"}
>>> pdict(1, **dict)
>>> a1: 1, a2: two, a3: 3
```

Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 96 2025-02-06

Disclaimer: The last couple of features of Python functions are a bit more advanced than would usually be expected from a Python programming introduction in a course such as IWGS. But one of the goals of IWGS is to empower students to be able to read Python code of more experienced authors. And that kind of code may very well contain these features, so we need to cover them in IWGS.

So the last couple of slides should be considered as an "early exposure for understanding" rather than "essential to know for IWGS" content.

3.5 Regular Expressions: Patterns in Strings

Now we can come to the main topic of this section: regular expressions, A domain-specific language for describing string patterns. Regular expressions are extremely useful, but also quite cryptical at first. They should be understood as a powerful tool, that relies on a language with a very limited vocabulary. It is more important to understand what this tool can do and how it works in principle than memorizing the vocabulary – that can be looked up on demand.

Problem: Text/Data File Manipulation

- ▷ **Problem 1 (Information Extraction):** We often want to extract information from large document collections, e.g.
  - ▷ e-mail addresses or dates from collected correspondencesrtts
  - ▷ dates and places from newsfeeds
  - ▷ links from web pages
- ▷ **Problem 2 (Data Cleaning):** The representation in data files is often too noisy and inconsistent for directly importing into an application; e.g.
  - ▷ standardizing different spellings of e.g. city names, (Nuremberg vs. Nürnberg)
  - ▷ eliminating higher UniCode characters, when the application only accepts ASCII,
  - ▷ separating structured texts into data blocks. (e.g. in x-separated lists)
- ▷ **Enabling Technology:** Specifying text/data fragments $\rightsquigarrow$ regular expressions.

There are several dialects of regular expression languages that differ in details, but share the general setup and syntax. Here we introduce the Python variant and recommend [PyRegex] for a cheat-sheet on Python regular expressions (and an integrated regular expression tester).

Regular Expressions, see [Pyt]

- ▷ **Definition 3.5.1.** A regular expression (also called regular expression) is a formal expression that specifies a set of strings.
- ▷ **Definition 3.5.2 (Meta-Characters for Regexps).**

char	denotes
.	any single character (except a newline)
^	beginning of a string
$	end of a string
[. . .]/[^. . .]	any single character in/not in the brackets
[x−y]/[^x−y]	any single character in/not in range x to y
(. . .)	marks a capture group
\\n	the n^{th} captured group
|	disjunction
∗	matches preceding element zero or more times
+	matches preceding element one or more times
?	matches preceding element zero or one times
{n,m}	matches the preceding element between n and m times
\S/\s	non-/whitespace character
\W/\w	non-/word character
\D/\d	non-/digit (not only 0-9, but also e.g. arabic digits)

All other characters match themselves, to match e.g. a ?, escape with a \: \\?.

Let us now fortify our intuition with some (simple) examples and a more complex one.

Regular Expression Examples

▷ **Example 3.5.3 (Regular Expressions and their Values).**

regexp	values
car	car
.at	cat, hat, mat, . . .
[hc]at	cat, hat
[^c]at	hat, mat, . . . (but not cat)
^[hc]at	hat, cat, but only at the beginning of the line
[0−9]	Digits
[1−9][0−9]∗	natural numbers
(.∗)\1	mama, papa, wakawaka
cat|dog	cat, dog

▷ A regular expression can be interpreted by a regular expression processor (a program that identifies parts that match the provided specification) or a compiled by a parser generator.

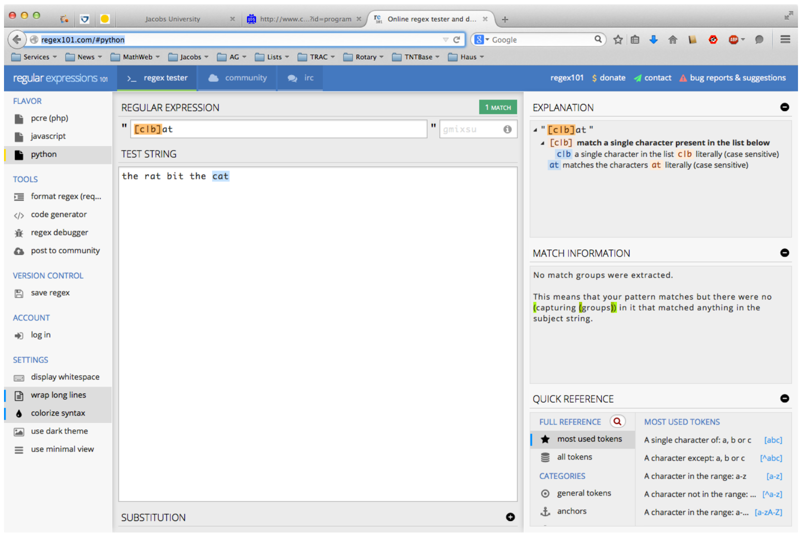
▷ **Example 3.5.4 (A more complex example).** The following regular expression matches times in a variety of formats, such as 10:22am, 21:10, 08h55, and 7.15 pm.

```
^(?:([0]?\d|1[012])|(?:1[3−9]|2[0−3]))[.:h]?[0−5]\d(?:\s?(?(1)(am|AM|pm|PM)))?$
```

As we have seen regular expressions can become quite cryptic and long (cf. e.g. ??), so we need help in developing them. One way is to use one of the many regexp testers online

Playing with Regular Expressions

▷ If you want to play with regular expressions, go e.g. to `http://regex101.com`



REGULAR EXPRESSION 1 MATCH

```
" [c|b]at " gmixsu
```

TEST STRING

```
the rat bit the cat
```

EXPLANATION

" [c|b]at "
[c|b] **match a single character present in the list below**
c|b a single character in the list c|b literally (case sensitive)
at matches the characters at literally (case sensitive)

MATCH INFORMATION

No match groups were extracted.

This means that your pattern matches but there were no (capturing (groups)) in it that matched anything in the subject string.

QUICK REFERENCE

FULL REFERENCE	MOST USED TOKENS	
most used tokens	A single character of: a, b or c	[abc]
all tokens	A character except: a, b or c	[^abc]
CATEGORIES	A character in the range: a-z	[a-z]
general tokens	A character not in the range: ...	[^a-z]
anchors	A character in the range: a-...	[a-zA-Z]

After covering regular expressions in the abstract, we will see how they are integrated into programming languages to solve problems. Of course we take Python as an example.

Regular Expressions in Python

- We can use regular expressions directly in Python by importing the re module (just add **import** re at the beginning)
- As Python has UniCode strings, regular expressions support UniCode as well.
- Useful Python functions that use regular expressions.
  - re.findall(⟨⟨pat⟩⟩,⟨⟨str⟩⟩): Return a list of non-overlapping matches of ⟨⟨pat⟩⟩ in ⟨⟨str⟩⟩.

    ```
    >>> re.findall(r"[h|c|r]at",'the cat ate the rat on the mat')
    ['cat','rat']
    ```

  - re.sub(⟨⟨pat⟩⟩,⟨⟨sub⟩⟩,⟨⟨str⟩⟩): Replace substrings that match ⟨⟨pat⟩⟩ in ⟨⟨str⟩⟩ by ⟨⟨sub⟩⟩.

    ```
    >>> re.sub(r'\sAND|and\s', '  ', 'Baked Beans and Spam')'Baked Beans  Spam'
    ```

  - re.split(⟨⟨pat⟩⟩,⟨⟨str⟩⟩): Split ⟨⟨str⟩⟩ into substrings that match *pmetavarpat*.

    ```
    >>> re.split(r'\s+','When shall we three meet again?'))
    ['When','shall','we','three','meet','again?']
    >>> re.split(r'\s+|\?|\.|!|,|:|;|','When shall we three meet again?'))
    ['When','shall','we','three','meet','again']
    ```

As regular expressions form a special language for describing sets of strings, it is not surprising that they are used in all kinds of searching, splitting, and substring replacement operations. As the language of regular expressions is well standardized, these more or less work the same in all programming languages, so what you learn for Python, you can re-use in other languages.

We will now see what we can do with regular expressions in a practical example. You should consider it as a "code reading/understanding" exercise, not think of it as something you should (easily) be able to do yourself. But **??** could serve as a quarry of ideas for things you can do to texts with regular expressions.

Example: Correcting and Anonymizing Documents

▷ **Example 3.5.5 (Document Cleanup).**

We write a function that makes simple corrections on documents and also crosses out all names to anonymize.

- ▷ *The worst president of the US,arguably was George W. Bush, right?*
- ▷ *However,are you famILIar with Paul Erdős or Henri Poincaré?* (Unicode)

Here is the function

- ▷ we import the regular expressions library and start the function

```
import re
def corranon (s)
```

- ▷ we first add blanks after commata

```
    s = re.sub(r",(\S)", r", \1", s)
```

- ▷ capitalize the first letter of a new sentence,

```
    s = re.sub(r"([\.\?!])\w*(\S)",
                 lambda m:m.group(1),r" ".upper()+m.group(2),
                 s)
```

This program is just a series of stepwise regular expression computations that are assigned to the variable s. For the last one, we use the **lambda** operator that constructs a function as an argument (the second) to re.sub. We use the anonymous functions because this function is only used once. This worked well, so we just continue along these lines.

Example: Correcting and Anonymizing Documents (cont.)

▷ **Example 3.5.6 (Document Cleanup (continued)).**

- ▷ next we make abbreviations for regular expressions to save space

```
    c = "[A–Z]"
    l = "[a–z]"
```

- ▷ remove capital letters in the middle of words

```
    s = re.sub(f"({l})({c}+)({l})",
                lambda m:f"{m.group(1)}{m.group(2).lower()}{m.group(3)}",
                s) #
```

- ▷ and we cross-out for official public versions of government documents,

```
    s = re.sub(f"({c}{l}+ ({c}{l}*(\.?) )?{c}{l}+)", #
                lambda m:re.sub("\S", "X", m.group(1)),
                s)
```

▷ finally, we return the result

```
    s
```

The worst president of the US,arguably was George W. Bush, right?
becomes
The worst president of the US, arguably was XXXXXX XX XXXX, right?

We show the whole program again, to see that it is relatively small (thanks to the very compact – if cryptic – regular expressions), when we leave out all the comments.

Example: Correcting and Anonymizing Documents (all)

▷ **Example 3.5.7 (Document Cleanup (overview)).**

```
import re
def corranon (s)
    s = re.sub(r",(\S)", r", \1", s)
    s = re.sub(r"([\.\?!])\w*(\S)",
               lambda m:m.group(1),r" ".upper()+m.group(2),
               s)
    c = "[A−Z]"
    l = "[a−z]"
    s = re.sub(f"({l})({c}+)({l})",
               lambda m:f"{m.group(1)}{m.group(2).lower()}{m.group(3)}",
               s) #
    s = re.sub(f"({c}{l}+ ({c}{l}*(\.?) )?{c}{l}+)", #
               lambda m:re.sub("\S", "X", m.group(1)),
               s)
    s
```

3.6 Exercises

Problem 6.1 (Basic Lists)

When working with lists, the first and the last elements of the list are often of special interest or significance.

1. Write a Python function that, when given a list as an argument, prints (on two separate lines, with some explanatory text) the first and last elements of the list.

2. Is it possible to do this without iterate over the entire list to find the last element?

3. What happens when you give this function a list of only one element?

4. What happens when you give it the empty list?

Problem 6.2 (User Input II)

Often, when you are taking input from the user, it becomes important that the input is one of a certain set of "acceptable" answers.

Write a Python program that asks the user for their favourite deadly sin. If the input it receives is not one of the acceptable answers (i.e. the strings `"lust"`, `"gluttony"`, `"greed"`, `"sloth"`, `"wrath"`, `"envy"` and `"pride"`), it should keep asking again and again.

When the input is (finally) correct, it should print a message either complimenting or deriding the user on their pick (your choice!).

Problem 6.3 (Dictionaries)

In programming, it is important to gain familiarity with the most commonly used data structures. This exercise will make you more familiar with the dictionary data structure.

1. Write a Python dictionary that associates names of famous peoples (i.e. strings as keys) with their year of birth (i.e. `int`s as values). The entries can be real or fictional people, as long as they have a clear year of birth.

2. Write a program that finds the oldest person (i.e. lowest year of birth) in that dictionary. (How) can you iterate over all keys of a dictionary? Finally, your program should print in what year the oldest person in your dictionary was born (it does not have to say who that person is).

Problem 6.4 (Egyptian Hieroglyphs 1: Numerals)

Programming is a versatile discipline and applicable to a lot of very different fields, from space satelites to fast pizza delivery to Egyptian hieroglyphs. In the following exercises, you will take a closer look at the latter to familiarise yourselves with the unicode character encoding.

The Egyptian numeral system[2] is decimal, like our system, but is not position-based (similar to Roman numerals). Each hieroglyph has a certain unicode encoding[3], i.e. a certain number that people have agreed upon to represent a certain hieroglyph.

The Egyptian number system is relatively simple (for number up to $1,000,000$ or so). Learn about it. Then, write a Python function `arabic2Egyptian` that takes a standard (positive) integer and returns a unicode string of a corresponging Egyptian number.

Note: The code here will be structurally similar to a previous exercise. Also recall that the Universal Character Set assigns every character a hexadecimal number n, e.g. $1F607$ (smiling face with halo). If we want to use character n in a string in Python, just use "\U0001F607" (i.e. n filled up with leading zeros to make it 8 hex digits).

Note: Note that we will *not* be awarding / deducting points on precise hieroglyph choice. As long as the hieroglyphs you chose roughly align with those presented in the number systems article, we will assume them correct. This goes for all exercises on this sheet.

[2]See, for example: `https://en.wikipedia.org/wiki/Egyptian_numerals`

[3]See `https://en.wikipedia.org/wiki/Egyptian_Hieroglyphs_(Unicode_block)` for details

Problem 6.5 (Character Encodings)
Briefly introduce and discuss the relative merits of

1. the ASCII code,
2. the ISO Latin codes,
3. the universal character set, and
4. the unicode encodings UTF-8, UTF-16, and UTF-32

Problem 6.6 (Egyptian Hieroglyphs 2: Text)
Supose that word has gotten around that you know how to handle unicode in Python and one of your friends who is also an egyptology enthusiast wants your help.

The standard method of displaying Egyptian hieroglyphs (etched into stone or clay) can be slow in writing and just remembering longer messages can be hard to do[4]. A digital format would be so much simpler!

First, write a Python dictionary that associates English or German words (keys) to fitting unicode symbols (values). Your dictionary obviously does not need to translate *all* hieroglyphs, but should at least inlcude five different ones.

Second, write a program that, using this dictionary, will ask the user again and again for input, looks up the value associated with that input in your dictionary and appends it to a string variable. When some special phrase to end the program is entered (e.g. `"exit"` or `"quit"`), the program should print the variable and exit.

This way, you can take a message that's easy to write on a Western keyboard and easily turn it into proper Egyptian hieroglyphs.

Problem 6.7 (Egyptian Hieroglyphs 3: Input Sanitising)
Whenever you ask a user for input that you want to use in a meaningful way later in your program, it is vital that you make sure the user has actually entered something sensible. Because often, they won't.

Concretely, if you look up a key in a dictionary that was never assigned a value, Python will print an error message and your program will crash.

Amend your program from the previous exercise to check if the entered string is actually a key in the dictionary you are using. If it isn't, you can print an error message or simply ask again. Entering garbage should no longer crash your program.

Problem 6.8 (Basechange)
Colours are important for a plethora of things in software development and there are many ways of describing just which colour you are talking about.

Maybe the most common way to specify a colour is by giving a triple of numbers between 0 and 255, signifying the how strong the red, green and blue (RGB) components in the colour are. Often, these are given as hexadecimal values (i.e. `00` to `FF`).

First, make sure that you understand how a hexadecimal number system works. Then, write a function that takes a string as an argument. This string will only have one (hexadecimal) character, either of the following:

`["0","1","2","3","4","5","6","7","8","9","A","B","C","D","E","F"]`.

The function should return the decimal value of the input as a regular integer.

Then, using the function you just finished, write a program that takes strings of six hexadecimal characters (two for red, two for green and two for blue, in that order, e.g. `00FF88` or `326496`) and prints their correct RGB components in decimal.

[4] Compare “Ente, Auge, Zickzack” (ZDF, German): `https://www.youtube.com/watch?v=SbZXiDE6GO4`

Problem 6.9 (Regular Expressions 1)

In this exercise we will explore regular expression. Regular expressions allow us to find patterns in a given text and even modify the matched substrings. To use regular expressions, you need to import the “re” library. This is done by typing “**import** re” at the top of your Python file.

In the imperial unit system, mass is measured in pounds (lb). As Central Europeans are more used to expressing mass in kilograms (kg), we will use regular expressions to find occurences of mass measurements in a text and convert it.

Consider the following text[5]:

> Two-thirds of Americans report that their actual weight is more than their ideal weight, although for many, the difference between actual and ideal is only 10 pounds or less. But 30% of women and 18% of men say their current weight is more than 20 pounds more than their ideal weight. The average American today weighs 17 pounds above what he or she considers to be ideal, with women reporting a bigger difference between actual and ideal than men.

Use regular expressions to find all NonNegativeRealNumbers in the text. Use the re.findall() function[6], which returns a list of matches.

Take into consideration, that numbers can consist of more than one digit. Print the list of matches. Amend the program, such that it only matches occurences of pound measurements, i.e. only numbers followed by the string "pounds". The list for the above text should now be ["10 pounds", "20 pounds", "17 pounds"].

In regular expression, you can group certain parts of the pattern by enclosing it in parentheses. This can be useful, if you want to further process the results of the matching.

Amend your program, such that findall() returns the following list: ["10", "20", "17"]. Note that these are still only the numbers followed by "pounds", but the "pounds"-part is stripped away automatically.

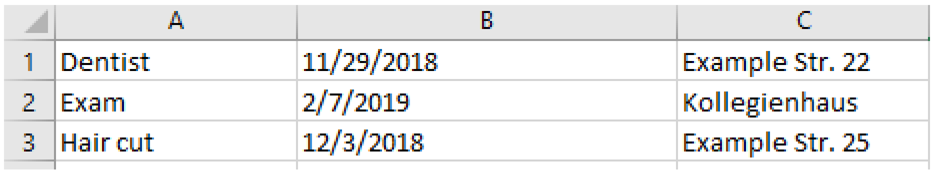
Iterate over your list of measurements. For each entry, convert the entry to kilograms using the following formula:

$$[kg] = [lb]/2.2046$$

Print the conversion with some explanatory text, i.e. "10lb are 4.535970244035199kg".

Problem 6.10 (Regular Expressions 2)

In the real world, data processed by computers often comes from files read from the hard disk. Consider the following spreadsheet table:



	A	B	C
1	Dentist	11/29/2018	Example Str. 22
2	Exam	2/7/2019	Kollegienhaus
3	Hair cut	12/3/2018	Example Str. 25

It lists appointments line by line. Each line consists of the type of appointment, the date and the place. A common data format is the CSV file format. Most spreadsheets (like OO Calc or MS Excel) support exporting to this format.

The resulting CSV file (also supplied for this exercise) looks like this:

```
Dentist;11/29/2018;Example Str. 22
Exam;7/2/2019;Kollegienhaus
Hair cut;12/3/2018;Example Str. 25
```

[5]Source: https://news.gallup.com/poll/102919/average-american-weighs-pounds-more-than-ideal.aspx

[6]https://docs.python.org/3/library/re.html#re.findall

CSV is short for “Comma Separated Values”. As the name implies it lists the entries, separated by commata (actually it’s semicolons in this case).

The dates in this example are given in the American notation: Month/Day/Year. We will use regular expressions to convert it into German notation: Day.Month.Year, i.e. day before month and separated by dots instead of slashes.

Open the file using Python’s File I/O (input/output) functionality[7]. Read the whole file using the readlines() function, which returns a list of lines. Print this list.

Now iterate over the list and perform the following for each entry: Use the string split() method[8] to separate individual entries at the semicolons.

For example, splitting the entry "Dentist;11/29/2018;Example Str. 22" at the semicolons should give you the list ["Dentist", "11/29/2018", "Example Str. 22"].

The second value is the date we would like to convert. Use the re.sub() function[9] to extract the day, month and year and reassemble them in the German notation. Afterward print some useful text for the appointment containing the converted date.

Problem 6.11 (Regular Expressions 3)

One of the best uses of a computer’s enormous processing power is to have it filter quickly through large amounts of data that would otherwise take a human a long time to sift through. This is also often a task where regular expressions shine.

Along with this exercise, you will be supplied with a text file that contains the entire text of Lev Tolstoy’s “War and peace”[10], slightly modified.[11] This will serve as our “corpus data” for this exercise.

Somewhere in this text (more than 500.000 words), you know that there are a few e-mail adresses and a few hexadecimal colour codes (in a format like the following: #10FFAA). Write a Python program that reads the file and uses regular expressions to find these addresses and colour codes. Afterwards, display the result with some explanatory text.

Note: Simply searching for "#" or "@" will not help you here, because since the data is sadly a bit “degraded”, those characters are also interspersed a few hundred times at random intervals.

---

[7]If you need a refresher about file input/output, see: https://www.pythonforbeginners.com/cheatsheet/python-file-handling

[8]https://docs.python.org/3/library/stdtypes.html#str.split

[9]https://docs.python.org/3/library/re.html#re.sub

[10]As found on Project Gutenberg: https://www.gutenberg.org (currently not accessible from Germany due to copyright disputes)

[11]Found here: https://kwarc.info/teaching/IWGS/materials/war-and-peace_modified.txt

Chapter 4

Documents as Digital Objects

In this chapter we take a first look at documents and how they are represented on the computer.

4.1 Representing & Manipulating Documents on a Computer

Now that we can represent characters as bit sequences, we can represent text documents. In principle text documents are just sequences of characters; they can be represented by just concatenating them.

Electronic Documents

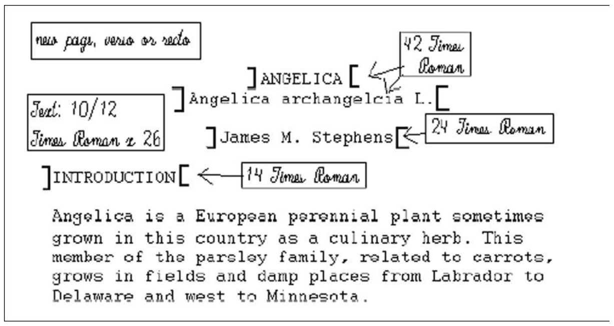
- ▷ **Definition 4.1.1.** An electronic document is any media content that is intended to be used via a document renderer, i.e. a program or computing device that transforms it into a form that can be directy perceived by the end user.
- ▷ **Example 4.1.2.** PDFs, digital images, videos, audio recordings, web pages, . . .
- ▷ **Definition 4.1.3.** An electronic document that contains a digital encoding of textual material that can be read by the end user by simply presenting the encoded characters is called digital text.
- ▷ **Definition 4.1.4.** Digital text is subdivided into plain text, where all characters carry the textual information and formatted text, which also contains instructions to the document renderer.
- ▷ **Example 4.1.5.** Python programs are plain text, PDFs are formatted.

We will now establish a nomenclature for giving instructions to a document renderer. This has originated from movable (lead) type based typesetting but carries over well to electronic documents.

Document Markup

- ▷ **Definition 4.1.6.** Document markup (or just markup) is the process of adding control words (special character sequences also called markup code) to a plain text to control the structure, formatting, or the relationship among its parts, making it a formatted text. All characters of a formatted text that are not control words constitute its textual content.

▷ **Example 4.1.7.** A text with markup codes (for printing)


new page, verso or recto
42 Times Roman
Text: 10/12 Times Roman x 26
24 Times Roman
14 Times Roman

]ANGELICA[

]Angelica archangelcia L.[

]James M. Stephens[

]INTRODUCTION[

Angelica is a European perennial plant sometimes grown in this country as a culinary herb. This member of the parsley family, related to carrots, grows in fields and damp places from Labrador to Delaware and west to Minnesota.

▷ **Definition 4.1.8.** The control words and composition rules for a particular kind of markup system determine a markup format (also called a markup language). The markup format used in an electronic document is called its document type.

▷ *Remark 4.1.9.* Markup turns plain text into formatted text.

There are many systems for document markup, ranging from informal ones as in ?? that specify the intended document appearance to humans – in this case the printer – to technical ones which can be understood by machines but serving the same purpose.
Markup is by no means limited to visual markup for documents intended for printing as ?? may suggest. There are aural markup formats that instruct document renderers that transform documents to audio streams of e.g. reading speeds, intonation, and stress.
We now come to another aspect of electronic documents: We mostly interact with them in the form of files. Again, we fix our nomenclature.

File Types

▷ **Observation 4.1.10.** *We mostly encounter electronic documents in the form of files on some storage medium.*

▷ **Definition 4.1.11.** A text file is a file that contains text data, a binary file one that contains binary data

Definition 4.1.12. Text files are often processed as a sequence of text line (or just lines), i.e. sub string separated by the line feed character $\mathrm{U+000A}$; LINEFEED(LF). The line number is just the position in the sequence.

▷ *Remark 4.1.13.* Text files are usually encoded with ASCII, ISO Latin, or increasingly UniCode encodings like UTF-8.

▷ **Example 4.1.14.** Python programs are stored in text files.

▷ In practice, text files are often processed as a sequence of text line (or just lines), i.e. sub strings separated by the line feed character $\mathrm{U+000A}$; LINEFEED(LF). The line number is just the position in the sequence.

Remark 4.1.15. Plain text is different from formatted text, which includes markup code, and binary files in which some portions must be interpreted as binary data (encoded integers, real numbers, digital images, etc.)
As we have seen above, it does not take much to render a text file: we only need to guess the right encoding scheme so we can decode the file and show the character sequence to the user. Indeed the `UNIX` cat just prints the contents of a text file to a shell. But we need much more, we need tools with which we can compose and edit text files; we do this with text editors, which we will discuss now.

Text Editors

▷ **Definition 4.1.16.** A text editor is a program used for rendering and manipulating text files.

▷ **Example 4.1.17.** Popular text editors include

- ▷ `Notepad` is a simple editor distributed with `MSWindows`.
- ▷ `emacs` and `vi` are powerful editors originating from `UNIX` and optimized for programming.
- ▷ `sublime` is a sophisticated programming editor for multiple operating systems.
- ▷ `EtherPad` is a browser-based real-time collaborative editor.

▷ **Example 4.1.18.** Even though it can save documents as text files, `MSWord` is not usually considered a text editor, since it is optimized towards formatted text; such "editors" are called word processors.

What text editors do for text files, word processors do for other electronic documents.

Word Processors and Formatted Text

▷ **Definition 4.1.19.** A word processor is a software application, that – apart from being a document renderer – also supports the tasks of composition, editing, formatting, printing of electronic documents.

▷ **Example 4.1.20.** Popular word processors include

- ▷ `MSWord`, an elaborated word processor for `MSWindows`, whose native format is Office Open XML (OOXML; file extension .docx).
- ▷ `OpenOffice` and `LibreOffice` are similar word processors using the ODF format (Open Office Format; file extension .odf) natively, but can also import other formats..
- ▷ `Pages`, a word processors for `macOS` it uses a proprietary format.
- ▷ `OfficeOnline` and `GoogleDocs` are browser-based real-time collaborative word processors.

▷ **Example 4.1.21.** Text editor are usually not considered to be word processors, even though they can sometimes be used to edit markup based formatted text.

Before we go on, let us first get into some basics: how do we measure information, and how does this relate to units of information we know.

4.2 Measuring Sizes of Documents/Units of Information

Having represented documents as sequences of characters, we can use that to measure the sizes of documents. In this section we will have a look at the underlying units of information and try to get an intuition about what we can store in files.

⚠ We will take a very generous stance towards what a document is, in particular, we will include pictures, audio files, spreadsheets, computer aided designs, ...

Units for Information

- ▷ **Observation:** The smallest unit of information is knowing the state of a system with only two states.
- ▷ **Definition 4.2.1.** A bit (a contraction of "binary digit") is the basic unit of capacity of a data storage device or communication channel. The capacity of a system which can exist in only two states, is one bit (written as $1\mathrm{b}$)
- ▷ **Note:** In the ASCII encoding, one character is encoded as $8\mathrm{b}$, so we introduce another basic unit:
- ▷ **Definition 4.2.2.** The byte is a derived unit for information capacity: $1\mathrm{B} = 8\mathrm{b}$.

From the basic units of information, we can make prefixed units for prefixed units for larger chunks of information. But note that the usual SI unit prefixes are inconvenient for application to information measures, since powers of two are much more natural to realize.

Larger Units of Information via Binary Prefixes

- ▷ We will see that memory comes naturally in powers to 2, as we address memory cell by binary number, therefore the derived information units are prefixed by special prefixes that are based on powers of 2.
- ▷ **Definition 4.2.3 (Binary Prefixes).** The following binary unit prefixes are used for information units because they are similar to the SI unit prefixes.

prefix	symbol	2^n	decimal	~SI prefix	Symbol
kibi	Ki	2^{10}	1024	kilo	k
mebi	Mi	2^{20}	1048576	mega	M
gibi	Gi	2^{30}	1.074×10^9	giga	G
tebi	Ti	2^{40}	1.1×10^{12}	tera	T
pebi	Pi	2^{50}	1.125×10^{15}	peta	P
exbi	Ei	2^{60}	1.153×10^{18}	exa	E
zebi	Zi	2^{70}	1.181×10^{21}	zetta	Z
yobi	Yi	2^{80}	1.209×10^{24}	yotta	Y

▷ **Note:** The correspondence works better on the smaller prefixes; for yobi vs. yotta there is a 20% difference in magnitude.

▷ The SI unit prefixes (and their operators) are often used instead of the correct binary ones defined here.

▷ **Example 4.2.4.** You can buy hard-disks that say that their capacity is "one terabyte", but they actually have a capacity of one tebibyte.

Let us now look at some information quantities and their real-world counterparts to get an intuition for the information content.

How much Information?

Bit (b)	*binary digit 0/1*
Byte (B)	*8 bit*
2 Bytes	A UniCode character in UTF.
10 Bytes	your name.
Kilobyte (kB)	*1,000 bytes OR 10^3 bytes*
2 Kilobytes	A Typewritten page.
100 Kilobytes	A low-resolution photograph.
Megabyte (MB)	*1,000,000 bytes OR 10^6 bytes*
1 Megabyte	A small novel or a 3.5 inch floppy disk.
2 Megabytes	A high-resolution photograph.
5 Megabytes	The complete works of Shakespeare.
10 Megabytes	A minute of high-fidelity sound.
100 Megabytes	1 meter of shelved books.
500 Megabytes	A CD-ROM.
Gigabyte (GB)	*1,000,000,000 bytes or 10^9 bytes*
1 Gigabyte	a pickup truck filled with books.
20 Gigabytes	A good collection of the works of Beethoven.
100 Gigabytes	A library floor of academic journals.

How much Information?

Terabyte (TB)	*1,000,000,000,000 bytes or 10^{12} bytes*
1 Terabyte	50000 trees made into paper and printed.
2 Terabytes	An academic research library.
10 Terabytes	The print collections of the U.S. Library of Congress.
400 Terabytes	National Climate Data Center (NOAA) database.
Petabyte (PB)	*1,000,000,000,000,000 bytes or 10^{15} bytes*
1 Petabyte	3 years of EOS data (2001).
2 Petabytes	All U.S. academic research libraries.
20 Petabytes	Production of hard-disk drives in 1995.
200 Petabytes	All printed material (ever).
Exabyte (EB)	*1,000,000,000,000,000,000 bytes or 10^{18} bytes*
2 Exabytes	Total volume of information generated in 1999.
5 Exabytes	All words ever spoken by human beings ever.
300 Exabytes	All data stored digitally in 2007.
Zettabyte (ZB)	*1,000,000,000,000,000,000,000 bytes or 10^{21} bytes*
2 Zettabytes	Total volume digital data transmitted in 2011
100 Zettabytes	Data equivalent to the human Genome in one body.

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 113 2025-02-06

The information in this table is compiled from various studies, most recently [HL11].

Note: Information content of real-world artifacts can be assessed differently, depending on the view. Consider for instance a text typewritten on a single page. According to our definition, this has ca. 2kB, but if we fax it, the image of the page has 2MB or more, and a recording of a text read out loud is ca. 50MB. Whether this is a terrible waste of bandwidth depends on the application. On a fax, we can use the shape of the signature for identification (here we actually care more about the shape of the ink mark than the letters it encodes) or can see the shape of a coffee stain. In the audio recording we can hear the inflections and sentence melodies to gain an impression on the emotions that come with text.

4.3 Hypertext Markup Language

WWW documents have a specialized document type that mixes markup for document structure with layout markup, hyper-references, and interaction. The HTML markup elements always concern text fragments, they can be nested but may not otherwise overlap. This essentially turns a text into a document tree. In IWGS, we discuss HTML mostly as a way to build interfaces of web applications. Therefore we will prioritize those aspects of HTML that have to do with "programming documents" over the creation of nice-looking web pages. Therefore we will pick up the notion of nested text fragments marked up by well-bracketed tags and elements in **??** and generalize these ideas to XML as a general representation paradigm for semi-structured data in **??**.

We will also postpone the discussion of cascading stylesheets, which have evolved as the dominant technology for the specification of presentation (layout, colors, and fonts) for marked-up documents, to **??**.

4.3.1 Introduction

HTML was created in 1990 and standardized in version 4 in 1997 [RHJ98]. Since then the WWW has evolved considerably from a web of static web pages to a Web in which highly dynamic web pages become user interfaces for web-based applications and even mobile applets. HTML5 standardized the necessary infrastructure in 2014 [Hic+14].

HTML: Hypertext Markup Language

▷ **Definition 4.3.1.** The HyperText Markup Language (HTML), is a representation format for web pages [Hic+14].

▷ **Definition 4.3.2 (Main markup elements of HTML).** HTML marks up the structure and appearance of text with tags of the form <el> (begin tag), </el> (end tag), and <el/> (empty tag), where el is one of the following

structure	`html`,`head`, `body`	metadata	`title`, `link`, `meta`
headings	`h1`, `h2`, . . . , `h6`	paragraphs	`p`, `br`
lists	`ul`, `ol`, `dl`, . . . , `li`	hyperlinks	`a`
multimedia	`img`, `video`, `audio`	tables	`table`, `th`, `tr`, `td`, . . .
styling	`style`, `div`, `span`	old style	`b`, `u`, `tt`, `i`, . . .
interaction	`script`	forms	`form`, `input`, `button`
Math	MathML (formulae)	interactive graphics	vector graphics (SVG) and `canvas` (2D bitmapped)

▷ **Example 4.3.3.** A (very simple) HTML file with a single paragraph.

```
<html>
  <body>
    <p>Hello IWGS students!</p>
  </body>
</html>
```

The thing to understand here is that HTML uses the characters <, >, and / to delimit the markup. All markup is in the form of tags, so anything that is not between < and > is the textual content.

We will not give a complete introduction to the various tags and elements of the HTML language here, but refer the reader to the HTML recommendation [Hic+14] and the plethora of excellent web tutorials. Instead we will introduce the concepts of HTML markup by way of examples.

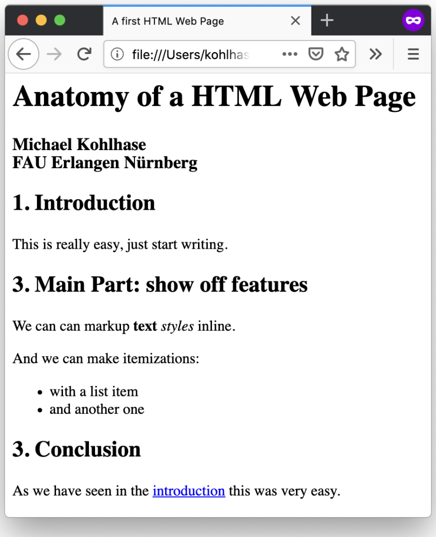
The best way to understand HTML is via an example. Here we have prepared a simple file that shows off some of the basic functionality of HTML.

A very first HTML Example (Source)

```
<html xmlns="http:www.w3.org/1999/xhtml">
  <head>
    <title>A first HTML Web Page</title>
  </head>
  <body>
    <h1>Anatomy of a HTML Web Page</h1>
    <h3>Michael Kohlhase<br/>FAU Erlangen Nuernberg</h3>
    <h2 id="intro">1. Introduction</h2>
    <p>This is really easy, just start writing.</p>
    <h2>3. Main Part: show off features</h2>
    <p>We can can markup <b>text</b> <em>styles</em> inline.</p>
    <p> And we can make itemizations:
      <ul>
        <li> with a list item</li>
        <li> and another one</li>
      </ul>
    </p>
    <h2>4. Conclusion</h2>
    <p> As we have seen in the <a href="#intro">introduction</a> this
    was very easy.</p>
  </body>
</html>
```

The thing to understand here is that HTML markup is itself a well-balanced structure of begin and end tags. That wrap other balanced HTML structures and – eventually – textual content. The HTML recommendation [Hic+14] specifies the visual appearance expectation and interactions afforded by the respective tags, which HTML-aware software systems – e.g. a web browser – then execute. In the next slide we see how FireFox displays the HTML document from the previous.

A very first HTML Example (Result)



Anatomy of a HTML Web Page

Michael Kohlhase
FAU Erlangen Nürnberg

1. Introduction

This is really easy, just start writing.

3. Main Part: show off features

We can can markup **text** *styles* inline.

And we can make itemizations:

- with a list item
- and another one

3. Conclusion

As we have seen in the introduction this was very easy.

4.3.2 Interacting with HTML in Web Broswers

In the last slide, we have seen FireFox as a document renderer for HTML. We will now introduce this class of programs in general and point out a few others.

Web Browsers

▷ **Definition 4.3.4.** A web browser is a software application for retrieving (via HTTP), presenting, and traversing information resources on the WWW, enabling users to view web pages and to jump from one page to another.

Definition 4.3.5. A web browser usually supplies user tools like

▷ history that gives the user access to web pages visited earlier and

▷ bookmark to remember web pages.

Definition 4.3.6. A web browser usually supplies developer tools like

- ▷ the console that logs system-level events in the browser and
- ▷ an inspector that gives access to the structure and content of the DOM.

▷ **Practical Browser Tools:**

- ▷ Status Bar: security info, page load progress
- ▷ Favorites (bookmarks)
- ▷ View Source: view the code of a web page
- ▷ Tools/Internet Options, history, temporary Internet files, home page, auto complete, security settings, programs, etc.

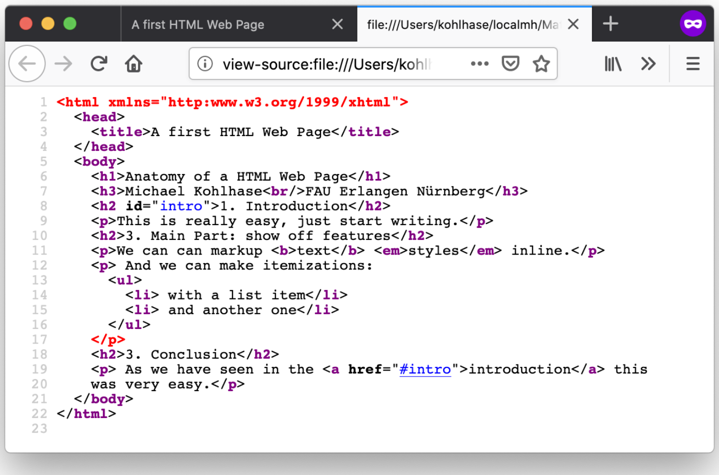
▷ **Example 4.3.7 (Common Browsers).**

- ▷ `MSInternetExplorer` is an once dominant, now obsolete browser for `MSWindows`.
- ▷ `Edge` is provided by Microsoft for `MSWindows`.(replaces `MSInternetExplorer`)
- ▷ `FireFox` is an open source browser for all platforms, it is known for its standards compliance.
- ▷ `Safari` is provided by Apple for `macOS` and `MSWindows`.
- ▷ `Chrome` is a lean and mean browser provided by Google Inc. (very common)
- ▷ `WebKit` is a library that forms the open source basis for `Safari` and `Chrome`.

Let us now look at a couple of more advanced tools available in most web browsers for dealing with the underlying HTML document.

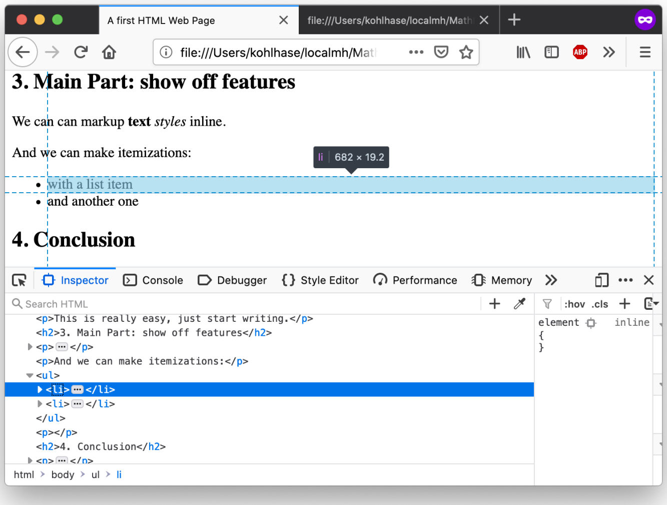
Browser Tools for dealing with HTML, e.g. in `FireFox`

▷ Hit Control-U to see the page source in the browser


A first HTML Web Page
file:///Users/kohlhase/localmh/Ma
view-source:file:///Users/kohl

```
<html xmlns="http:www.w3.org/1999/xhtml">
  <head>
    <title>A first HTML Web Page</title>
  </head>
  <body>
    <h1>Anatomy of a HTML Web Page</h1>
    <h3>Michael Kohlhase<br/>FAU Erlangen Nürnberg</h3>
    <h2 id="intro">1. Introduction</h2>
    <p>This is really easy, just start writing.</p>
    <h2>3. Main Part: show off features</h2>
    <p>We can can markup <b>text</b> <em>styles</em> inline.</p>
    <p> And we can make itemizations:
      <ul>
        <li> with a list item</li>
        <li> and another one</li>
      </ul>
    </p>
    <h2>3. Conclusion</h2>
    <p> As we have seen in the <a href="#intro">introduction</a> this
    was very easy.</p>
  </body>
</html>
```

▷ go to an element and right-click ⇝ "Inspect element"



3. Main Part: show off features

We can can markup **text** *styles* inline.

And we can make itemizations:

- with a list item
- and another one

4. Conclusion

```
<p>This is really easy, just start writing.</p>
<h2>3. Main Part: show off features</h2>
▸ <p>⋯</p>
<p>And we can make itemizations:</p>
▾ <ul>
  ▸ <li>⋯</li>
  ▸ <li>⋯</li>
</ul>
<p></p>
<h2>4. Conclusion</h2>
▸ <p>⋯</p>
```

```
element {
}
```

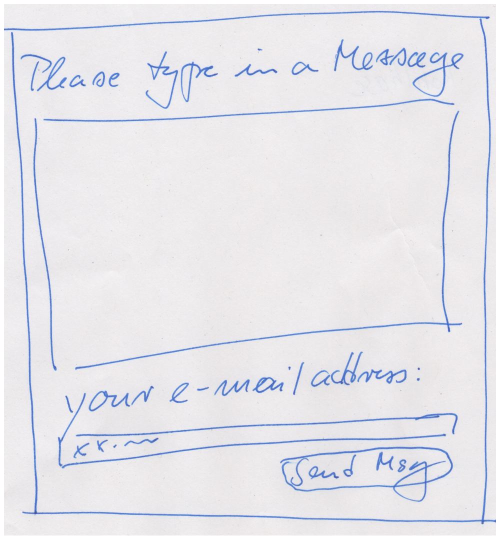
We have used FireFox as an example here, but these tools are available in some form in all major browsers the browser vendors want to make their offerings attractive to web developers, so that web pages and web applications get tested and debugged in them and therefore work as expected.

4.3.3 A Worked Example: The Contact Form

After this simple example, we will come to a more complex one: a little "contact form" as we find on many web sites that can be used for sending a message to the owner of the site. Let us only look a the design of the form document before we go into the interaction facilities afforded it.

HTML in Practice: Worked Example

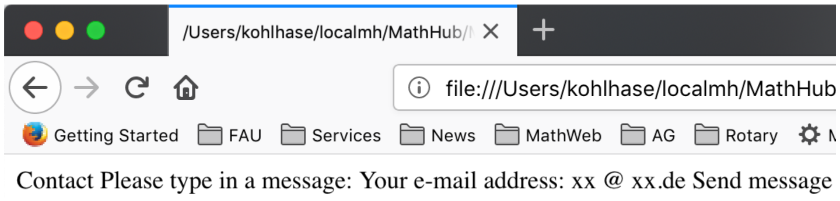
▷ Make a design and "paper prototype" of the page:



▷ Put the intended text into a file: contact.html:

```
Contact
Please enter a message:
Your e—mail address: xx @ xx.de
Send message
```

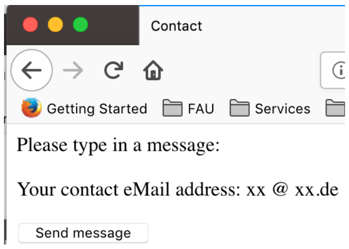
▷ Load into your browser to check the state:


/Users/kohlhase/localmh/MathHub/
file:///Users/kohlhase/localmh/MathHub
Getting Started
FAU
Services
News
MathWeb
AG
Rotary

Contact Please type in a message: Your e-mail address: xx @ xx.de Send message

▷ Add title, paragraph and button markup:

```
<title>Contact</title>
<h2>Please enter a message:</h2>
<h3>Your e—mail address: xx @ xx.de</h3>
<button>Send message</button>
```

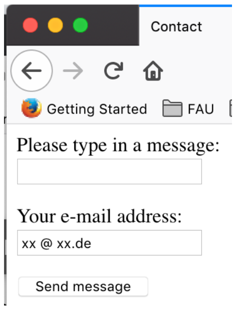


Please type in a message:

Your contact eMail address: xx @ xx.de

▷ Add input fields and breaks:

```
<title>Contact</title>
<h2>Please enter a message:</h2>
<input name="msg" type="text"/>
<h3> Your e−mail address:</h3>
<input name="addr" type="text"
        value="xx @ xx.de"/>
<br/>
<button>Send message</button>
```



Please type in a message:

Your e-mail address:

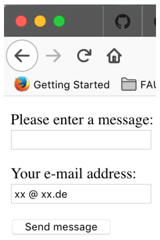
xx @ xx.de

Send message

▷ Convert into a HTML form with action (message receipt):

```
<title>Contact</title>
<form action="contact−after.html">
  <h2>Please enter a message:</h2>
    <input name="msg" type="text"/>
  <h3>Your e−mail address:</h3>
  <input name="addr" type="text"
          value="xx @ xx.de"/>
  <br/>
  <input type="submit"
         value="Send message"/>
</form>
```

```
<title>
  Contact − Message Confirmed
</title>
<form action="contact4.html">
  <h2>
    Your message has been submitted!
  </h2>
  <input type="submit"
         value="Continue"/>
</form>
```

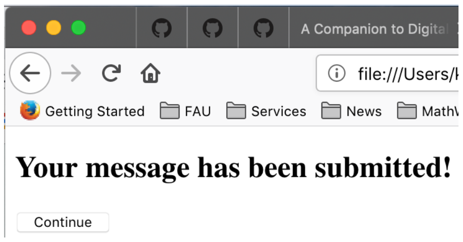


Please enter a message:

Your e-mail address:

xx @ xx.de

Send message



Your message has been submitted!

Continue

▷ That's as far as we will go, the rest is page layout and interaction. (up next)

After designing the functional (what are the text blocks) structure of the contact form, we will need to understand the interaction with the contact form.

HTML Forms

▷ **Question:** But how does the interaction with the contact form really work?

▷ **Definition 4.3.8.** A HTML form is realized by the HTML form tags, which groups the layout, form action specification and input fields:

- ▹ **<form action="**⟪URI⟫"...> specifies the form action (as a web page address).
- ▹ the input field **<input type=**"submit".../> triggers the form action: it sends the form data to web page specified there.

▷ **Example 4.3.9 (In the Contact Form).** We send the request

```
GET contact—after.html?
    msg=Hi;addr=foo@bar.de
```

We current ignore the form data (the part after the ?)

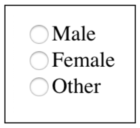
▷ We will come to the full story of processing actions later.

Unfortunately, we can only see what the browser sends to the server at the current state of play, not what the server does with the information. But we will get to this when we take up the example again.
For the moment, we made use of the fact that we can just specify the page contact—after.html, which the browser displays next. That ignores the query part and – via a form tags of its own gets the user back to the original contact form.

More useful types of Input fields

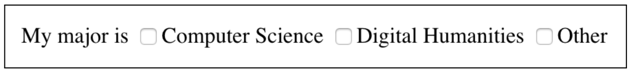
▷ Radio buttons: type="radio" (grouped by name attribute)

```
<input type="radio" name="gender" value="male"/>Male<br/>
<input type="radio" name="gender" value="female"/>Female<br/>
<input type="radio" name="gender" value="other"/>Other
```



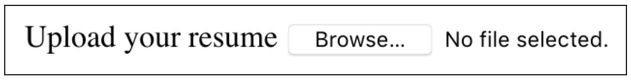
▷ Check boxes: type="checkbox"

```
My major is
<input type="checkbox" name="major" value="cs"/>Computer Science
<input type="checkbox" name="major" value="dh"/>Digital Humanities
<input type="checkbox" name="major" value="other"/>Other
```



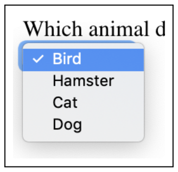
▷ File selector dialogs (interaction is system specific here for MacOS Mojave)

```
<p> Upload your resume <input type="file" name="resume"/></p>
```



▷ Drop down menus: select and option

```
Which animal do you like?<br/>
<select name="animals">
  <option value="bird">Bird</option>
  <option value="hamster">Hamster</option>
  <option value="cat">Cat</option>
  <option value="dog">Dog</option>
</select>
```



4.4 Documents as Trees

We have concentrated on HTML as a document type for interactive multimedia documents. Before we progress, we want to discuss an important feature: all practical document types that employ control words are in some sense well-bracketed. Well-bracketed structures are well-understood in

CS and mathematics: they are called trees and come with a rich and useful collection of descriptive concepts and tools. We will present the concepts in this section and the tools they enable in ??.

Well-Bracketed Structures in Computer Science

▷ **Observation 4.4.1.** *We often deal with well-bracketed structures in CS, e.g.*

- ▷ *Expressions: e.g.* $\frac{3\cdot(a+5)}{2x+7}$ *(numerator an denominator in fractions implicitly bracketed)*
- ▷ *Markup languages like HTML:*

```
<html>
  <head><script>.emph {color:red}</script></head>
  <body><p>Hello IWGS</p></body>
</html>
```

- ▷ *Programming languages like python:*

```
answer = input("Are you happy? ")
if answer == 'No' or answer == 'no':
    print("Have a chocolate!")
else:
    print("Good!")
print("Can I help you with something else?")
```

▷ **Idea:** Come up with a common data structure that allows to program the same algorithms for all of them. (common approach to scaling in computer science)

A Common Data Structure for Well Bracketed Structures

▷ **Observation 4.4.2.** *In well-bracketed strutures, brackets contain two kinds of objects*

- ▷ *bracket-less objects*
- ▷ *well-bracketed structures themselves*

▷ **Idea:** Write bracket pairs and bracket-less objects as nodes, connect with an arrow when contained. (let arrows point downwards)

▷ **Example 4.4.3.** Let's try this for HTML creating nodes top to bottom

```
<html>
  <head>
    <script>.emph {color:red}</script>
  </head>
  <body>
    <p>Hello IWGS</p>
  </body>
</html>
```

$$\begin{tikzcd}
 & \langle\text{html}\rangle \arrow[dl] \arrow[dr] & \\
\langle\text{head}\rangle \arrow[d] & & \langle\text{body}\rangle \arrow[d] \\
\langle\text{script}\rangle \arrow[d] & & \langle\text{p}\rangle \arrow[d] \\
\boxed{\text{.emph \{color:red\}}} & & \boxed{\text{Hello IWGS}}
\end{tikzcd}$$

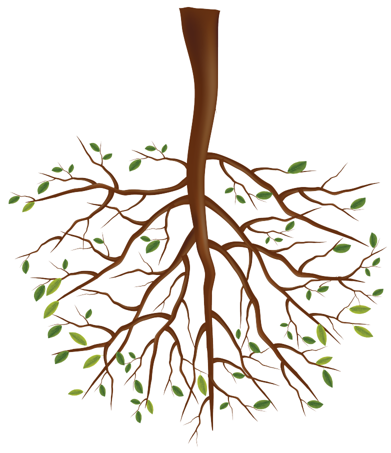
▷ **Definition 4.4.4.** We call such structures tree. (more on trees next)

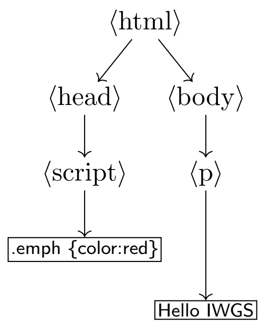
Trees are well understood mathematical objects and tree data structures are very commonly used in computer science and programming. As such they have a well-developed nomenclature, which we will introduce now.

Well-Bracketed Structures: Tree Nomenclature

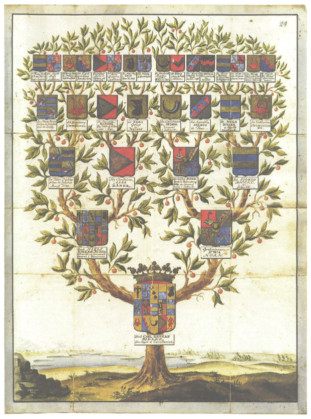
- ▷ **Definition 4.4.5.** In mathematics and CS, such well-bracketed structures are called trees (with root, branches, leaves, and height). (but written upside down)

- ▷ **Example 4.4.6.** In a tree, there is only one path from the root to the leaves





- ▷ **Definition 4.4.7.** We speak of parent, child, ancestor, and descendant nodes (genealogy nomenclature).

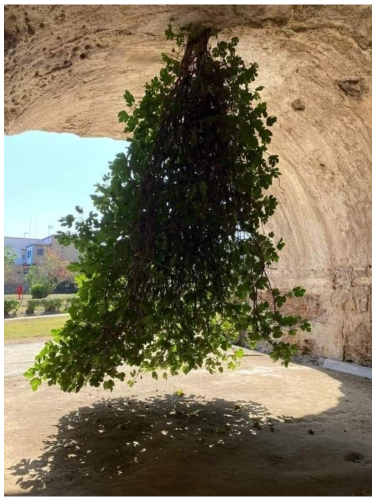


$$\begin{tikzcd}
& \langle \text{html} \rangle \arrow[dl] \arrow[dr] & \\
\langle \text{head} \rangle \arrow[d] & & \langle \text{body} \rangle \arrow[d] \\
\langle \text{script} \rangle \arrow[d] & & \langle \text{p} \rangle \arrow[d] \\
\boxed{\text{.emph \{color:red\}}} & & \boxed{\text{Hello IWGS}}
\end{tikzcd}$$

Why are trees written upside-down?: The main answer is that we want to draw tree diagrams in text. And we naturally start drawing a tree at the root. So, if a tree grows from the root and we do not exactly know the tree height, then we do not know how much space to leave. When we write trees upside down, we can directly start from the root and grow the tree downward as long as we need. We will keep to this tradition in the IWGS course.

Upside Down Trees in Nature

▷ Actually, upside down trees exist in nature (though rarely):



This is a fig tree in Bacoli, Italy; see `https://www.atlasobscura.com/places/upside-down-fig-tree`

We will now make use of the tree structure for computation. Even if the computing tasks we pursue here may seem a bit abstract, they show very nicely how tree algorithms typically work.

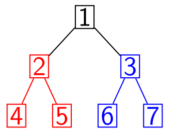
Computing with Trees in Python

▷ **Observation 4.4.8.** *All connected substructures of trees are trees themselves.*

▷ **Idea:** Operate on the tree by "Divide and Conquer"

▷ operate on the two subtrees

▷ combine results, taking root into account



This approach lends itself very well to recursive programming (functions that call themselves)

▷ **Idea:** Represent trees as lists of tree labels and lists (of subtrees).

▷ **Example 4.4.9 (The tree above).** Represented as [1,[2,[[4],[5]]],[3,[[6],[7]]]]
compute the tree height by the following Python functions:

```
def height (tree):
    return maxh(tree[1:]) + 1

height([1,[2,[[4],[5]]],[3,[[6],[7]]]])
>>> 3
```

```
def maxh (l):
    if l == []:
        return 0
    else
        return max(height(l[0]),maxh(l[1:]))
```

Let use have a closer look at ??. The algorithm consists of two functions:

1. height, which computes the height of an input tree by delegating the computation of the maximal height of its children to maxh and then incrementing the value by 1.

2. maxh, which takes a list of trees and computes the maximum of their heights by calling height on the first input tree and then comparing with the maximal height of the remaining trees.

Note that maxh and height each call the other. We call such functions mutually recursive. Here this behavior poses no problem, since the arguments in the recursive calls are smaller than the inputs: for maxh it is the rest list, and for height the "list of children" of the input tree.
?? was complex for two reasons: mutual recursion and the somewhat cryptic encoding of trees as lists of lists of integers. We claim that recursive programming is "not a bug, but a feature", as it allows to succinctly capture the "divide-and-conquer" approach afforded by trees. For the cryptic encoding of trees we can do better.

Computing with Trees in Python (Dictionaries)

▷ **That was a bit cryptic:** i.e. very difficult to read/debug

▷ **Idea:** why not use dictionaries? (they are more explicit)

▷ **Example 4.4.10.** Compute the tree weight (the sum of all labels) by

```
t =
{"label": = 1,
 "children": = [{
       "label": = 2,
       "children": = [{
             "label": = 4,
             "children": = []},
           {"label": = 5,
             "children": = []}]},
       {"label": = 3,
         "children": = [{
             "label": = 6,
             "children": = []},
           {"label": = 7,
             "children": = []}]}]}
```

```
def wsum (tl):
    if tl == []:
        return 0;
    else
        return weight(tl[0]) + wsum(tl[1:])

def weight (tree):
    return tree["label"] + wsum(tree["children"]);

weight(t);
>>> 28
```

Again, we have two mutually recursive functions: weight that takes a tree, and wsum that takes a list and the recursion goes analogously. Only that this time, the list of children is a dictionary value and the calls are clearer. The only real difference, is that in wsum we have to add up the weight of the head of the list an the joint sum of the rest list.

The Document Object Model

▷ **Definition 4.4.11.** The document object model (DOM) is a data structure for storing marked up electronic documents as trees together with a standardized set of access methods for manipulating them.

▷ **Idea:** When a web browser loads a HTML page, it directly parses it into a DOM and then works exclusively on that. In particular, the HTML document is immediately discarded; documents are rendered from the DOM.

4.5 An Overview over XML Technologies

We have seen that many of the technologies that deal with marked-up documents utilize the tree-like structure of (the DOM) of HTML documents. Indeed, it is possible to abstract from the concrete vocabulary of HTML that implements the intended layout of hypertexts and the function of its fragments, and build a generic framework for document trees. This is what we will study in this section.

4.5.1 Introduction to XML

XML (EXtensible Markup Language)

▷ **Definition 4.5.1.** XML (short for Extensible Markup Language) is a framework for markup formats for documents and structured data.

- ▷ Tree representation language (begin/end brackets)
- ▷ Restrict instances by *Doc. Type Def. (DTD)* or *Schema* (Grammar)
- ▷ Presentation markup by *style files* (XSL: XML Style Language)

- ▷ **Intuition:** XML is extensible HTML
- ▷ logic annotation (*markup*) instead of presentation!
- ▷ many tools available: parsers, compression, data bases, ...
- ▷ **conceptually**: transfer of trees instead of strings.
- ▷ details at `http://w3c.org` (XML is standardize by the WWW Consortium)

The idea of XML being an "extensible" markup language may be a bit of a misnomer. It is made "extensible" by giving language designers ways of specifying their own vocabularies. As such XML does not have a vocabulary of its own, so we could have also it an "empty" markup language that can be filled with a vocabulary.

XML is Everywhere (E.g. Web Pages)

▷ **Example 4.5.2.** Open web page file in `FireFox`, then click on $View \searrow PageSource$, you get the following text: (showing only a small part and reformatting)

```
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Michael Kohlhase</title>
    <meta name="generator"
          content="Page generated from XML sources with the WSML package"/>
  </head>
  <body>...
    <p>
      <i>Professor of Computer Science</i><br/>
      Jacobs University<br/><br/>
      <strong>Mailing address - Jacobs (except Thursdays)</strong><br/>
      <a href="http://www.jacobs-university.de/schools/ses">
        School of Engineering amp; Science</a><br/>...</p>...</body></html>
```

▷ **Definition 4.5.3.** XHTML is the XML version of HTML.(just make it valid XML)

Now we see an example of an XML file that is used for communicating data in a machine-readable, but human-understandable way.

XML is Everywhere (E.g. Catalogs)

▷ **Example 4.5.4 (The NYC Galleries Catalog).** A public XML file at `https://data.cityofnewyork.us/download/kcrmj9hh/application/xml`

```
<?xml version="1.0" encoding="UTF−8"?>
<museums>
  <museum>
    <name>American Folk Art Museum</name>
    <phone>212−265−1040</phone>
    <address>45 W. 53rd St. (at Fifth Ave.)</address>
    <closing>Closed: Monday</closing>
    <rates>admission: $9; seniors/students, $7; under 12, free</rates>
    <specials>
      Pay−what−you−wish: Friday after 5:30pm;
      refreshments and music available
    </specials>
```

```
</museum>
<museum>
  <name>American Museum of Natural History</name>
  <phone>212−769−5200</phone>
  <address>Central Park West (at W. 79th St.)</address>
  <closing>Closed: Thanksgiving Day and Christmas Day</closing>
```

This XML uses an ad hoc markup language: Every <museum> element represents one museum in New York City (NYC). Its children convey the detailed information as "key value pairs".
And now, if you still need proof that XML is really used almost everywhere, here is the ultimate example.

XML is Everywhere (E.g. Office Suites)

▷ **Example 4.5.5 (MS Office uses XML).** The MSOffice suite and LibreOffice use compressed XML as an electronic document format.

1. Save a MSOffice file test.docx, add the extension .zip to obtain test.docx.zip.
2. Uncompress with unzip (UNIX) or open File Explorer, right-click ⇝ "Extract All" (MSWindows)
3. You obtain a folder with 15+ files, the content is in word/contents.xml
4. Other files have packaging information, images, and other objects.

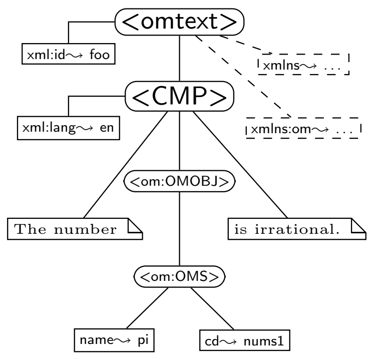
⚠ This is huge and offensively ugly.

- ▷ But you have everything you wanted and more
- ▷ In particular, you can process the contents via a program now.

XML Documents as Trees

▷ **Idea:** An XML Document is a Tree

```
<omtext xml:id="foo"
   xmlns="..."
   xmlns:om="...">
  <CMP xml:lang='en'>
   The number
   <om:OMOBJ>
     <om:OMS cd="nums1"
             name="pi"/>
     </om:OMOBJ>
   is irrational.
  </CMP>
</omtext>
```



▷ **Definition 4.5.6.** The XML document tree is made up of XML elements, attribute nodes, text nodes (and namespace declarations, comments,...)

XML Documents as Trees (continued)

▷ **Definition 4.5.7.** For communication this tree is serialized into a balanced bracketing structure, where

- ▷ an inner XML element nodes is represented by the brackets <el> (called the opening tag) and </el> (called the closing tag),
- ▷ the leaves of the XML tree are represented by empty element tags (serialized as <el></el>, which can be abbreviated as <el/>,
- ▷ and text node (serialized as a sequence of UniCode characters).
- ▷ An XML element node can be annotated by further information using attribute nodes serialized as an attribute in its opening tag.

▷ **Note:** As a document is a tree, the XML specification mandates that there must be a unique document root.

4.5.2 Computing with XML in Python

We have claimed above that the tree nature of XML documents is one of the main advantages. Let us now see how Python makes good on this promise.

We use the external lxml library [LXMLa] in IWGS, even though the Python distribution includes the standard library ElementTree library [ET] for dealing with XML. lxml subsumes ElementTree and extends it by functionality for XPath and can parse a large set of HTML documents even though they are not valid XML. This makes lxml a better basis for practical applications in the Digital Humanities.

Acknowledgements: Many of the examples and the flow of exposition in the next slides has been adapted from the lxml tutorial [LXMLc].

Computing with XML in Python (Elements)

▷ The lxml library [LXMLa] provides Python bindings for the (low-level) LibXML2 library. (install it with pip3 install lxml)

▷ The ElementTree API is the main way to programmatically interact with XML. Activate it by importing etree from lxml:

```
>>> from lxml import etree
```

▷ Elements are easily created, their properties are accessed with special accessor methods

```
>>> root = etree.Element("root")
>>> print(root.tag)
root
```

▷ Elements are organised in an XML tree structure. To create child element nodes and add them to a parent element nodes, you can use the append() method:

```
>>> root.append( etree.Element("child1") )
```

▷ **Abbreviation:** create a child element node and add it to a parent.

```
>>> child2 = etree.SubElement(root, "child2")
>>> child3 = etree.SubElement(root, "child3")
```

Computing with XML in Python (Result)

▷ Here is the resulting XML tree so far; we serialize it via etree.tostring

```
>>> print(etree.tostring(root, pretty_print=True))
<root>
  <child1/>
  <child2/>
  <child3/>
</root>
```

▷ BTW, the etree.tostring is highly configurable via default arguments.

```
tostring(element_or_tree,
         encoding=None, method="xml", xml_declaration=None, doctype=None,
         pretty_print=False, with_tail=True, standalone=None, exclusive=False,
         inclusive_ns_prefixes=None, with_comments=True, strip_text=False)
```

The lxml API documentation [LXMLb] has the details.

This method of "manually" producing XML trees in memory by applying etree methods may seem very clumsy and tedious. But the power of lxml lies in the fact that these can be embedded in Python programs. And as always, programming gives us the power to do things very efficiently.

Computing with XML in Python (Automation)

▷ This may seem trivial and/or tedious, but we have Python power now:

```
def nchildren (n):
    root = etree.Element("root")
    for i in range(1,n):
        root.append(f"child{i}")
```

produces a tree with 1000 children without much effort.

```
>>> t = nchildren(1000)
>>> print(len(t))
>>> 1000
```

We abstain from printing the XML tree (too large) and only check the length.

But XML documents that only have elements, are boring; let's do XML attributes next. Recall that attributes are essentially string-valued key/value pairs. So what could be more natural than treating them like dictionaries.

Computing with XML in Python (Attributes)

▷ Attributes can directly be added in the Element function

```
>>> root = etree.Element("root", interesting="totally")
>>> etree.tostring(root)
b'<root interesting="totally"/>'
```

▷ The .get method returns attributes in a dictionary-like object:

```
>>> print(root.get("interesting"))
totally
```

We can set them with the .set method:

```
>>> root.set("hello", "Huhu")
>>> print(root.get("hello"))
Huhu
```

This results in a changed element:

```
>>> etree.tostring(root)
b'<root interesting="totally" hello="Huhu"/>'
```

Recall that we could use Python dictionaries for iterating over in a for loop. We can do the same for attributes:

Computing with XML in Python (Attributes; continued)

▷ We can access attributes by the keys, values, and items methods, known from dictionaries:

```
>>> sorted(root.keys())
['hello', 'interesting']

>>> for name, value in sorted(root.items()):
... print(f'{name} = {value}')
hello = 'Huhu'
interesting = 'totally'
```

▷ ⚠ To get a 'real' dictionary, use the attrib method (e.g. to pass around)

```
>>> attributes = root.attrib
```

Note that attributes participates in any changes to root and vice versa.

▷ ⚠ To get an independent snapshot of the attributes that does not depend on the XML tree, copy it into a dict:

```
>>> d = dict(root.attrib)
>>> sorted(d.items())
[('hello', 'Guten Tag'), ('interesting', 'totally')]
```

The last two items touch a somewhat delicate subject in programming. Mutable an immutable data structures: the former can be changed in place as we have above with the .set method, and the latter cannot. Both have their justification and respective advantages. Immutable data structures are “safe” in the sense that they cannot be changed unexpectedly by another part of the program, they have the disadvantage that every time we want to have a variant, we have to copy the whole object. Mutable ones do not – we can change in place – but we have to be very careful about who accesses them when.

This is also the reason why we spoke of “dictionary-like interface” to XML trees in lxml: dictionaries are immutable, while XML trees are not.

The main remaining functionality in XML is the treatment of text. XML treats text as special kinds of node in the tree: text nodes. They can be treated just like any other node in the XML tree in the etree library.

Computing with XML in Python (Text nodes)

- ▷ XML elements can contain text: we use the .text property to access and set it.

```
>>> root = etree.Element("root")
>>> root.text = "TEXT"
>>> print(root.text)
TEXT
>>> etree.tostring(root)
b'<root>TEXT</root>'
```

To get a real intuition about what is happening, let us see how we can use all the functionality so far: we programmatically construct an HTML tree.

Case Study: Creating an HTML document

- ▷ We create nested html and body elements

```
>>> html = etree.Element("html")
>>> body = etree.SubElement(html, "body")
```

- ▷ Then we inject a text node into the latter using the .text property.

```
>>> body.text = "TEXT"
```

- ▷ Let's check the result

```
>>> etree.tostring(html)
b'<html><body>TEXT</body></html>'
```

- ▷ We add another element: a line break and check the result

```
>>> br = etree.SubElement(body, "br")
>>> etree.tostring(html)
b'<html><body>TEXT<br/></body></html>'
```

- ▷ Finally, we can add trailing text via the .tail property

```
>>> br.tail = "TAIL"
>>> etree.tostring(html)
b'<html><body>TEXT<br/>TAIL</body></html>'
```

Note the use of the .tail property here? While the .text property can be used to set "all" the text in an XML element, we have to use the .tail property to add trailing text (e.g. after the <br/> element).
Notwithstanding the "Python power" argument from above, there are situations, where we just want to write down XML fragments and insert them into (programmatically created) XML trees. lxml as functionality for this: XML literals, which we introduce now.

Computing with XML in Python (XML Literals)

▷ **Definition 4.5.8.** We call any string that is well-formed XML an XML literal.

▷ We can use the XML function to read XML literals.

```
>>> root = etree.XML("<root>data</root>")
```

The result is a first-class element tree, which we can use as above

```
>>> print(root.tag)
root
>>> etree.tostring(root)
b'<root>data</root>'
```

BTW, the fromstring function does the same.

▷ There is a variant html that also supplies the necessary HTML decoration.

```
>>> root = etree.HTML("<p>data<br/>more</p>")
>>> etree.tostring(root)
b'<html><body><p>data<br/>more</p></body></html>'
```

▷ **BTW:** If you want to read only the text content of an XML element, i.e. without any intermediate tags, use the method keyword in tostring:

```
>>> etree.tostring(root, method="text")
b'datamore'
```

4.5.3 XML Namespaces

We now come to a topic that is considered very difficult, confusing, and un-necessary by many people: XML namespaces. But it really is not, if you approach it with an open mind. Indeed it is probably what you would have come up with if you had been presented with the problem of mixing vocabularies, which is in turn a consequence of the fact that XML is used pervasively in the computing world and especially in Digital Humanities, where we often need to aggregate semi-structured data from multiple sources (and this multiple XML vocabularies).

XML is Everywhere (E.g. document metadata)

▷ **Example 4.5.9.** Open a PDF file in `AcrobatReader`, then click on

$$File \searrow DocumentProperties \searrow DocumentMetadata \searrow ViewSource$$

you get the following text: (showing only a small part)

```
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
         xmlns:iX='http://ns.adobe.com/iX/1.0/'>
```

```
  <rdf:Description xmlns:pdf='http://ns.adobe.com/pdf/1.3/'>
    <pdf:CreationDate>2004-09-08T16:14:07Z</pdf:CreationDate>
    <pdf:ModDate>2004-09-08T16:14:07Z</pdf:ModDate>
    <pdf:Producer>Acrobat Distiller 5.0 (Windows)</pdf:Producer>
    <pdf:Author>Herbert Jaeger</pdf:Author>
    <pdf:Creator>Acrobat PDFMaker 5.0 for Word</pdf:Creator>
    <pdf:Title>Exercises for ACS 1, Fall 2003</pdf:Title>
  </rdf:Description>
  ...
  <rdf:Description xmlns:dc='http://purl.org/dc/elements/1.1/'>
    <dc:creator>Herbert Jaeger</dc:creator>
    <dc:title>Exercises for ACS 1, Fall 2003</dc:title>
  </rdf:Description>
</rdf:RDF>
```

▷ **Example 4.5.10.** ?? mixes elements from three different vocabularies:

- ▷ RDF: xmlns:rdf for the "Resource Descritpion Format",
- ▷ PDF: xmlns:pdf for the "Portable Document Format", and
- ▷ DC: xmlns:dc for the "Dublin Core" vocabulary

This is an excerpt from the document metadata which AcrobatDistiller saves along with each PDF document it creates. It contains various kinds of information about the creator of the document, its title, the software version used in creating it and much more. Document metadata is useful for libraries, bookselling companies, all kind of text databases, book search engines, and generally all institutions or persons or programs that wish to get an overview of some set of books, documents, texts. The important thing about this document metadata text is that it is not written in an arbitrary, PDF proprietary format. Document metadata only make sense if these metadata are independent of the specific format of the text. The metadata that MSWord saves with each Word document should be in the same format as the metadata that Amazon saves with each of its book records, and again the same that the British library uses, etc.
We will now reflect what we have seen in ?? and fully define the namespacing mechanisms involved. Note that these definitions are technically involved, but conceptually quite natural. As a consequence they should be read more with an eye towards "what are we trying to achieve" than the technical details.

Mixing Vocabularies via XML Namespaces

▷ **Problem:** We would like to reuse elements from different XML vocabularies
What happens if element names coincide, but have different meanings?

▷ **Idea:** Disambiguate them by vocabulary name. (prefix)

▷ **Problem:** What if vocabulary names are not unique? (e.g. different versions)

▷ **Idea:** Use a long string for identification and a short prefix for referencing

▷ **Definition 4.5.11.** An XML namespace is a string that identifies an XML vocabulary. Every element and attribute name in XML consists of a local name and a namespace.

▷ **Definition 4.5.12.** A namespace declaration is an attribute xmlns:⟪prefix⟫= whose value is an XML namespace n on an XML element e. The first associates the namespace prefix ⟪prefix⟫ with the namespace n in e: Then, any XML element in e with a prefixed name ⟪prefix⟫:⟪name⟫ has namespace n and local name ⟪name⟫.

A default namespace declaration xmlns=d on an element e gives all elements in e whose name is not prefixed, the namepsace d.

Namespace declarations on subtrees shadow the ones on supertrees.

4.5.4 XPath: Specifying XML Subtrees

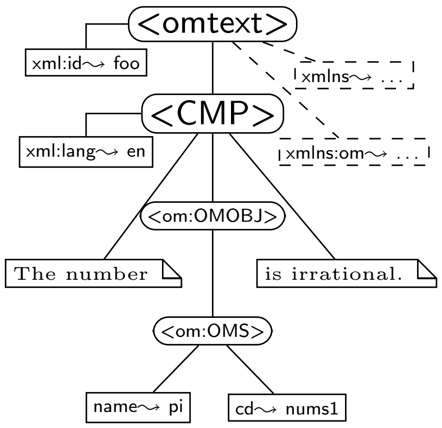
One of the great advantages of viewing marked-up documents as trees is that we can describe subsets of its nodes.

XPath, A Language for talking about XML Tree Fragments

▷ **Definition 4.5.13.** The XML path language (XPath) is a framework for specifying (sets of) fragments of XML trees by specifying paths from the root.

▷ **Intuition:** XPath is for trees what regular expressions are for strings.

▷ **Example 4.5.14.**



XPath exp.	fragment
/	root
omtext/CMP/∗	all <CMP> children
//@name	the name attribute on the <OMS> element
//CMP/∗[1]	the first child of all <CMP> elements
//∗[@cd='nums1']	all elements whose cd has value nums1

An XPath processor is an application or library that reads an XML file into a DOM and given an XPath expression returns (pointers to) the set of nodes in the DOM that satisfy the expression.

Computing with XML in Python (XPath)

▷ Say we have an XML tree:

```
>>> f = StringIO('<foo><bar></bar></foo>')
>>> tree = etree.parse(f)
```

▷ Then xpath() selects the list of matching elements for an XPath:

```
>>> r = tree.xpath('/foo/bar')
>>> len(r)
1
>>> r[0].tag
'bar'
```

▷ And we can do it again, . . .

```
>>> r = tree.xpath('bar')
>>> r[0].tag
'bar'
```

▷ The xpath() method has support for XPath variables:

```
>>> expr = "//*[local-name() = $name]"
>>> print(root.xpath(expr, name = "foo")[0].tag)
foo
>>> print(root.xpath(expr, name = "bar")[0].tag)
bar
```

To see that XPath is not just a plaything, we will now look at at a typical example where we can identify useful subtrees in a large HTML document: the Wikipedia page on paintings by Leonardo da Vinci.

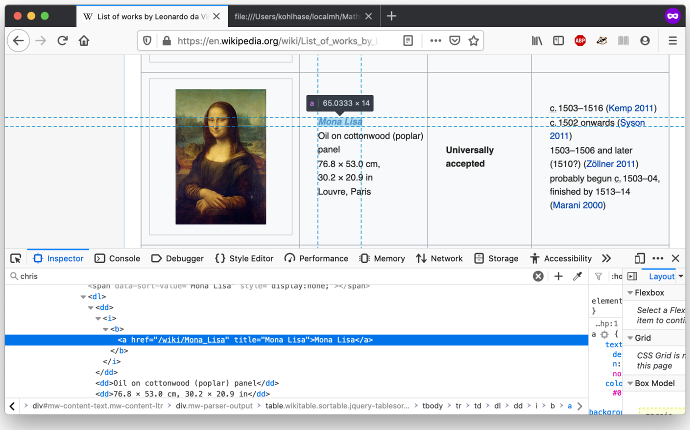
XPath Example: Scraping Wikipedia

▷ **Example 4.5.15 (Extracting Information from HTML).**

▷ We want a list of all titles of paintings by Leonardo da Vinci.

▷ open `https://en.wikipedia.org/wiki/List_of_works_by_Leonardo_da_Vinci` in FireFox. (save it into a file leo.html)

▷ call DOM inspector to get an idea of the XPath of titles. (bottom line)


List of works by Leonardo da Vi
file:///Users/kohlhase/localmh/Math
https://en.wikipedia.org/wiki/List_of_works_by_L
a 65.0333 × 14
Mona Lisa
Oil on cottonwood (poplar) panel
76.8 × 53.0 cm, 30.2 × 20.9 in
Louvre, Paris
Universally accepted
c. 1503–1516 (Kemp 2011)
c. 1502 onwards (Syson 2011)
1503–1506 and later (1510?) (Zöllner 2011)
probably begun c. 1503–04, finished by 1513–14 (Marani 2000)
Inspector
Console
Debugger
Style Editor
Performance
Memory
Network
Storage
Accessibility
chris
Layout
Flexbox
Select a Flex item to conti
Grid
CSS Grid is r this page
Box Model

```
<span data-sort-value="Mona Lisa" style="display:none;"></span>
<dl>
  <dd>
    <i>
      <b>
        <a href="/wiki/Mona_Lisa" title="Mona Lisa">Mona Lisa</a>
      </b>
    </i>
  </dd>
  <dd>Oil on cottonwood (poplar) panel</dd>
  <dd>76.8 × 53.0 cm, 30.2 × 20.9 in</dd>
```

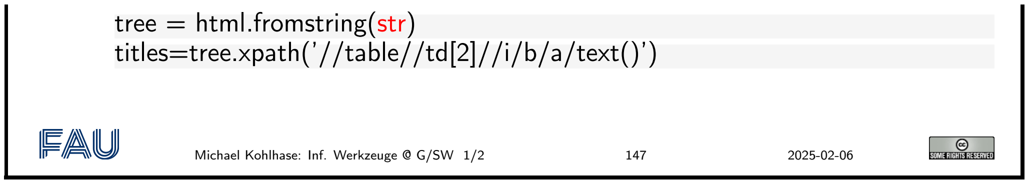
```
div#mw-content-text.mw-content-ltr > div.mw-parser-output > table.wikitable.sortable.jquery-tablesor... > tbody > tr > td > dl > dd > i > b > a
```

The path is table > tbody > tr > td > dl > dd > i > b > a
Alternatively: right-click on highlighted line, $\rightsquigarrow$ "copy" $\rightsquigarrow$ "XPath", gives /html/body/div[3]/div[3]/div[4]/div/table[4]/tbody/tr[3]/td[2]/dl/dd/i/b/a.

▷ **Idea**: We want to use the second table cells td[2].

▷ Program it in Python using the lxml library: titles is list of title strings.

```
from lxml import html

with open('leo.html', 'r') as m:
    str = m.read()
```



```
tree = html.fromstring(str)
titles=tree.xpath('//table//td[2]//i/b/a/text()')
```

Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 147 2025-02-06

If the task of writing an XPath for extracting the 50+ titles from this page does not convince you as worth learning XPath for, consider that Wikipedia has ca. 30 such lists, which apparently have exactly the same tree structure, so the XPath developed once for da Vinci, probably works for all the others as well.

4.6 Exercises

Problem 6.1 (HTML table)

In the lecture you saw the overview table for HTML below.

purpose	**elements**	**purpose**	**elements**
structure	`html`,`head`, `body`	metadata	`title`, `link`, `meta`
headings	`h1`, `h2`, ..., `h6`	paragraphs	`p`, `br`
lists	`ul`, `ol`, `dl`, ..., `li`	hyperlinks	`a`
multimedia	`img`, `video`, `audio`	tables	`table`, `th`, `tr`, `td`, ...
styling	`style`, `div`, `span`	old style	`b`, `u`, `tt`, `i`, ...
interaction	`script`	forms	`form`, `input`, `button`
Math	MathML (formulae)	interactive graphics	vector graphics (SVG) and `canvas` (2D bitmapped)

Make a HTML file `htmltable.html` that re creates this table in HTML. Note that the table heading is boldface and all of the HTML element names in the right column are in typewriter font (but the commata, ellipses, and explanations are not.)

Problem 6.2 (A Simple HTML Page)

Have a look at `https://www.izdigital.fau.de/efi-digitale-souveraenitaet/`. This page has header and footer parts (in blue) and two columns of text in between. The left one has the main text of the page (the page payload) and the right one some information about other pages on the same web site.

Make a simple web page from the payload text and the page heading "EFI-Förderung für das Forschungsprojekt „Diskurse und Praktiken einer digitalen Souveränität“".

1. Download the file `https://kwarc.info/teaching/IWGS/materials/efi.txt`, save it, and rename it to `efi.html`.
2. With the HTML tags we have introduced in the lecture mark up all structural parts: paragraphs, itemized lists, hyperlinks (Hint: you can obtain the link target by right-clicking on the hyperlink and selecting "Copy Link Address". You only need to mark up five links total.)
3. Load your `.html` file into a browser of your choice (this acts as the HTML document viewer) and export the contents to PDF (call the file `efi.pdf`).
4. Use the HTML checker at `https://validator.w3.org/#validate_by_upload` to see what it thinks of your HTML. Correct your errors reported there (as much as reasonable). Briefly discuss what your experience has been with this tool.

Submit `efi.html`, `efi.pdf`, and your discussion from 4.

Problem 6.3 (Simple HTML Form)

For this exercise, you will construct a very simple HTML page with a basic form. Suppose you want to establish a basic pizza delivery service only for **FAU** staff and students. It is your task to make the first version of the website for the "front-end" (that is, the user-facing part of the application).

Create a `.html` file[3] with a title, a heading, a paragraph or so of descriptive text and a `<form>`-element that contains the following inputs:

- a text input field for people to enter their name,
- a dropdown menu with (at least three) FAU-related addresses,
- (at least three) radio buttons labeled with different pizza options (for the moment, we only allow one pizza to be ordered at a time).
- a form-submit button.

When the submit button is clicked by the user, they should be redirected to a second HTML page (hand this in, too, in a separate file), that tells the user their order has been received. Use the form `action` attribute to accomplish this. This second page does *not* need to use the data from the form.

Problem 6.4 (Regex Parsing)

Suppose that you are now working on the Python "back-end" (that is, the part of the software that is managing and manipulating the data) of your **FAU**-internal pizza delivery sercive from **??**.

Say you have a log file where in each line there is a percent-encoded[4] HTML POST request to your web site. Each of them encodes the *name*, *address* and *pizza choice* of one order, like in the following examples:

```
POST name%3DTheo+McTestPerson%26address%3Dkollegienhaus%26pizza%3Dsalame
POST name%3DMax+Musterfrau%26address%3Dkollegienhaus%26pizza%3Dvegetaria
POST name%3DBea+Beispielname%26address%3Dmartensstrasse%26pizza%3Dsalame
...
```

Such a file is also being provided along with this exercise.[5] Write a program that first reads that file and creates a list of Python dictionaries (one for each order, with the keys `"name"`, `"address"` and `"pizza"`) out of the included data.[6] Use regular expressions to find the corresponding values in the data.

The program should then do the following:

- Your program needs to compute (and print) what sorts of pizzas were ordered and how many of each are needed in total.
- Your program should also print all adresses that the delivery driver needs to go to.
- Lastly, your program should compute and display the total amount of money that you would expect to be paid for this delivery (you can assign an arbitrary price to each variety of pizza for this exercise).

Problem 6.5 (Trees in Python & Recursion)

[3] If you need a refresher: there is excellent documentation on how the basics work at `https://www.w3schools.com/html/html_intro.asp` and related pages.

[4] See: `https://en.wikipedia.org/wiki/Percent-encoding`

[5] Found here: `https://kwarc.info/teaching/IWGS/materials/console.log`

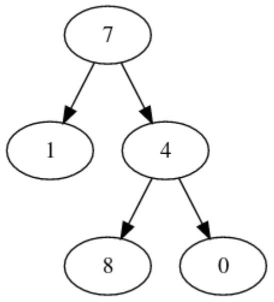
[6] You can read up on how to create and/or add key/value pairs to dictionaries in a program here: `https://www.w3schools.com/python/python_dictionaries.asp`

During the lecture, you learned about the very important data structure of trees. In this exercise we will be taking a closer look at binary trees (trees where every non leaf node has exactly two children) of integers.

One way of implementing trees in Python is by nesting dictionaries. Every nodes in the tree is either the empty dictionarys ({}, this is called a leaf of the tree) or a dictionary with the keys "value" (which for this exercise will be an integer), "left" and "right". The latter two are both dictionaries that are again either empty or trees with a value and two children.

You can find an example tree constructed in this manner in the code snippet below and a visualization of the same tree below.

```
# Example for a tree as nested dictionaries.
treeA = {"value":1, "left":{}, "right":{}}
treeB = {"value":8, "left":{}, "right":{}}
treeC = {"value":0, "left":{}, "right":{}}

treeD = {
  "value" : 4,
  "left" : treeB, "right" : treeC
}

exampleTree = {
  "value" : 7,
  "left" : treeA, "right" : treeD
}
```



A visual representation of the tree encoded as dictionaries on the left.

Write a Python function called treeMinimum that takes a (non-empty) tree as input (you can take exampleTree from above as a test case, but it needs to work for all trees constructed this way) and finds the *smallest* integer that any node in the tree carries. For example, for the tree above, your function should return 0.

Problem 6.6 (XML)

In this exercise, we will discuss the XML language family. Please answer the following questions (at most a few sentences each):

1. What is the difference between XML and HTML?
2. What roles do trees play for those two?
3. Name at least three uses of XML.

Give a short example of valid XML code that you have written yourself. Also give a small example of *incorrect* XML and explain why exactly your example is incorrect.

Problem 6.7 (Generating HTML elements)

One of the biggest advantages of programming is automation, recognising structured tasks that come up a lot and replacing human effort with computation. In these exercises we will try and automate the “boring” parts of generating simple web sites in HTML.

First, write two functions, wrapH1 and wrapP, that take one argument and return (not to be confused with "print"!) a string. The return string should be an opening tag (`<h1>` and `<p>` tags respectively), followed by the argument to the function, and then the matching closing tag.

Problem 6.8 (Generating a Website Skeleton)

Next, write a function wrapQuickFacts that takes 5 string arguments and returns a string describing a HTML table[9] listing these arguments under the categories “Name”, “Job Title”, “Date of Birth”, “Email”, and “Website”.

[9]If you need a refresher on this, you can find this structure here: https://www.w3schools.com/html/html_tables.asp

Finally, write a Python function wrapSkeleton that analogous to those in ??, return the general structure of a basic HTML page[10] as a string. The function should also take a string as an argument that is inserted between the opening and closing tag `<body>` tags in the returned string.

Problem 6.9 (Generating Complete Websites)

After we have solved the smaller problems, it is now time to combine the solutions into a (slightly) bigger program.

Using your result from ?? and ??, write a Python function generateWebsite that, given a dictionary with appropriate data[11] as input, generates (i.e. returns the HTML string that describes) the complete web site including a heading, the table and a paragraph of flavour text and saves it into a `.html` file.

Generate one of these web sites for all entires in peopleList using the functions you wrote.

[10]See also: `https://www.w3schools.com/html/html_intro.asp`

[11]You can find a file with example data here: `https://kwarc.info/teaching/IWGS/materials/people.py` You can either copy-paste these or have the file next to yours and use **import** people in your file to be able to use people.peopleList.

Chapter 5

Web Applications

In this chapter we will see how we can turn HTML pages into web-based applications that can be used without having to install additional software.

For that we discuss the basics of the World Wide Web as the client server architecture that enables such applications. Then we take up the contact form example to get an understanding how information is passed between client and server in interactive web pages. This motivates a discussion of server-side computation of web pages that can react to such information. A discussion of CSS styling shows how to make the web pages that are generated can be made visually appealing. We conclude the chapter by a discussion of client-side computation that allows making web pages interactive without recurring to the server. **Excursion:** The World Wide Web as we introduce it here is based on the Internet infrastructure and protocols. In some places it may be useful to read up on this in??.

5.1 Web Applications: The Idea

Web Applications: Using Applications without Installing

▷ **Definition 5.1.1.** A web application (also called webapp) is a program that runs on a web server and delivers its user interface as a web site consisting of programmatically generated web pages using a web browser as the client.

▷ **Example 5.1.2.** Commonly used web applications include

- ▷ `http://ebay.com`; auction pages are generated from databases.
- ▷ `http://www.weather.com`; weather information generated from weather feeds.
- ▷ `http://slashdot.org`; aggregation of news feeds/discussions.
- ▷ `http://github.com`; source code hosting and project management.
- ▷ `http://studon`; course/exam management from students records.

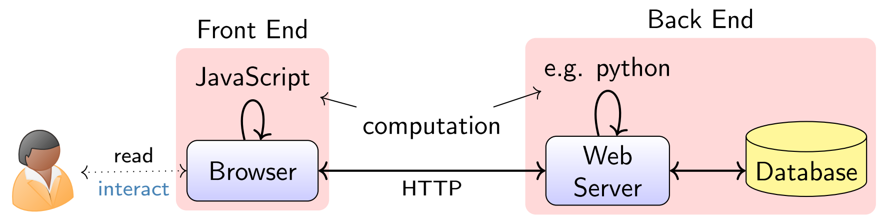
▷ **Common Traits:** Pages generated from databases and external feeds, content submission via HTML forms, file upload, dynamic HTML.

We have seen that web applications are a common way of building application software. To understand how this works let us now have a look at the components.

Anatomy of a Web Application

▷ **Definition 5.1.3.** A web application consists of two parts:

- ▷ A front end that handles the user interaction.
- ▷ A back end that stores, computes and serves the application content.



Both parts rely on (separate) computational facilities.
A database as a persistence layer is optional.

▷ **Note:** The web browser, web server, and database can

- ▷ be deployed on different computers, (high throughput)
- ▷ all run on your laptop (e.g. for development)

To understand web applications, we will first need to understand

1. how we can express web pages in HTML and (see **??**) interact with them for data input (we recap this in **??**),
2. the basics of how the World Wide Web works as a distribution framework (see **??**),
3. how we can generate HTML documents programmatically (in our case in Python; see **??**) as answer pages, and finally
4. how we can make HTML pages dynamic by client side manipulation (see **??**).

5.2 Basic Concepts of the World Wide Web

We will now present a very brief introduction into the concepts, mechanisms, and technologies that underlie the World Wide Web and thus web applications, which are our interest here.

5.2.1 Preliminaries

The WWW is the hypertext/multimedia part of the internet. It is implemented as a service on top of the internet (at the application level) based on specific protocols and markup formats for documents.

The Internet and the Web

▷ **Definition 5.2.1.** The Internet is a global computer network that connects hundreds of thousands of smaller networks.

- ▷ **Definition 5.2.2.** The World Wide Web (WWW) is an open source information space where documents and other web resources are identified by URLs, interlinked by hypertext links, and can be accessed via the Internet.
- ▷ **Intuition:** The WWW is the multimedia part of the internet, they form critical infrastructure for modern society and commerce.
- ▷ The internet/WWW is huge:

Year	Web	Deep Web	eMail
1999	21 TB	100 TB	11TB
2003	167 TB	92 PB	447 PB
2010	????	?????	?????

- ▷ We want to understand how it works. (services and scalability issues)

Given this recap we can now introduce some vocabulary to help us discuss the phenomena.

Concepts of the World Wide Web

- ▷ **Definition 5.2.3.** A web page is a document on the WWW that can include multimedia data and hyperlinks.
- ▷ **Note:** Web pages are usually marked up in in HTML.
- ▷ **Definition 5.2.4.** A web site is a collection of related web pages usually designed or controlled by the same individual or organization.
- ▷ A web site generally shares a common domain name.
- ▷ **Definition 5.2.5.** A hyperlink is a reference to data that can immediately be followed by the user or that is followed automatically by a user agent.
- ▷ **Definition 5.2.6.** A collection text documents with hyperlinks that point to text fragments within the collection is called a hypertext. The action of following hyperlinks in a hypertext is called browsing or navigating the hypertext.
- ▷ In this sense, the WWW is a multimedia hypertext.

5.2.2 Addressing on the World Wide Web

The essential idea is that the World Wide Web consists of a set of resources (documents, images, movies, etc.) that are connected by links (like a spider-web). In the WWW, the links consist of pointers to addresses of resources. To realize them, we only need addresses of resources (much as we have IP numbers as addresses to hosts on the internet).

Uniform Resource Identifier (URI), Plumbing of the Web

- ▷ **Definition 5.2.7.** A uniform resource identifier (URI) is a global identifiers of local or network-retrievable documents, or media files (web resources). URIs adhere a uniform syntax (grammar) defined in RFC-3986 [BLFM05].

A URI is made up of the following components:

- ▷ a scheme that specifies the protocol governing the resource,
- ▷ an authority: the host (authentication there) that provides the resource,
- ▷ a path in the hierarchically organized resources on the host,
- ▷ a query in the non-hierarchically organized part of the host data, and
- ▷ a fragment identifier in the resource.

▷ **Example 5.2.8.** The following are two example URIs and their component parts:

```
http://example.com:8042/over/there?name=ferret#nose
\__/   \______________/\_________/ \_________/ \__/
 |           |              |           |        |
scheme   authority        path        query  fragment
 |___    _________________|_
 /    \ /                   \
mailto:michael.kohlhase@fau.de
```

▷ **Note:** URIs only identify documents, they do not have to provide access to them (e.g. in a browser).

The definition above only specifies the structure of a URI and its functional parts. It is designed to cover and unify a lot of existing addressing schemes, including URLs (which we cover next), ISBN numbers (book identifiers), and mail addresses.
In many situations URIs still have to be entered by hand, so they can become quite unwieldy. Therefore there is a way to abbreviate them.

Relative URIs

▷ **Definition 5.2.9.** URIs can be abbreviated to relative URIs; missing parts are filled in from the context.

▷ **Example 5.2.10.** Relative URIs are more convenient to write

relative URI	abbreviates	in context
#foo	⟪current − file⟫#foo	curent file
bar.txt	file:///home/kohlhase/foo/bar.txt	file system
../bar/bar.html	http://example.org/bar/bar.html	on the web

▷ **Definition 5.2.11.** To distinguish them from relative URIs, we call URIs absolute URIs.

The important concept to grasp for relative URIs is that the missing parts can be reconstructed from the context they are found in: the document itself and how it was retrieved.
For the file system example, we are assuming that the document is a file foo.html that was loaded from the file system – under the file system URI file:///home/kohlhase/foo/foo.html – and for the web example via the URI //example.org/foo/foo.html. Note that in the last example, the relative URI ../bar/ goes up one segment of the path component (that is the meaning of ../), and specifies the file bar.html in the directory bar.

But relative URIs have another advantage over absolute URIs: they make a web page or web site easier to move. If a web site only has links using relative URIs internally, then those do not mention e.g. authority (this is recovered from context and therefore variable), so we can freely move the web-site e.g. between domains.
Note that some forms of URIs can be used for actually locating (or accessing) the identified resources, e.g. for retrieval, if the resource is a document or sending to, if the resource is a mailbox. Such URIs are called "uniform resource *locators*", all others "uniform resource *locators*".

Uniform Resource Names and Locators

▷ **Definition 5.2.12.** A uniform resource locator (URL) is a URI that gives access to a web resource, by specifying an access method or location. All other URIs are called uniform resource name (URN).

▷ **Idea:** A URN defines the identity of a resource, a URL provides a method for finding it.

▷ **Example 5.2.13.**
The following URI is a URL (try it in your browser)
http://kwarc.info/kohlhase/index.html

▷ **Example 5.2.14.** urn:isbn:978—3—540—37897—6 only identifies [Koh06] (it is in the library)

▷ URNs can be turned into URLs via a catalog service, e.g. `http://wm-urn.org/urn:isbn:978-3-540-37897-6`

▷ **Note:** URIs are one of the core features of the web infrastructure, they are considered to be the plumbing of the WWW. (direct the flow of data)

Historically, started out as URLs as short strings used for locating documents on the internet. The generalization to identifiers (and the addition of URNs) as a concept only came about when the concepts evolved and the application layer of the internet grew and needed more structure.
Note that there are two ways in URI can fail to be resource locators: first, the scheme does not support direct access (as the ISBN scheme in our example), or the scheme specifies an access method, but address does not point to an actual resource that could be accessed. Of course, the problem of "dangling links" occurs everywhere we have addressing (and change), and so we will neglect it from our discussion. In practice, the URL/URN distinction is mainly driven by the scheme part of a URI, which specifies the access/identification scheme.

Internationalized Resource Identifiers

▷ *Remark 5.2.15.* URIs are ASCII strings.

▷ **Problem:** This is awkward e.g. for *France Télécom*, worse in Asia.

▷ **Solution?:** Use unicode! (no, too young/unsafe)

▷ **Definition 5.2.16.** Internationalized resource identifiers (IRIs) extend the ASCII-based URIs to the universal character set.

▷ **Definition 5.2.17.** URI-encoding maps non-ASCII characters to ASCII strings:

1. Map each character to its UTF-8 representation.
2. Represent each byte of the UTF-8 representation by three characters.
3. The first character is the percent sign (%),
4. and the other two characters are the hexadecimal representation of the byte.

URI-decoding is the dual operation.

▷ **Example 5.2.18.** The letter "ł" ($\mathrm{U}+142$) would be represented as `%C5%82`.

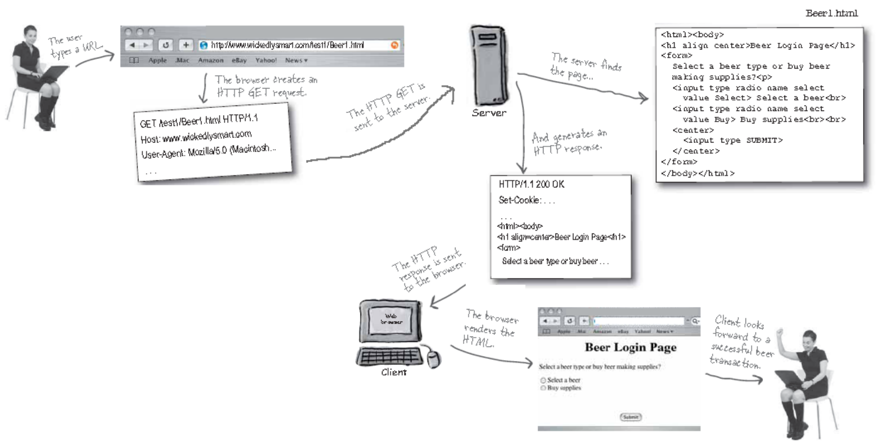
▷ **Example 5.2.19.** `http://www.Übergrößen.de` becomes `http://www.%C3%9Cbergr%C3%B6%C3%9Fen.de`

▷ *Remark 5.2.20.* Your browser can still show the URI-decoded version (so you can read it)

5.2.3 Running the World Wide Web

The infrastructure of the WWW relies on a client-server architecture, where the servers (called web servers) provide documents and the clients (usually web browsers) present the documents to the (human) users. Clients and servers communicate via the HTTPs and HTTPSs protocols. We give an overview via a concrete example before we go into details.

The World Wide Web as a Client/Server System


The user types a URL.
http://www.wickedlysmart.com/test1/Beer1.html
Apple .Mac Amazon eBay Yahoo! News
The browser creates an HTTP GET request.
The HTTP GET is sent to the server.
Server
The server finds the page...
Beer1.html
And generates an HTTP response.
The HTTP response is sent to the browser.
Web browser
Client
The browser renders the HTML.
Beer Login Page
Select a beer type or buy beer making supplies?
Select a beer
Buy supplies
Submit
Client looks forward to a successful beer transaction.

```
GET /test1/Beer1.html HTTP/1.1
Host: www.wickedlysmart.com
User-Agent: Mozilla/5.0 (Macintosh...
...
```

```
<html><body>
<h1 align center>Beer Login Page</h1>
<form>
  Select a beer type or buy beer
  making supplies?<p>
  <input type radio name select
    value Select> Select a beer<br>
  <input type radio name select
    value Buy> Buy supplies<br><br>
  <center>
    <input type SUBMIT>
  </center>
</form>
</body></html>
```

```
HTTP/1.1 200 OK
Set-Cookie: ...
...
<html><body>
<h1 align=center>Beer Login Page</h1>
<form>
 Select a beer type or buy beer ...
```

The web browser communicates with the web server through a specialized protocol, the hypertext transfer protocol, which we cover now.

HTTP: Hypertext Transfer Protocol

▷ **Definition 5.2.21.** The Hypertext Transfer Protocol (HTTP) is an application layer protocol for distributed, collaborative, hypermedia information systems.

▷ June 1999: HTTP/1.1 is defined in RFC 2616 [Fie+99].

▷ **Preview/Recap:** HTTP is used by a client (called user agent) to access web web resources (addressed by uniform resource locators (URLs)) via a HTTP request. The web server answers by supplying the web resource (and metadata).

▷ **Definition 5.2.22.** Most important HTTP request methods. (5 more less prominent)

GET	Requests a representation of the specified resource.	safe
PUT	Uploads a representation of the specified resource.	idempotent
DELETE	Deletes the specified resource.	idempotent
POST	Submits data to be processed (e.g., from a web form) to the identified resource.	

▷ **Definition 5.2.23.** We call a HTTP request safe, iff it does not change the state in the web server. (except for server logs, counters,. . . ; no side effects)

▷ **Definition 5.2.24.** We call a HTTP request idempotent, iff executing it twice has the same effect as executing it once.

▷ HTTP is a stateless protocol. (very memory efficient for the server.)

Finally, we come to the last component, the web server, which is responsible for providing the web page requested by the user.

Web Servers

▷ **Definition 5.2.25.** A web server is a network program (a server in a client server architecture of the WWW) that delivers web resources to and receives content from clients via the Hypertext Transfer Protocol (HTTP).

▷ **Example 5.2.26 (Common Web Servers).**

- ▷ `apache` is an open source web server that serves about 50% of the WWW.
- ▷ `nginx` is a lightweight open source web server. (ca. 35%)
- ▷ `IIS` is a proprietary web server provided by Microsoft Inc.

▷ **Definition 5.2.27.** A web server can host – i.e serve web resources for multiple domains (via configurable hostnames) that can be addressed in the authority components of URLs. This usually includes the special hostname localhost which is interpreted as "this computer".

▷ Even though web servers are very complex software systems, they come preinstalled on most `UNIX` systems and can be downloaded for `MSWindows` [Xam].

Now that we have seen all the components we fortify our intuition of what actually goes down the net by tracing the HTTP messages.

Example: An HTTP request in real life

▷ Send off a GET request for `http://www.nowhere123.com/doc/index.html`

```
GET /docs/index.html HTTP/1.1
Host: www.nowhere123.com
Accept: image/gif, image/jpeg, */*
Accept-Language: en-us
Accept-Encoding: gzip, deflate
User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
(blank line)
```

▷ The response from the server

```
HTTP/1.1 200 OK
Date: Sun, 18 Oct 2009 08:56:53 GMT
Server: Apache/2.2.14 (Win32)
Last-Modified: Sat, 20 Nov 2004 07:16:26 GMT
ETag: "10000000565a5-2c-3e94b66c2e680"
Accept-Ranges: bytes
Content-Length: 44
Connection: close
Content-Type: text/html
X-Pad: avoid browser bug

<html><body><h1>It works!</h1></body></html>
```

▷ **Note:** As you can seen, these are clear-text messages that go over an unprotected network. A consequence is that everyone on this network can intercept this communication and see what you are doing/reading/watching.

5.3 Recap: HTML Forms Data Transmission

EdN:2 The first two requirement for web applications above are already met by HTML in terms of HTML forms (see slide 120 ff.). Let us recap and extend[2]

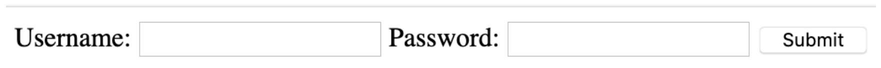
Recap HTML Forms: Submitting Data to the Web Server

▷ **Recall:** HTML forms collect data via named input elements, the submit event triggers a HTTP request to the URL specified in the action attribute.

▷ **Example 5.3.1.** Forms contain input fields and explanations.

```
<form name="input" action="login.html" method="get">
  Username: <input type="text" name="user"/>
  Password: <input type="password" name="pass"/>
  <input type="submit" value="Submit"/>
</form>
```

yields the following in a web browser:



Pressing the submit button activates a HTTP GET request to the URL login.html?user=⟪name⟫&pass=⟪passwd⟫

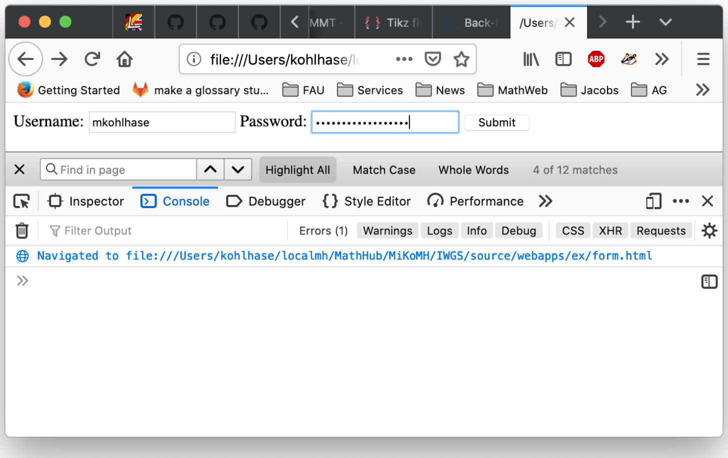
▷ ⚠ Never use the GET method for submitting passwords (see below)

[2]EdNote: continue

We can now use the tools any modern browser supplies to check up on this claim. In fact, using the browser tools is essential for advanced web development. Here we use the web console, that monitors any activity, to check upon what really happens when we interact with the web page.

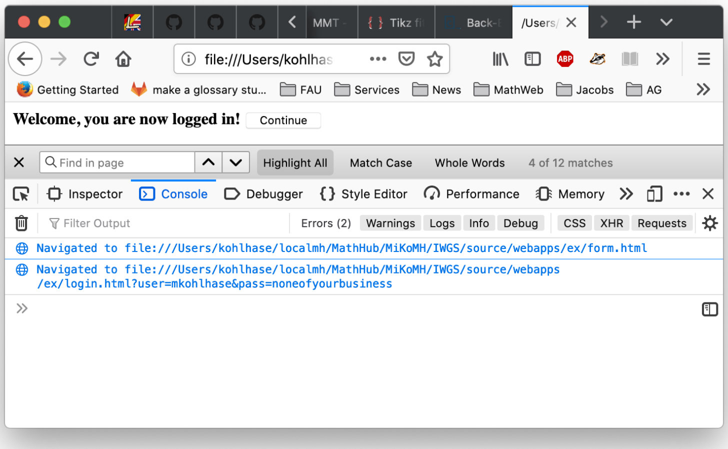
Checking up on the Transmission

- ▷ Let's verify the claims above using browser tools (here the web console)
- ▷ Loading the file and filling in the form: (console logs file URI)


file:///Users/kohlhase/l
Username: mkohlhase Password: •••••••••••••••••• Submit
Inspector Console Debugger Style Editor Performance
Errors (1) Warnings Logs Info Debug CSS XHR Requests

```
Navigated to file:///Users/kohlhase/localmh/MathHub/MiKoMH/IWGS/source/webapps/ex/form.html
```

- ▷ After submitting the form: (console logs the HTTP request)


file:///Users/kohlhas
Welcome, you are now logged in! Continue
Inspector Console Debugger Style Editor Performance Memory
Errors (2) Warnings Logs Info Debug CSS XHR Requests

```
Navigated to file:///Users/kohlhase/localmh/MathHub/MiKoMH/IWGS/source/webapps/ex/form.html
Navigated to file:///Users/kohlhase/localmh/MathHub/MiKoMH/IWGS/source/webapps
/ex/login.html?user=mkohlhase&pass=noneofyourbusiness
```

A side effect of re-playing our development in the browser is that we see another type of input field: A password field, which hides user input from un-authorized eyes. We also see that the GET request incorporates the form data which contains the password into the URI of the request, which is visible to everyone on the web. We will come back to this problem later.

Let us now look at the data transmission mechanism in more detail to see what is actually transmitted and how.

HTML Forms and Form Data Transmission

▷ We specify the HTTP communication of HTML forms in detail.

▷ **Definition 5.3.2.** The HTML form element groups the layout and input elements:

- ▷ <form action="$\langle\!\langle$URI$\rangle\!\rangle$" method="$\langle\!\langle$req$\rangle\!\rangle$"> specifies the form action in terms of a HTTP request $\langle\!\langle$req$\rangle\!\rangle$ to the URI $\langle\!\langle$URI$\rangle\!\rangle$.
- ▷ The form data consists of a string $\langle\!\langle$data$\rangle\!\rangle$ of the form $n_1{=}v_1\&\cdots\&n_k{=}v_k$, where
  - ▷ n_i are the values of the name attributes of the input fields
  - ▷ and v_i are their values at the time of submission.
- ▷ <input type="submit" .../> triggers the form action: it composes a HTTP request
  - ▷ If $\langle\!\langle$req$\rangle\!\rangle$ is get (the default), then the browser issues a GET request $\langle\!\langle$URI$\rangle\!\rangle$?$\langle\!\langle$data$\rangle\!\rangle$.
  - ▷ If $\langle\!\langle$req$\rangle\!\rangle$ is post, then the browser issues a POST request to $\langle\!\langle$URI$\rangle\!\rangle$ with document content $\langle\!\langle$data$\rangle\!\rangle$.

▷ We now also understand the form action, but should we use GET or POST.

To understand whether we should use the GET or POST methods, we have to look into the details, which we will now summarize.

Practical Differences between HTTP GET and POST

▷ **Using GET vs. POST in HTML Forms:**

	GET	POST
Caching	possible	never
Browser History	Yes	never
Bookmarking	Yes	No
Change Server Data	No	Yes
Size Restrictions	$\leq 2KB$	No
Encryption	No	HTTPS

▷ **Upshot:** HTTP GET is more convenient, but less potent.

▷ ⚠ Always use POST for sensitive data! (passwords, personal data, etc.)
GET data is part of the URI and thus unencrypted, POST data via HTTPS is.

5.4 Generating HTML on the Server

As the WWW is based on a client server architecture, computation in web applications can be executed either on the client (the web browser) or the server (the web server). For both we have a special technology; we start with computation on the web server.

Server-Side Scripting: Programming Web pages

- ▷ **Idea:** Why write HTML pages if we can also program them! (easy to do)
- ▷ **Definition 5.4.1.** A server-side scripting framework is a web server extension that generates web pages upon HTTP requests.
- ▷ **Example 5.4.2.** `perl` is a scripting language with good string manipulation facilities. PERL CGI is an early server-side scripting framework based on this.
- ▷ **Example 5.4.3.** Python is a scripting language with good string manipulation facilities. And bottle WSGI is a simple but powerful server-side scripting framework based on this.
- ▷ **Observation:** Server-side scripting frameworks allow to make use of external resources (e.g. databases or data feeds) and computational services during web page generation.
- ▷ **Observation:** A server-side scripting framework solves two problems:
  1. making the development of functionality that generates HTML pages convenient and efficient, usually via a template engine, and
  2. binding such functionality to URLs the routes, we call this routing.

We will look at the second problem: routing first. There is a dedicated Python library for that.

5.4.1 Routing and Argument Passing in Bottle

We wil now introduce the bottle library, which supplies a lightweight web server and server-side scripting framework implemented in Python. It is already installed on the JuptyerLab cloud IDE at `http://jupyter.kwarc.info`. To install it on your laptop, just type pip install bottle in a shell.

The Web Server and Routing in Bottle WSGI

- ▷ **Definition 5.4.4.** Serverside routing (or simply routing) is the process by which a web server connects a HTTP request to a function (called the route function) that provides a web resource. A single URI path/route function pair is called a route.
- ▷ The bottle WSGI library supplies a simple Python web server and routing.
  - ▷ The run(⟪keys⟫) function starts the web server with the configuration in ⟪keys⟫.
  - ▷ The @route decorator connects path components to Python function that return strings. Decorators change functions. A decorator @route(⟪path⟫) augments the following function f to answer to HTTP requests to the ⟪path⟫ and return f's return value.

▷ **Example 5.4.5 (A Hello World route).** . . . for localhost on port 8080

```
from bottle import route, run

@route('/hello')
def hello():
    return "Hello IWGS!"

run(host='localhost', port=8080, debug=True)
```

This web server answers to HTTP GET requests for the URL `http://localhost:8080/hello`

Let us understand **??** line by line: The first line imports the library. The second establishes a route with the name hello and binds it to the Python function hello in line 3 and 4. The last line configures the bottle web server: it serves content via the HTTP protocol for localhost on port 8080.

So, if we run the program from **??**, then we obtain a web server that will answer HTTP GET requests to the URL `http://localhost:8080/hello` with a HTTP answer with the content Hello IWGS!.

To keep the example simple, we have only returned a text string; A realistic application would have generated a full HTML page (see below).
In the last line of **??**, we have also configured the bottle web server to use "debug mode", which is very helpful during early development.

In this mode, the bottle web server is much more verbose and provides helpful debugging information whenever an error occurs. It also disables some optimisations that might get in your way and adds some checks that warn you about possible misconfiguration.

Note that debug mode should be disabled in a production server for efficiency.
But we can do more with routes!

Dynamic Routes in Bottle

▷ **Definition 5.4.6.** A dynamic route is a route annotation that contains named wildcards, which can be picked up in the route function.

▷ **Example 5.4.7.** Multiple @route annotations per route function f are allowed $\rightsquigarrow$ the web application uses f to answer multiple URLs.

```
@route('/')
@route('/hello/<name>')
def greet(name='Stranger'):
    return (f'Hello {name}, how are you?')
```

With the wildcard <name> we can bind the route function greet to all paths and via its argument name and customize the greeting.

Concretely: A HTTP GET request to

- ▷ `http://localhost` is answered with Hello Stranger, how are you?.
- ▷ `http://localhost/hello/MiKo` is answered with Hello MiKo, how are you?.

Requests to e.g `http://localhost/hello` or `http://localhost/hello/prof/kohlhase` lead to errors. (404: not found)

Often we want to have more control over the routes. We can get that by filters, which can involve data types and/or regular expressions.

Restricting Dynamic Routes

▷ **Definition 5.4.8.** A dynamic route can be restricted by a route filter to make it more selective.

▷ **Example 5.4.9 (Concrete Filters).** We use :int for integers and :re:⟪regex⟫ for regular expressions

```
@route('/tel/<id:int>') # local number
@route('/tel/<num:re:^\+[1−9]{1}[0−9]{3,14}$>') # international
```

Different route filters allow to classify paths and treat them differently.

▷ **Note:** Multiple named wildcards are also possible, in a dynamic route; with and without filters

▷ **Example 5.4.10 (A route with two wildcards).**

```
@route('/<action>/<user:re:[a−z]+>') # matches /follow/miko
def user_api(action, user):
    ...
```

We have already seen above that we want to use HTTP GET and POST request for different facets of transmitting HTML form data to the web server. This is supported by bottle WSGI in two ways: we can specify the HTTP method of a route and we have access to the form data (and other aspects of the request).

Method-Specific Routes: HTTP GET and POST

▷ **Definition 5.4.11.** The @route decorator takes a method keyword to specify the HTTP request method to be answered. (HTTP GET is the default)

- ▹ @get(⟪path⟫) abbreviates @route(⟪path⟫,method="GET")
- ▹ @post(⟪path⟫) abbreviates @route(⟪path⟫,method="POST")

▷ **Example 5.4.12 (Login 1).** Managing logins with HTTP GET and POST.

```
from bottle import get, post, request # or route

@get('/login') # or @route('/login')
def login():
    return '''
        <form action="/login" method="post">
            Username: <input name="username" type="text" />
            Password: <input name="password" type="password" />
            <input value="Login" type="submit" />
        </form>
    '''
```

▷ **Note:** We can also have a POST request to the same path; we use that for handling the form data transmitted by the POST action on submit. (up next)

Recall that we have already seen most of this in slide 160. The only new thing is that we return the HTML as a string in the route function as a request to a HTTP GET request. Now comes the interesting part: the form uses the POST method in the form action and we have to specify a route for that. Recall from **??** that this allows for encrypted transmission, so we are less naive than our solution from slide 160.

Bottle Request: Dealing with POST Data

▷ **Recall:** from a HTML form we get a GET or POST request with form data $n_1{=}v_1\&\cdots\&n_k{=}v_k$ (here user=mkohlhase&login=noneofyourbusiness)

▷ Bottle WSGI provides the request object for dealing with HTTP request data.

▷ **Example 5.4.13 (Login 2).** Continuing from **??**: we parse the request transmitted request and check password information:

```
@post('/login') # or @route('/login', method='POST')
def do_login():
    username = request.forms.get('username')
    password = request.forms.get('password')
    if check_login(username, password):
        return "<p>Your login information was correct.</p>"
    else:
        return "<p>Login failed.</p>"
```

We assume a Python function check_login that checks authentication credential and authenticator, and keeps a list of logged in users.

The main new thing in **??** is that we use the request.forms.get method to query the request object that comes with the HTTP request triggering the route for the form data.

5.4.2 Templating in Python via STPL

In IWGS, we use Python for programming, so let us see how we would generate HTML pages in Python.

What would we do in Python

▷ **Example 5.4.14 (HTML Hello World in Python).**

```
print("<html>")
print("<body>Hello world</body>")
print("</html>")
```

▷ **Problem 1:** Most web page content is static (page head, text blocks, etc.)

▷ **Example 5.4.15 (Python Solution).** . . . use Python functions:

```
def htmlpage (t,b):
    f"<html><head><title>{t}</title></head><body>{b}</body></html>"
```

```
htmlpage("Hello","Hello IWGS")
```

▷ **Problem 2:** If HTML markup dominates, want to use a HTML editor (mode),

▹ e.g. for HTML syntax highlighting/indentation/completion/checking

▷ **Idea:** Embed program snippets into HTML. (only execute these, copy rest)

We will now formalize and toolify the idea of "embedding code into HTML". What comes out of this idea is called "templating". It exists in many forms, and in most programming languages.

Template Processing for HTML

▷ **Definition 5.4.16.** A template engine (or template processor) for a document format F is a program that transforms templates, i.e. strings or files (a template file) ith a mixture of program constructs and F markup, into a F strings or F documents by executing the program constructs in the template (template processing).

▷ **Note:** No program code is left in the resulting web page after generation. (important security concern)

▷ **Remark:** We will be most interested in HTML template engines.

▷ **Observation:** We can turn a template engine into a server-side scripting framework by employing the URIs of template files on a server as routes and extending the web server by template processing.

▷ **Example 5.4.17.** PHP (originally "Programmable Home Page Tools") is a very successful server-side scripting framework following this model.

Naturally, Python comes with a template engine in fact multiple ones. We will use the one from the bottle web application framework for IWGS.

stpl: the "Simple Template Engine" from Bottle

▷ **Definition 5.4.18.** Bottle WSGI supplies the template engine stpl (Simple Template Engine) that processes the STPL (Simple Template Language) format. (documentation at [STPL])

▷ **Definition 5.4.19.** A template engine for a document format F is a program that transforms templates, i.e. strings or files through a mixture of program constructs and F markup, into F-strings or F-documents by executing the program constructs in the template (template processing).

▷ stpl uses the template function for template processing and {{...}} to embed program objects into a template; it returns a formatted unicode string.

```
>>> template('Hello {{name}}!', name='World')
u'Hello World!'

>>> my_dict={'number': '123', 'street': 'Fake St.', 'city': 'Fakeville'}
>>> template('I live at {{number}} {{street}}, {{city}}', **my_dict)
```

```
u'I live at 123 Fake St., Fakeville'
```

The stpl template function is a powerful enabling basic functionality in Python, but it does not satisfy our goal of writing “HTML with embedded Python”. Fortunately, that can easily be built on top of the template functionality:

stpl Syntax and Template Files

- **But what about. . . :** HTML files with embedded Python?
- stpl uses template files (extension .tpl) for that.
- **Definition 5.4.20.** A stpl template file mixes HTML with stpl python:
  - stpl python is exactly like Python but ignores indentation and closes bodies with end instead.
  - stpl python can be embedded into the HTML as
    - a code lines starting with a %,
    - a code blocks surrounded with <% and %>, and
    - an expressions {{⟨⟨exp⟩⟩}} as long as ⟨⟨exp⟩⟩ evaluates to a string.
- **Example 5.4.21.** Two template files

```
<!-- next: a line of python code -->
% course = "Informatische werkzeuge ..."
<p>Some plain text in between</p>
<%
  # A block of python code
  course = name.title().strip()
%>
<p>More plain text</p>
```

```
<ul>
  % for item in basket:
    <li>{{item}}</li>
  % end
</ul>
```

So now, we have template files. But experience shows that template files can be quite redundant; in fact, the better designed the web site we want to to create, the more fragments of the template files we want to reuse in multiple places – with and without adaptions to the particular use case.

Template Functions

- **Definition 5.4.22.** stpl python supplies the template functions
  1. include(⟨⟨tpl⟩⟩,⟨⟨vars⟩⟩), where ⟨⟨tpl⟩⟩ is another template file and ⟨⟨vars⟩⟩ a set of variable declarations (for ⟨⟨tpl⟩⟩).
  2. defined(⟨⟨var⟩⟩) for checking definedness ⟨⟨var⟩⟩
  3. get(⟨⟨var⟩⟩,⟨⟨default⟩⟩): return the value of ⟨⟨var⟩⟩, or ⟨⟨default⟩⟩.
  4. setdefault(⟨⟨name⟩⟩,⟨⟨val⟩⟩)
- **Example 5.4.23 (Including Header and Footer in a template).** In a coherent web site, the web pages often share common header and footer parts. Realize this

via the following page template:

```
% include('header.tpl', title='Page Title')
 ... Page Content ...
% include('footer.tpl')
```

▷ **Example 5.4.24 (Dealing with Variables and Defaults).**

```
% setdefault('text', 'No Text')
<h1>{{get('title', 'No Title')}}</h1>
<p> {{ text }} </p>
% if defined('author'):
  <p>By {{ author }}</p>
% end
```

5.4.3 Completing the Contact Form

We are now equipped to finish the contact form example

We now come back to our worked HTML example: the contact form from above. Here is the current state:

Back to our Contact Form (Current State)

▷ A contact form and message receipt (communicate via HTTP requests)

contact4.html

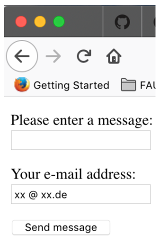
```
<title>Contact</title>
<form action="contact−after.html">
  <h2>Please enter a message:</h2>
    <input name="msg" type="text"/>
  <h3>Your e−mail address:</h3>
  <input name="addr" type="text"
           value="xx @ xx.de"/>
  <br/>
  <input type="submit"
          value="Send message"/>
</form>
```

```
GET contact−after.html?
    msg=Hi;addr=foo@bar.de
```

contact−after.html

```
<title>
  Contact − Message Confirmed
</title>
<form action="contact4.html">
  <h2>
    Your message has been submitted!
  </h2>
  <input type="submit"
          value="Continue"/>
</form>
```

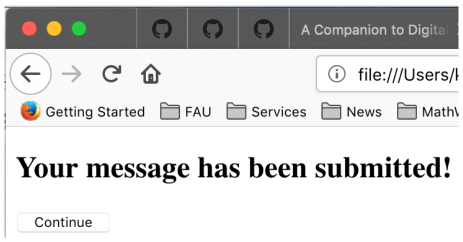
```
GET contact.html
```



Please enter a message:

Your e-mail address:

xx @ xx.de



Your message has been submitted!

▷ **Problem:** The answer is a static HTML document independent of form data.

▷ **Solution:** Generate the answer programmatically using the form data. (up next)

There are two great flaws in the current state of the contact form:

1. The "receipt page" contact−after.html is static and does not take the data it receives from the contact form into account. It would be polite to give some record on what happened. We can fix this using bottle WSGI using the methods we just learned.

2. Nothing actually happens with the message. It should be either entered into an internal message queue in a database0 or ticketing system, or fed into an e-mail to a sales person. As we do not have access to the first, we will just use a Python library to send an e-mail programmatically.

Completing the Contact Form

▷ bottle WSGI has functionality (request.GET and request.POST) to decode the form data from a HTTP request. (so we do not have to worry about the details)

▷ **Example 5.4.25 (Submitting a Contact Form).** We use a new route for contact−form−after.html with a corresponding template file:

contact.py

```
from bottle import route, run, debug,
                      template, request, get

@get('/contact−after.html')
def new_item():
    data = {'msg': request.GET.msg.strip(),
            'addr': request.GET.addr.strip()}
    send−contact−email(addr,msg)
    return template('contact−after',**data)

run(host="localhost", port=8080)
```

contact−after.tpl

```
<p>Message submitted!</p>
<table>
  <tr>
    <td>Return Address:</td>
    <td>{{addr}}</td>
  </tr>
  <tr>
    <td>Message Sent:</td>
    <td>{{msg}}</td>
  </tr>
</table>
```

Fortunately, the only remaining part: actually sending off an e-mail to the specified mailbox is very easy: using the smtplib library we just create an e-mail message object, and then specify all the components.

Sending off the e-mail

▷ We still need to implement the send−contact−email function, . . .

▷ Fortunately, there is a Python package for that: smtplib, which makes this relatively easy. (SMTP $\hat{=}$ Simple Mail Transfer Protocol")

▷ **Example 5.4.26 (Continuing).**

```
import smtplib
from email.message import EmailMessage

def send−contact−email (addr, text)
    msg = EmailMessage()
    msg.set_content(text)
    msg['Subject'] = 'Contact Form Result'
    msg['From'] = info@example.org
    msg['To'] = addr
    s = smtplib.SMTP('smtp.gmail.com', 587)
    s.send_message(msg)
    s.quit()
```

Actually, this does not quite work yet as google requires authentication and encryption, . . . ; (google for "python smtplib gmail")

Once we have the e-mail message object msg, we open a "SMTP connection" s send the message via its send_message method and close the connection by s.quit()). Again, the Python library hides all the gory details of the SMTP protocol.

5.5 Exercises

In the exercises in this section, we will take a closer look at web applications, templating and HTML routing. Concretely, we will be using the Bottle framework[1], as demonstrated in the course.

Problem 5.1 (Hello WebApp World)

Set up the following routes (pairs of URLs and Python functions that return strings):

- A client navigating to the root directory of your webapp (`"/"`) should receive a standard `"Hello World"` message.
- A client navigating to `"/hello/<name>"` should find a greeting message personalised with the name given in the URL (`"/hello/Philipp"` greets Philipp, `"/hello/Jonas"` greets Jonas, . . .).

  Have at least one name (your choice) be treated differently than all others (for example: all names get a nice message by default, but the name `"GrumpyCat"` gets an annoyed message).

Problem 5.2 (Routing a HTML form)

In the following exercises, we want to build a small, but complete (!) web application where users can submit reviews for media (books, movies, . . .) that get saved into a "database" and can be viewed later. A lot of these exercises will ask for HTML or Python code that is similar to previous exercises. The challenge is to integrate the familiar code into the new context of web-applications and the bottle framework.

Add a `"/submit"` route to your web app that delivers a HTML form. The form should at least have `input` elements for a title (text), a synopsis (text) and a rating from 1 to 5 (number or radio buttons).

When the submit button (which also needs to be included in the form) is pressed, the form should redirect the user to the `"/submitted"` route (see **??**) via the `action` attribute. Make sure that the method used for this is a GET request (how can you specify this?).

Problem 5.3 (HTML GET Requests)

Now, add a route specifically for GET requests at `"/submitted"` (the target of your submit-redirect from **??**). Since we're dealing with a GET request, the information submitted through the form will be encoded in the URL.

The corresponding function should read the title, synopsis and rating from the HTML request (see the bottle documentation or the course materials for examples) and append them to a file[2] called `database.txt`[3].

[1]See the documentation of bottle for reference: `https://bottlepy.org/docs/dev/tutorial.html`

[2]Even though the function must ultimately *return* a string from which a HTML page is constructed, it can write to a file before doing so as a side effect.

[3]This file will appear next to your other files in your `pythonAnywhere` directory. It is enough to simply append to the file, Python will create the file if it does not exist yet.

You can append one line of text to the file per entry in the database, with the title, synopsis and rating seperated by semicolons, for example.

Problem 5.4 (Displaying the database)

Finally, add a `"/database"` route to your web app that reads the aforementioned database file (`database.txt`) and displays its contents as a HTML page. This page should contain a heading and an unordered list (the `<ul>` element), in which each entry in the database (= line in the file) is one list item (`<li>` element).

Problem 5.5 (Simple CSS)

It is a well-known fact that nobody likes to buy from a pizza place that only uses plain HTML on their website. So now, we will improve upon the website from **??**.

Create an external stylesheet (in a CSS file called `styles.css`) to change the look of your website. You can load this stylesheet by placing the following `head`-element into your website's `html`-element:

```
<head>
    <link rel="stylesheet" href="styles.css">
</head>
```

You can make this stylesheet as elaborate as you like. However, at least the following style changes should be implemented by your stylesheet:

- Center the heading.
- Give the `<body>` of your website a `background-color`.
- Set the `font-family` of all text to "Verdana".
- Set the font size of your descriptive text to 14.

Chapter 6

Frontend Technologies

We introduce three important concepts for building modern web front ends for web applications:

1. Client-side computation: manipulating the browser DOM via JavaScript.
2. Cascading Stylesheets (CSS) for styling the layout of HTML (and XML).
3. The jQuery library: a symbiosis of JavaScript and CSS ideas to make JavaScript coding easier and more efficient.

6.1 Dynamic HTML: Client-side Manipulation of HTML Documents

We now turn to client-side computation:
One of the main advantages of moving documents from their traditional ink-on-paper form into an electronic form is that we can interact with them more directly. But there are many more interactions than just browsing hyperlinks we can think of: adding margin notes, looking up definitions or translations of particular words, or copy-and-pasting mathematical formulae into a computer algebra system. All of them (and many more) can be made, if we make documents programmable. For that we need three ingredients:

i) a machine-accessible representation of the document structure, and

ii) a program interpreter in the web browser, and

iii) a way to send programs to the browser together with the document.

We will sketch the WWW solution to this in the following.

To understand client-side computation, we first need to understand the way browsers render HTML pages.

Background: Rendering Pipeline in browsers

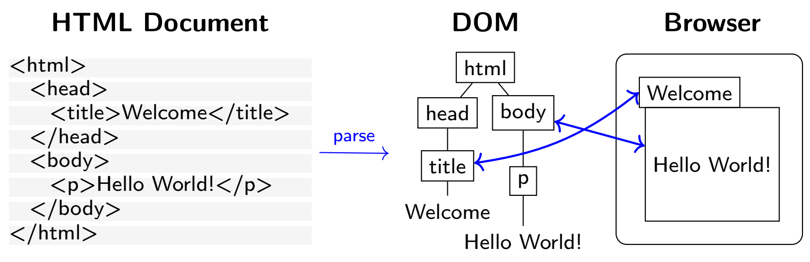
▷ **Observation:** The nested markup codes turn HTML documents into trees.

▷ **Definition 6.1.1.** The document object model (DOM) is a data structure for the HTML document tree together with a standardized set of access methods.

▷ **Rendering Pipeline:** Rendering a web page proceeds in three steps

1. the browser receives a HTML document,
2. parses it into an internal data structure, the DOM,

3. which is then painted to the screen. (repaint whenever DOM changes)



```
<html>
  <head>
    <title>Welcome</title>
  </head>
  <body>
    <p>Hello World!</p>
  </body>
</html>
```

The DOM is notified of any user events. (resizing, clicks, hover,. . .)

The most important concept to grasp here is the tight synchronization between the DOM and the screen. The DOM is first established by parsing (i.e. interpreting) the input, and is synchronized with the browser UI and document viewport. As the DOM is persistent and synchronized, any change in the DOM is directly mirrored in the browser viewpoint, as a consequence we only need to change the DOM to change its presentation in the browser. This exactly is the purpose of the client side scripting language, which we will go into next.

6.1.1 JavaScript in HTML

Dynamic HTML

▷ **Definition 6.1.2.** We call a web page dynamic, if its presentation can change without the web browser loading new content.

▷ **Idea:** Generate parts of the web page dynamically by manipulating the DOM.

▷ **Definition 6.1.3.** JavaScript is an object-oriented scripting language mostly used to enable programmatic access to the DOM in a web browser.

▷ JavaScript is standardized by ECMA in [Ecm].

▷ **Example 6.1.4.** We write the some text into HTML DOM.

```
<html>
 <head>
  <script type="text/javascript">document.write("Dynamic HTML!");</script>
 </head>
 <body><!-- nothing here; will be added by the script later --></body>
</html>
```

▷ **Application:** Write “gmail” or “google docs” as JavaScript enhanced web applications. (client-side computation for immediate reaction)

▷ **Current Megatrend:** Computation in the “cloud”, browsers (or “apps”) as user interfaces.

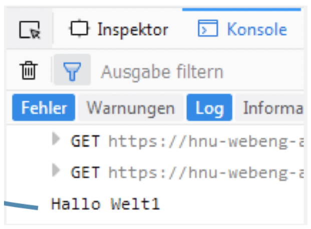
The example above already shows a JavaScript command: document.write, which replaces the content of the <body> element with its argument – this is only useful for testing and debugging purposes.

Current web applications include simple office software (word processors, online spreadsheets, and presentation tools), but can also include more advanced applications such as project management, computer-aided design, video editing and point-of-sale. These are only possible if we carefully balance the effects of server-side and client-side computation. The former is needed for computational resources and data persistence (data can be stored on the server) and the latter to keep personal information near the user and react to local context (e.g. screen size).
Here are three browser level functions that can be used for user interaction (and finer debugging as they do not change the DOM).

Browser-level JavaScript functions: 1

▷ **Example 6.1.5 (Logging to the browser console).**

```
console.log("hello IWGS")
```



```
GET https://hnu-webeng-a
GET https://hnu-webeng-a
Hallo Welt1
```

The function console.log writes its argument into the console of the web browser.
It is primarily used for debugging the source code of a web page.

Example 6.1.6. If we want to know whether a function square has been executed we add calls to console.log like this:

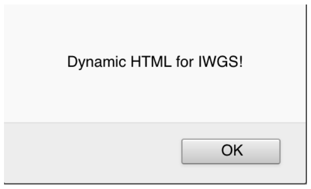
```
function square (n) {
  console.log ("entered function square with argument " + n);
  return (n ∗ n);
  console.log ("exited function square with result " + n ∗ n);
  }
```

In the console we can check whether the content contains e.g. entered function square and moreover whether argument and value are as expected.

Browser-level JavaScript functions: 2

▷ **Example 6.1.7 (Raising a Popup).**

```
alert("Dynamic HTML for IWGS!")
```

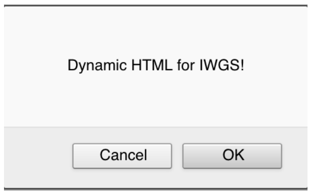


The function alert creates a popup that contains the argument.

Browser-level JavaScript functions: 3

▷ **Example 6.1.8 (Asking for Confirmation).**

```
var returnvalue = confirm("Dynamic HTML for IWGS!")
```



The function confirm creates a popup that contains the argument and a confirmation/cancel button pair and returns the corresponding Boolean value.
If the user clicks on the confirmation button, the returned value will be false and true for the cancel button.

Example 6.1.9. You can play with this in the following frizzle:

```
<html>
<head>
    <title>confirm</title>
    <script src="./client-js/jquery-3.6.4.min.js" type="application/javascript"></script>
    <style>
        .emph{
            color: blue;
        }
        .code{
            font-size: 110%;
        }
    </style>
</head>
<body>
    <h2>Live Demo of the JavaScript <span class="code emph">confirm</span> Function</h2>

    <textarea id="output" style="width:400px">
    </textarea>
    <textarea id="code" style="width:400px; height:400px">
    </textarea>
    <p>
      Click <button onclick="openPopup()">here</button> to execute
      the <span class="code">confirm</span> function again!
    </p>
    <p>
        Show <button onclick="showCode()">source code</button>
    </p>

    <script type="application/javascript">
        function openPopup(){
            console.log("executed openPopup function");
            var output="";
            var returnValue=confirm ("Hello World!");
            if(returnValue==true){
                output="You clicked the OK button!" + "(return value: " + returnValue + ")";

            } else {
                output="You clicked the Cancel button!" + "(return value: " + returnValue
+ ")";
            }
            console.log(output);
            $("#output").html(output);
            $("p").show();
        }
        openPopup();

        function showCode(){
            console.log("executed showCode function");
            var func=openPopup.toString();
            //alert(func);
            $("#code").html(func);
        }
    </script>
</body>
</html>
```

JavaScript is a client side programming language, that means that the programs are delivered to the browser with the HTML documents and is executed in the browser. There are essentially three ways of embedding JavaScript into HTML documents:

Embedding JavaScript into HTML

▷ In a <script> element in HTML, e.g.

```
<script type="text/javascript">
   function sayHello() { console.log('Hello IWGS!'); }
</script>
```

▷ External JavaScript file via a <**script**> element with src attribute:

```
<script type="text/javascript" src="../js/foo.js"/>
```

Advantage: HTML and JavaScript code are clearly separated.

▷ In event handler attributes of various HTML elements, e.g.

```
<input type="button" value="Hallo" onclick="alert('Hello IWGS')"/>
```

A related – and equally important – question is, *when* the various embedded JavaScript fragments are executed. Here, the situation is more varied

Execution of JavaScript Code

▷ **Question:** When and how is JavaScript code executed?

▷ **Answer:** While loading the HTML page or afterwards triggered by events.

▷ JavaScript in a script element: during page load: (not in a function)

```
<script type="text/javascript">alert('Huhu');</script>
```

▷ JavaScript in an event handler attribute onclick, ondblclick, onmouseover, ...” whenever the corresponding event occurs.

▷ JavaScript in a “special link”: when the anchor is clicked:

```
<a href="javascript:..."/>
```

The first key concept we need to understand here is that the browser essentially acts as an user interface: it presents the HTML pages to the user, waits for actions by the user – usually mouse clicks, drags, or gestures; we call them events – and reacts to them.

The second is that all events can be associated to an element node in the DOM: consider an HTML anchor node, as we have seen above, this corresponds to a rectangular area in the browser window. Conversely, for any point p in the browser window, there is a minimal DOM element $e(p)$ that contains p recall that the DOM is a tree. So, if the user clicks while the mouse is at point p, then the browser triggers a click event in $e(p)$, determines how $e(p)$ handles a click event, and if $e(p)$ does not, bubbles the click event up to the parent of $e(p)$ in the DOM tree.

There are multiple ways a DOM element can handle an event: some elements have default event handlers, e.g. an HTML anchor <a href="⟪URI⟫"> will handle a click event by issuing a HTTP GET request for ⟪URI⟫. Other HTML elements can carry event handler attributes whose JavaScript content is executed when the corresponding event is triggered on this element.

Actually there are more events than one might think at first, they include:

1. Mouse events; click when the mouse clicks on an element (touchscreen devices generate it on a tap); contextmenu: when the mouse right-clicks on an element; mouseover / mouseout: when the mouse cursor comes over / leaves an element; mousedown / mouseup: when the mouse button is pressed / released over an element; mousemove: when the mouse is moved.

2. Form element events; submit: when the visitor submits a <form>; focus: when the visitor focuses on an element, e.g. on an <input>.

3. keyboard events; keydown and keyup: when the visitor presses and then releases the button.

4. Document events; DOMContentLoaded:– when the HTML is loaded and processed, DOM is fully built, but external resources like pictures <img> and stylesheets may be not yet loaded. load: the browser loaded all resources (images, styles etc); beforeunload / unload: when the user is leaving the page.

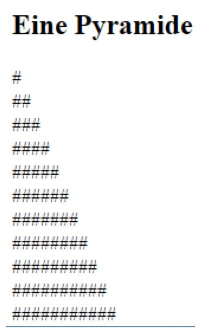
5. resource loading events; onload: successful load, onerror: an error occurred.

Let us now use all we have learned in an example to fortify our intuition about using JavaScript to change the DOM.

Example: Changing Web Pages Programmatically

▷ **Example 6.1.10 (Stupid but Fun).**

```
<body>
<h2>A Pyramid</h2>
<div id="pyramid"/>

<script type="text/javascript">
  var char = "#";
  var triangle = "";
  var str = "";
  for(var i=0;i<=10;i++){
      str = str + char;
      triangle = triangle + str + "<br/>"
      }
  var elem = document.getElementById("pyramid");
  elem.innerHTML=triangle;
</script>
</body>
</html>
```



Eine Pyramide

```
#
##
###
####
#####
######
#######
########
#########
##########
###########
```

The HTML document in ?? contains an empty <div> element whose id attribute has the value pyramid. The subsequent script element contains some code that builds a DOM node-set of 10 text and <br/> nodes in the triangle variable. Then it assigns the DOM node for the <div> to the variable elem and deposits the triangle node-set as children into it via the JavaScript innerHTML method.

We see the result on the right of ??. It is the same as if the #-strings and <br/> sequence had been written in the HTML which at least for pyramids of greater depth would have been quite tedious for the author.

6.2 Cascading Stylesheets

In this section we introduce a technology of digital documents which naturally belongs into ??: the specification of presentation (layout, colors, and fonts) for marked-up documents.

6.2.1 Separating Content from Layout

As the WWW evolved from a hypertext system purely aimed at human readers to a Web of multimedia documents, where machines perform added-value services like searching or aggregating, it became more important that machines could understand critical aspects web pages. One way to facilitate this is to separate markup that specifies the content and functionality from markup that specifies human-oriented layout and presentation (together called “styling”). This is what “cascading style sheets” set out to do.

Another motivation for CSS is that we often want the styling of a web page to be customizable (e.g. for vision impaired readers).

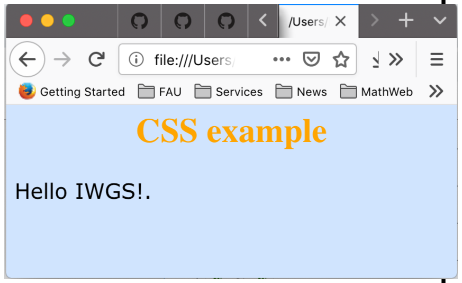
CSS: Cascading Style Sheets

▷ **Idea:** Separate structure/function from appearance.

▷ **Definition 6.2.1.** Cascading Style Sheets (CSS) is a style language that allows authors and users to attach style (e.g., fonts, colors, and spacing) to HTML and XML documents.

▷ **Example 6.2.2.** Our text file from ?? with embedded CSS:

```
<html>
  <head>
    <style type="text/css">
       body {background−color:#d0e4fe;}
       h1 {color:orange;
             text−align:center;}
       p {font−family:"Verdana";
             font−size:20px;}
     </style>
  </head>
  <body>
    <h1>CSS example</h1>
    <p>Hello IWGS!.</p>
  </body>
</html>
```



CSS example

Hello IWGS!.

Now that we have seen the example, let us fix the basic terminology of CSS.

CSS: Rules, Selectors, and Declarations

▷ **Definition 6.2.3.** A CSS style sheet consists of a sequence of rules that in turn consist of a set of selectors that determine which XML elements the rule applies to and a declaration block that specifies intended presentation.

▷ **Definition 6.2.4.** A CSS declaration block consists of a semicolon separated list of declarations in curly braces. Each declaration itself consists of a property, a colon, and a value.

▷ **Example 6.2.5.** In ?? we have three rules, they address color and font properties:

```
       body {background−color:#d0e4fe;}
       h1 {color:orange;
             text−align:center;}
       p {font−family:"Verdana";
```

▷ **Observation:** In modern web sites, CSS contributes as much – if not more – to the appearance as the choice of HTML elements.

In ?? the selectors are just element names, they specify that the respecive declaration blocks apply to all elements of this name.

We explore this new technology by way of an example. We rework the title box from the HTML example above – after all treating author/affiliation information as headers is not very semantic. Here we use div and span elements, which are generic block-level (i.e. paragraph-like) and inline containers, which can be styled via CSS classes. The class titlebox is represented by the CSS selector .titlebox.

A Styled HTML Title Box (Source)

▷ **Example 6.2.6 (A style Title Box).** The HTML source:

```
<head>
  <title>A Styled HTML Title</title>
  <link rel="stylesheet" type="text/css" href="style.css"/>
</head>
```

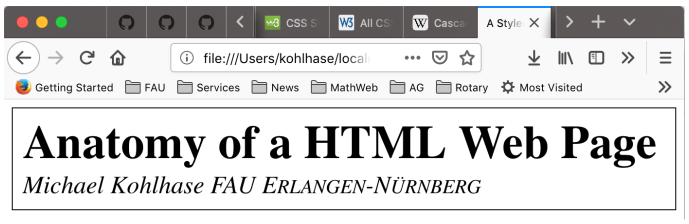
```
  <body>
    <div class="titlebox">
      <div class="title">Anatomy of a HTML Web Page</div>
      <div class="author">
        <span class="name">Michael Kohlhase</span>
        <span class="affil">FAU Erlangen−Nuernberg</span>
      </div>
    </div>
    ...
```

And the CSS file referenced in the <**link**> element in line 3:

```
.titlebox {border: 1px solid black;padding: 10px;
           text−align: center
           font−family: verdana;}
.title {font−size: 300%;font−weight: bold}
.author {font−size: 160%;font−style: italic;}
.affil {font−variant: small−caps;}
```

And here is the result in the browser:

A Styled HTML Title Box (Result)



Anatomy of a HTML Web Page

Michael Kohlhase FAU ERLANGEN-NÜRNBERG

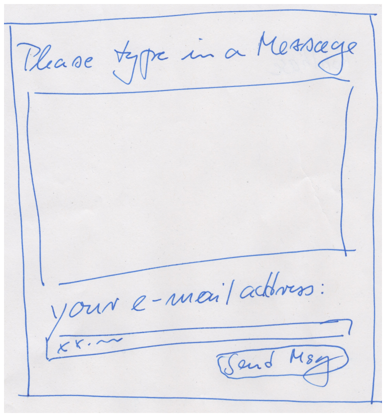
6.2.2 Worked Example: The Contact Form

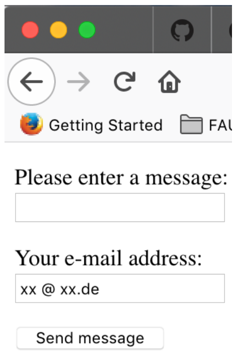
To fortify our intuition on CSS, we take up the “contact form” example from above and improve the layout in a step-by-step process concentrating on one aspect at a time.

CSS in Practice: The Contact Form Example (Continued)

▷ Recap: The unstyled contact form – Dream vs. Reality

```
<title>Contact</title>
<form action="contact-after.html">
  <h2>Please enter a message:</h2>
    <input name="msg" type="text"/>
  <h3>Your e-mail address:</h3>
  <input name="addr" type="text"
         value="xx @ xx.de"/>
  <br/>
  <input type="submit"
        value="Send message"/>
</form>
```





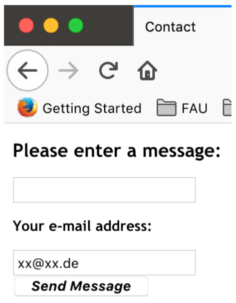
Please enter a message:

Your e-mail address:

xx @ xx.de

▷ Add a CSS file with font information

```
<link rel="stylesheet" type="text/css"
      href="csscontact1.css" />
  <input class="important" type="submit"
         value="Send Message"/>
```



Please enter a message:

Your e-mail address:

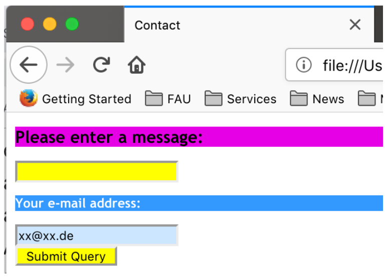
xx@xx.de

```
body {font-size: 62.5%;
      font-family: "Trebuchet MS",
          "Arial", "Helvetica",
          "Verdana", "sans-serif"}
.important{font-style: italic;}
input[type="submit"]{font-weight: bold;}
```

▷ Add lots of color (ooops, what about the size)

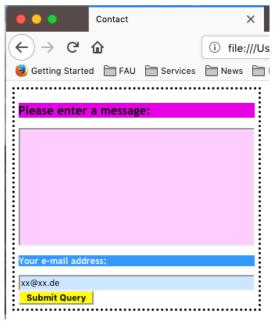
```
<h2>Please enter a message:</h2>
<h3>Your e−mail address:</h3>
<input class="important" name="addr"
       style="background−color:#cce6ff"
       type="text" value="xx@xx.de"/>
```

```
h2 {background−color: #e600e6;}
h3 {background−color: #3399ff;
    color: white;}
input{background−color:yellow}
```



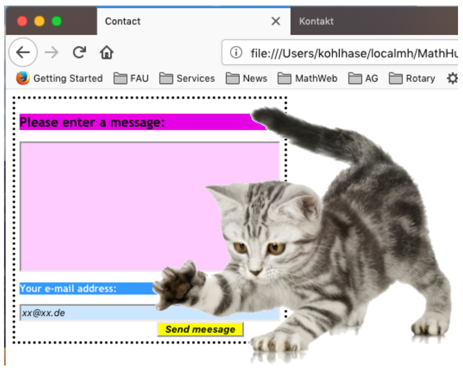
▷ Add size information and a dotted frame

```
<form action="contact−after.html"
      style="width:8cm;border:dotted;padding:5px">
  <h2>Please enter a message:</h2>
  <input name="msg" type="text"
         style="height:4cm;width:8cm;
                background−color:#ffccff"/>
  <br/>
  <h3>Your e−mail address:</h3>
  <input class="important" name="addr"
         type="text"
         value="xx@xx.de" style="width:8cm;
                 background−color:#cce6ff"/>
```



▷ Add a cat that plays with the submit button (because we can)

```
<img id="cat" src="cat.png"
     style="position:absolute;
                left:170px;top: 15px;
                width=300px"/>
```



This worked example should be enough to cover most layout needs in practice. Note that in most use cases, these generally layout primitives will have to be combined in different and may be even new ways.
Actually, the last "improvement" may have gone a bit overboard; but we used it to show how absolute positioning of images (or actually any CSS boxes for that matter) works in practice.

6.2.3 A small but useful Fragment of CSS

CSS is a huge ecosystem of technologies, which is spread out over about 100 particular specifications – see [CSSa] for an overview.

We will now go over a small fragment of CSS that is already very useful for web applications in more detail and introduce it by example. For a more complete introduction, see e.g. [CSSc].

Recall that selectors are the part of CSS rules that determine what elements a rule affects. We now give the most important cases for our applications.

CSS Selectors

▷ **Question:** Which elements are affected by a CSS rule?

▷ Elements of a given name (optionally with given attributes)

- ▹ Selectors: name $\widehat{=}$ ⟪elname⟫, attributes $\widehat{=}$ [⟪attname⟫=⟪attval⟫]

▷ **Example 6.2.7.** p[xml:lang='de'] applies to <**p** xml:**lang**="de">...</**p**>

▷ Any element with a given class attributes

- ▹ Selector: .⟪classname⟫

▷ **Example 6.2.8.** .important applies to <⟪el⟫ class='important'>...</⟪el⟫>

▷ The element with a given id attribute

- ▹ Selector: #⟪id⟫

▷ **Example 6.2.9.** #myRoot applies to <⟪el⟫ id='myRoot'>...</⟪el⟫>

▷ **Note:** Multiple selectors can be combined in a comma separated list.

▷ For a full list see `https://www.w3schools.com/cssref/css_selectors.asp`.

SOME RIGHTS RESERVED

We now come to one of the most important conceptual parts of CSS: the box model. Understanding it is essential for dealing with CSS based layouts.

The CSS Box Model

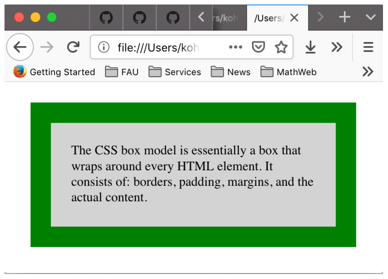
▷ **Definition 6.2.10.** For layout, CSS considers all HTML elements as boxes, i.e. document areas with a given width and height. A CSS box has four parts:

- ▹ content: the content of the box, where text and images appear.
- ▹ padding: clears an area around the content. The padding is transparent.
- ▹ border a border that goes around the padding and content.
- ▹ margin clears an area outside the border. The margin is transparent.

The latter three wrap around the content and add to its size.

▷ All parts of a box can be customized with suitable CSS properties:

```
div {
    background—color: lightgrey;
    width: 300px;
    border: 25px solid green;
    padding: 25px;
    margin: 25px;
}
```

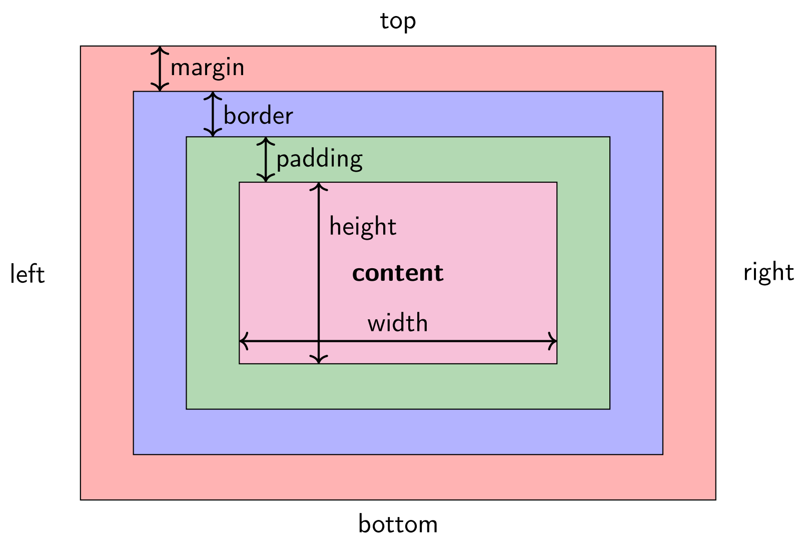

The CSS box model is essentially a box that wraps around every HTML element. It consists of: borders, padding, margins, and the actual content.

Note that the overall width of the CSS box is $300 + 2 \cdot 3 \cdot 25 = 450$ pixels.

As a summary of the above, we can visualize the CSS box model in a diagram:

The CSS Box Model: Diagram

▷ **Definition 6.2.11.** The following diagram summarizes the CSS box model



We now come to a topic that is quite mind-boggling at first: The "cascading" aspect of CSS stylesheets. Technically, the story is quite simple, there are two independent mechanisms at work:

- inheritance: if an element is fully contained in another, the inner (usually) inherits all properties of the outer.
- rule priorization: if more than one selector applies to an element (e.g. one by element name and one by id attribute), then we have to determine what rule applies.

Technically, priorization takes care of them in an integrated fashion.

Cascading of selectors in CSS: Priorization

▷ Multiple CSS selectors apply with the following priorizations:

1. important (i.e. marked with !important) before unimportant
2. inline (specified via the style attribute)
3. media-specific rules before general ones
4. user-defined CSS stylesheet (e.g. in the `FireFox` profile)
5. specialized before general selectors (complicated; see e.g. [CSSb])
6. rule order: later before earlier selectors
7. parent inheritance: unspecified properties are inherited from the parent.
8. Style sheet included or referenced in the HTML document.
9. browser default

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 194 2025-02-06

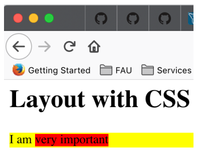
But do not despair with this technical specification, you do not have to remember it to be effective with CSS practically, because the rules just encode very natural "behavior". And if you need to understand what the browser – which implements these rules – really sees, use the integrated page inspector tool (see slide 199 for details).
We now look at an example to fortify our intuition.

Cascading of selectors in CSS: Priorization Example

▷ **Example 6.2.12.** Can you explain the colors in the web browsers below?

```
<h1>Layout with CSS</h1>
<div id="important" class="blue">
  I am <span class="markedimportant">very important</span>
</div>
```

```
.markedimportant {background—color:red !important}
#important {background—color:green}
.blue {background—color:blue}
#important {background—color:yellow}
```



Layout with CSS

I am very important

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 1/2 195 2025-02-06

For instance, the words *very important* get a red background, as the class markedimportant is marked as important by the CSS keyword !important, which makes (cf. rule 1 above) the color red win agains the color yellow inherited from the parent <div> element (rule 7 above).
Let us now look at CSS inheritance in a little more detail.

Cascading in CSS: Inheritance

▷ **Definition 6.2.13.** Child elements can inherits some properties (called inheritable) from their parents. In a nutshell:

- ▷ text-related properties are inheritable; e.g. color, font, letter—spacing, line—height, list—style, and text—align
- ▷ box-related properties are not; e.g. background, border, display, float, clear, height, width, margin, padding, position, and text—align.

▷ **Note:** Inheritance is integrated into priorization. (recall case 7. above)

▷ Inheritance makes for consistent text properties and smaller CSS stylesheets.

So far, we have looked at the mechanics of CSS from a very general perspective. We will now come to a set of CSS behaviors that are useful for specifying layouts of pages and texts. Recall that CSS is based on the box model, which understands HTML elements as boxes, and layouts as properties of boxes nested in boxes (as the corresponding HTML elements are).

If we can specify how inner boxes float inside outer boxes – via the CSS float rules, we can already do quite a lot, as the following examples show.

CSS Flow: How Boxes Float to their Place

▷ **Definition 6.2.14.** CSS Flow describes how different elements are distributed in the visible area. (how they flow; hence the name)
The float property allows to influence that.

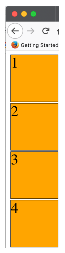
▷ **Example 6.2.15.** Block-level boxes (here div node) float to the left:

```
<div class="square">1</div>
<div class="square">2</div>
<div class="square">3</div>
<div class="square">4</div>
```

+

```
.square {font-size:200%;
         height:100px;
         width:100px;
         border:1px solid black;
         margin:2px;
         background-color:orange;}
```

=



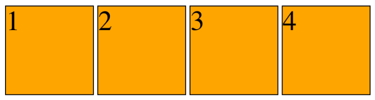
▷ **Example 6.2.16.** float:left floats boxes as far as they will go: (without overlap)

```
<div class="square">1</div>
<div class="square">2</div>
<div class="square">3</div>
<div class="square">4</div>
```

+

```
.square {font-size:200%;
         height:100px;
         width:100px;
         border:1px solid black;
         margin:2px;
         background-color:orange;
         float:left}
```

=



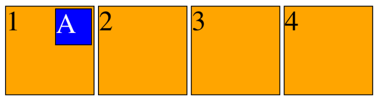
▷ **Example 6.2.17.** float:right in a div will float inside the corresponding box.

```
<div class="square">1
  <div class="smallsq">A</div>
</div>
<div class="square">2</div>
<div class="square">3</div>
<div class="square">4</div>
```

+

```
.smallsq {color:white;
  height: 40px;width: 40px;
  border: 1px solid black;
  margin: 2px;
  background-color: blue;
  float: right}
```

=



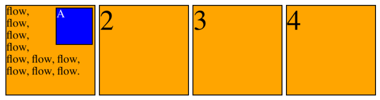
▷ **Example 6.2.18.** float:left will let contents flow around an obstacle

```
<div class="square"
  style="font−size:small">
  <div class="smallsq">A</div>
  flow, flow, flow, flow, flow,
  flow, flow, flow, flow, flow.
</div>
```

+

```
.smallsq {color:white;
  height: 40px;width: 40px;
  border: 1px solid black;
  margin: 2px;
  background−color: blue;
  float: right}
```

=

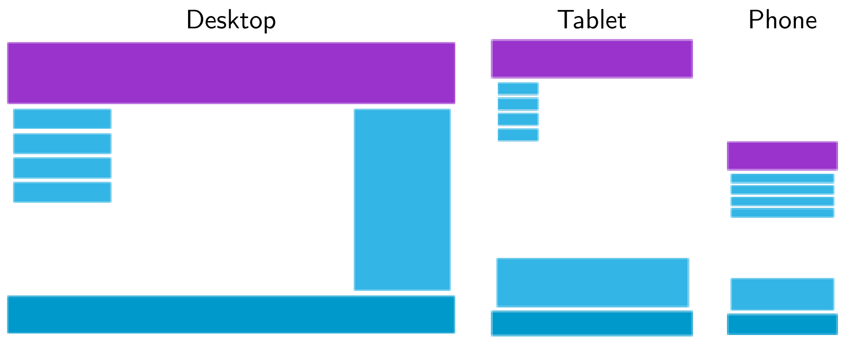


The large space (>2px) is caused because there is no linebreaking.

One of the important applications of the content/form separation made possible by CSS is to tailor web page layout to the screen size and resolution of the device it is viewed on. Of course, it would be possible to maintain multiple layouts for a web page one per screensize/resolution class, but a better way is to have one layout that changes according to the device context. This is what we will briefly look at now.

CSS Application: Responsive Design

- ▷ **Problem:** What is the screen size/resolution of my device?
- ▷ **Definition 6.2.19.** Responsive web design (RWD) designs web documents so that they can be viewed with a minimum of resizing, panning, and scrolling – across a wide range of devices (from desktop monitors to mobile phones).
- ▷ **Example 6.2.20.** A web page with content blocks



- ▷ **Implementation:** CSS based layout with relative sizes and media queries– CSS conditionals based on client screen size/resolution/. . .

6.2.4 CSS Tools

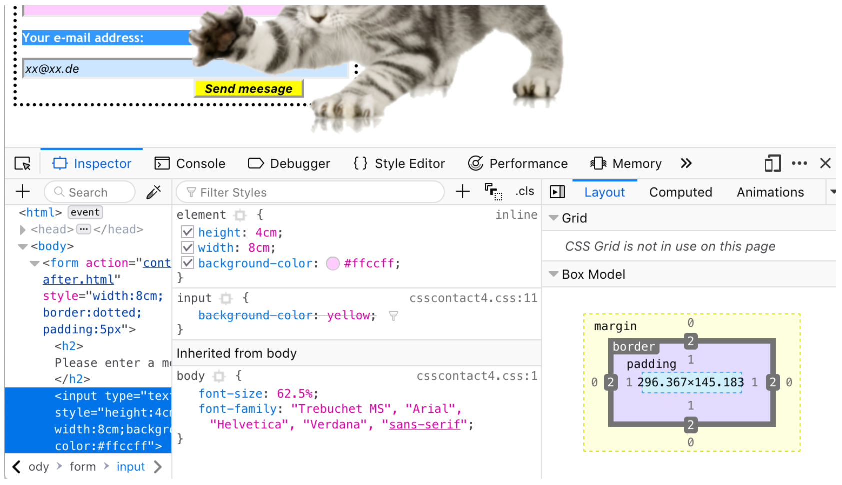
In this subsection we introduce a technology of digital documents which naturally As CSS has grown to be very complex and moreover, the browser DOM of which CSS is part can even be modified after loading the HTML (see **??**), we need tools to help us develop effective and maintainable CSS.

But how to find out what the web browser really sees?

▷ CSS has many interesting inheritance rules.

▷ **Definition 6.2.21.** The page inspector tool gives you an overview over the internal state of the web browser and its DOM.

▷ **Example 6.2.22.**



```
<html> event
<head>…</head>
<body>
  <form action="cont
  after.html"
  style="width:8cm;
  border:dotted;
  padding:5px">
    <h2>
    Please enter a me
    </h2>
    <input type="tex
    style="height:4cr
    width:8cm;backgr
    color:#ffccff">
```

```
element {                                    inline
  height: 4cm;
  width: 8cm;
  background-color: #ffccff;
}
input {                            csscontact4.css:11
  ~~background-color: yellow;~~
}
Inherited from body
body {                              csscontact4.css:1
  font-size: 62.5%;
  font-family: "Trebuchet MS", "Arial",
    "Helvetica", "Verdana", "sans-serif";
}
```

CSS Grid is not in use on this page

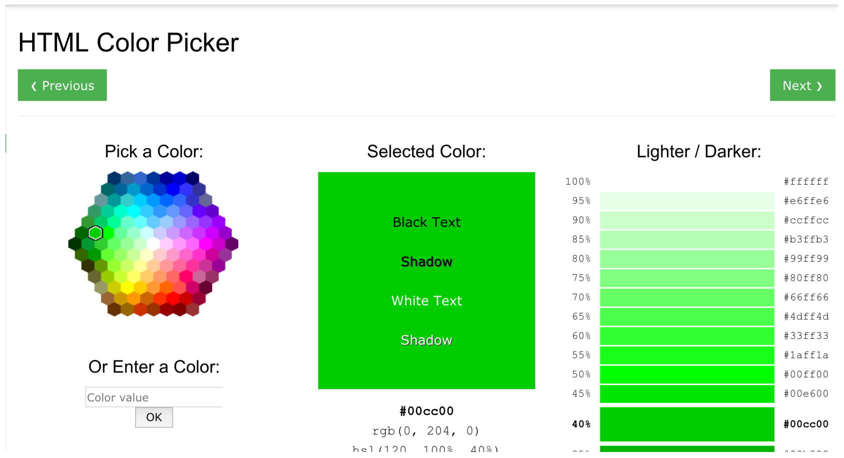
```
margin        0
  border      2
    padding   1
0 2 1 296.367×145.183 1 2 0
              1
              2
              0
```

In CSS we can specify colors by various names, but the full range of possible colors can only specified by numeric (usually hexadecimal) numbers. For instance in ??, we specified the background color of the page as #d0e4fe;, which is a pain for the author. Fortunately, there are tools that can help.

Picking CSS Colors

▷ **Problem:** Colors in CSS are specified by funny names (e.g. CornflowerBlue) or hexadecimal numbers, (e.g. #6495ED).

▷ **Solution:** Use an online color picker, e.g. `https://www.w3schools.com/colors/colors_picker.asp`



HTML Color Picker

Selected Color:

Black Text

Shadow

White Text

Shadow

#00cc00
rgb(0, 204, 0)
hsl(120, 100%, 40%)

Lighter / Darker:

100%	#ffffff
95%	#e6ffe6
90%	#ccffcc
85%	#b3ffb3
80%	#99ff99
75%	#80ff80
70%	#66ff66
65%	#4dff4d
60%	#33ff33
55%	#1aff1a
50%	#00ff00
45%	#00e600
40%	**#00cc00**

6.3 jQuery: Write Less, Do More

While JavaScript is fully sufficient to manipulate the HTML DOM, it is quite verbose and tedious to write. To remedy this, the web developer community has developed libraries that extend the JavaScript language by new functionalities that more concise programs and are often used Instead of pure JavaScript.

jQuery: Write Less, Do More

- **Definition 6.3.1.** JQuery is a feature-rich JavaScript library that simplifies tasks like HTML document traversal and manipulation, event handling, animation, and Ajax.
- **Using:**
  - Download from `https://jquery.com/download/`, save on your system (remember where)
  - integrate into your HTML (usually in the <head>)

    ```
    <script type="text/javascript" src="client−js/jquery−3.2.1.min.js"/>
    ```

    or from the internet directly (only works if you are online)

    ```
    <script src="https://ajax.googleapis.com/ajax/libs/jquery/3.2.1/jquery.min.js" />
    ```

The key feature of jQuery is that it borrows the notion of “selectors” to describe HTML node sets from CSS actually, jQuery uses the CSS selectors directly and then uses JavaScript-like methods to act on them. In fact, the name jQuery comes from the fact that selectors “query” for nodes in the DOM.

jQuery Philosophy and Layers

- **jQuery Philosophy:** Select a subtree from the DOM, and operate on it.

▷ **Syntax Convention:** jQuery instructions start with a $ to distinguish it from JavaScript.

▷ **Example 6.3.2.** The following jQuery command achieves a lot in four steps:

```
$("#myId").show().css("color", "green").slideDown();
```

1. Find elements in the DOM by CSS selectors, e.g. $("#myId")
2. do something to them, here show() (chaining of methods)
3. change their layout by changing CSS attributes, e.g. css("color","green")
4. change their behavior, e.g. slideDown()

▷ **Good News:** jQuery selectors $\hat{=}$ CSS selector.

We will now show a couple of jQuery methods for inserting material into HTML elements and discuss their behavior in examples

Inserting Material into the DOM

▷ **Inserting before the first child:**

```
$('#content').prepend(function(){return 'in front';});
```

▷ **Inserting after the last child:**

```
$('#content').append('<p>Hello</p>');
$('#content').append(function(){ return 'in the back'; });
```

▷ **Inserting before/after an element:**

```
$('#price').before('Price:');
$('#price').after(' EUR')
```

Let us fortify our intuition about dynamic HTML by going into a more involved example. We use the toggle method from the jQuery layout layer to change visibility of a DOM element. This method adds and removes a style="display:none" attribute to an HTML element and thus toggles the visibility in the browser window.

Applications and useful tricks in Dynamic HTML

▷ **Observation:** jQuery is not limited to adding material to the DOM.

▷ **Idea:** Use jQuery to change CSS properties in the DOM as well.

▷ **Example 6.3.3 (Visibility).** Hide document parts by setting CSS style attributes to display:none.

```
<html>
  <head>
    <title>Toggling</title>
     <style type="text/css">#dropper { display: none; }</style>
    <script src="https://ajax.googleapis.com/ajax/libs/jquery/3.2.1/jquery.min.js" />
    <script language="JavaScript" type="text/javascript">
      $("button").click(function(){$("#dropper").toggle();});
     </script>
  </head>
  <body>
    <h2>Toggling the visibility of material</h2>
     <button>...more </button>
     <div id="dropper"><p>Now you see it!</p></div>
  </body>
</html>
```

Fun with Buttons (Three easy Interactions)

▷ **Example 6.3.4 (A Button that Changes Color on Hover).**

```
  <div id="hoverPoint">
    <button id="hover">hover</button>
    <script type="text/javascript">
      $("#hover").hover(function () {$(this).css("background−color", "red");},
                           function () {$(this).css("background−color", "blue");});
    </script>
  </div>
```

- ▷ The HTML has a button with text "hover".
- ▷ The jQuery code selects it via its id and catches its hover event via the hover() method.
- ▷ This takes two functions as arguments:
  - ▷ The first is called when the mouse moves into the button, the second when it leaves.
  - ▷ The first changes changes the button color to red, the second reverts this.

Fun with Buttons (Three easy Interactions)

▷ **Example 6.3.5 (A Button that Uncovers Text).**

```
  <div id="readPoint">
    <button class="read" style="display:block">Read More</button>
    <button class="read" style="display:none">Read Less</button>
    <div id="rText" style="display:none; width:200px; clear:left">
      A read−more button is not only a call−to−action, but it also organizes
      the screen area management in a non−wasteful way. If and only if users are interested,
      they will use the button.<br/>
    </div>
    <script type="text/javascript">
      $(".read").click(function() {$("#rText").toggle("slow",function(){$(".read").toggle()});});
    </script>
  </div>
```

- ▷ The HTML has two buttons (one of them visible) and a text.
- ▷ The jQuery code selects both buttons via their read class.

- ▷ A click event activates the .click() method taking an event handler function:
  - ▷ This selects the text via its id attribute rTeX and
  - ▷ uses the toggle() method which changes the display between none and block.
  - ▷ The first parameter of toggle() is a duration for the animation.
  - ▷ The second is a completion function to be run after animation finishes.
  - ▷ Here complection function makes the respective other button visible (read more/less).

Fun with Buttons (Three easy Interactions)

▷ **Example 6.3.6 (A Button that Plays a Sound).**

```
<div id="soundPoint">
  <button id="sound" onclick="playSound('laugh.mp3')">Sound</button>
  <script type="text/javascript">
    function playSound(url) {
      console.log("Call playSound with " + url);
      const a = new Audio(url);
      a.play();
      }
  </script>
</div>
```

- ▷ The HTML has a button with text “sound” and an onclick attribute.
- ▷ That activates the playSound function on a URL:
- ▷ The playSound function is defined in the script element: it
  - ▷ logs the action and URL in the browser console,
  - ▷ makes a new audio object a, which
  - ▷ plays it via the play() method.

For reference, here is the full code of the examples in one file:

```
<html>
<head>
  <title>Buttons</title>
  <script src="https://code.jquery.com/jquery-3.4.1.min.js" type="text/javascript"></script>
  <style type="text/css">
    button {color: white; font-size: large; background-color: blue;
                 width: 110px; height: 40px; border-radius: 20px;}
    div[id$="Point"] {display: inline-block;}
  </style>
</head>

<body>
  <h1 id="top">Look how easy interaction is ... </h1>

  <div id="hoverPoint">
    <button id="hover">hover</button>
    <script type="text/javascript">
      $("#hover").hover(function () {$(this).css("background-color", "red");},
                         function () {$(this).css("background-color", "blue");});
    </script>
  </div>
  <div id="readPoint">
    <button class="read" style="display:block">Read More</button>
    <button class="read" style="display:none">Read Less</button>
    <div id="rText" style="display:none; width:200px; clear:left">
      A read-more button is not only a call-to-action, but it also organizes
      the screen area management in a non-wasteful way. If and only if users are interested,
      they will use the button.<br/>
    </div>
    <script type="text/javascript">
      $(".read").click(function() {$("#rText").toggle("slow",function(){$(".read").toggle()});});
    </script>
  </div>

  <div id="soundPoint">
    <button id="sound" onclick="playSound('laugh.mp3')">Sound</button>
    <script type="text/javascript">
      function playSound(url) {
        console.log("Call playSound with " + url);
        const a = new Audio(url);
        a.play();
        }
    </script>
  </div>
</body>
</html>
```

It has a bit more general CSS and includes jQuery in the beginning.

Chapter 7

What did we learn in IWGS-1?

Outline of IWGS 1:

▷ Programming in Python: (main tool in IWGS)

  ▹ Systematics and culture of programming
  ▹ Program and control structures
  ▹ Basic data structures like numbers and wordsstring, character encodings, unicode, and regular expressions

▷ Electronic documents and document processing:

  ▹ text files
  ▹ markup systems, HTML, and CSS
  ▹ XML: Documents are trees.

▷ Web technologies for interactive documents and web applications

  ▹ internet infrastructure: web browsers and server
  ▹ serverside computation: bottle routing and
  ▹ clientside interaction: dynamic HTML, JavaScript, HTML forms

▷ Web application project (fill in the blanks to obtain a working web app)

Outline of IWGS-II:

▷ Databases

  ▹ CRUD operations, querying, and python embedding
  ▹ XML and JSON for file based data storage

▷ BooksApp: a Books Application with persistent storage

▷ Image processing

  ▹ Basics

  - ▹ Image transformations, Image Understanding
- ▷ Ontologies, semantic web, and WissKI
  - ▹ Ontologies (inference $\rightsquigarrow$ get out more than you put in)
  - ▹ semantic web Technologies (standardize ontology formats and inference)
  - ▹ Using semantic web Tech for cultural heritage research data $\rightsquigarrow$ the WissKI System
- ▷ Legal Foundations of Information Systems
  - ▹ Copyright & Licensing
  - ▹ Data Protection (GDPR)

SOME RIGHTS RESERVED

Part II

IWGS-II: DH Project Tools

Chapter 8

Semester Change-Over

8.1 Administrativa

We will now go through the ground rules for the course. This is a kind of a social contract between the instructor and the students. Both have to keep their side of the deal to make learning as efficient and painless as possible.

Prerequisites

- ▷ **Formal Prerequisite:** IWGS-1 (If you did not take it, read the notes)
- ▷ **General Prerequisites:** Motivation, interest, curiosity, hard work.
  nothing else! (apart from IWGS-1)
  We will teach you all you need to know
- ▷ You can do this course if you want! (we will help)

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 210 2025-02-06

Now we come to a topic that is always interesting to the students: the grading scheme: The short story is that things are complicated. We have to strike a good balance between what is didactically useful and what is allowed by Bavarian law and the FAU rules.

Assessment, Grades

- ▷ **Grading Background/Theory:** Only modules are graded! (by the law)
  - ▷ Module "DH-Einführung" (DHE) $\hat{=}$ courses IWGS1/2, DH-Einführung. (7.5 ECTS)
  - ▷ DHE module grade $\rightsquigarrow$ pass/fail determined by "portfolio" $\hat{=}$ collection of contributions/assessments.
  - ▷ Module "DH-Einführung mit Übungen" (DHÜ) $\hat{=}$ courses IWGS1/2, (10 ECTS)
  - ▷ DHÜ module grade $\rightsquigarrow$ 1-5 50% exam, 50% homework assignments, 10% bonus points from prepquizzes.
- ▷ **Assessment Practice:** The IWGS assessments in the "portfolio" consist of

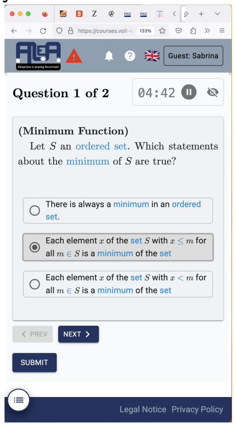
  - ▹ weekly homework assignments, (practice IWGS concepts and tools)
  - ▹ 60 minutes exam directly after lectures end: July 27. 2024.
- ▷ **Retake Exam:** 60 min exam at the end of the exam break. (October. 12. 2024)
- ▷ **To help you succeed:** We offer you
  - ▹ **External motivation**: informal points for homeworks and a grade for exam, (even though only pass/fail relevant in the end)
  - ▹ weekly online prepquizzes that help you prepare for the course. (check understanding/preparation)

Homework assignments, quizzes, and end-semester exam may seem like a lot of work – and indeed they are – but you will need practice (getting your hands dirty) to master the IWGS concepts. We will go into the details next.

Preparedness Quizzes

- ▷ **PrepQuizzes:** Before every lecture we offer a 10 min online quiz – the PrepQuiz – about the material from the previous week. (16:00-16:10; starts in week 2)
- ▷ **Motivations:** We do this to
  - ▹ keep you prepared and working continuously. (primary)
  - ▹ bonus points for the DHÜ students if you passed the exam with $\geq 50\%$ points. (potential part of your grade)
  - ▹ update the ALeA learner model. (fringe benefit)
- ▷ The prepquiz will be given in the ALeA system

- `https://courses.voll-ki.fau.de/quiz-dash/iwgs-2`
- You have to be logged into ALeA! (via FAU IDM)
- You can take the prepquiz on your laptop or phone, ...
- ... in the lecture or at home ...
- ... via WLAN or 4G Network. (do not overload)
- Prepquizzes will only be available 16:00-16:10!



Question 1 of 2 04:42

(Minimum Function)
Let S an ordered set. Which statements about the minimum of S are true?

- ○ There is always a minimum in an ordered set.
- ◉ Each element x of the set S with $x \leq m$ for all $m \in S$ is a minimum of the set
- ○ Each element x of the set S with $x < m$ for all $m \in S$ is a minimum of the set

SOME RIGHTS RESERVED

Outline of IWGS-II:

- Databases
  - CRUD operations, querying, and python embedding
  - XML and JSON for file based data storage
- BooksApp: a Books Application with persistent storage
- Image processing
  - Basics
  - Image transformations, Image Understanding
- Ontologies, semantic web, and WissKI
  - Ontologies (inference $\leadsto$ get out more than you put in)
  - semantic web Technologies (standardize ontology formats and inference)
  - Using semantic web Tech for cultural heritage research data $\leadsto$ the WissKI System
- Legal Foundations of Information Systems
  - Copyright & Licensing

▷ Data Protection (GDPR)

In IWGS-II, we want to consolidate the methods and technologies we learn in a small information system, which students build in groups, and which will serve as a running example for the course. These projects will consist of documents, data, and programs.

IWGS-II Project

▷ **Idea:** Consolidate the techniques from IWGS-I and IWGS-II into a prototypical information system for Art History @ FAU. (Practical Digital Humanities)

▷ **A Running Example:** Research image + metadata collection "Bauernkirmes" provided by Prof. Peter Bell

▷ **What will you do?:** Build a web-based image/data manager, test image algorithms, annotate ontologically, . . .

▷ **How will we organize this:** Mostly via the group homework assignments (together they will make the project)

Chapter 9

Databases

We now come to one of the core tools of computer science: databases give us a means to store large collections of data and organize them for efficient access. We will introduce the underlying concepts by example, go over the basics of relational database systems and the SQL language, and experiment with a concrete system: SQLite and its embedding into Python. **Acknowledgements:** We have borrowed and adapted examples and from [SSU04] and [PMDA] in this chapter.

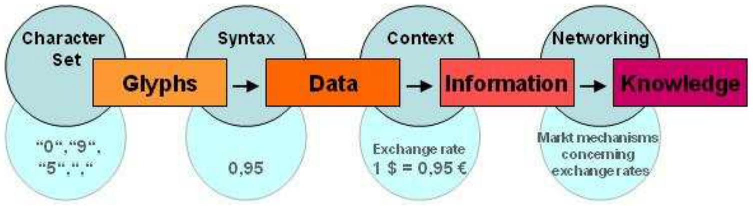
9.1 Introduction

Before we do anything else, we wil will look at various concepts around data to clarify concerns.

Databases, Data, Information, and Knowledge

▷ **Definition 9.1.1.** Discrete, objective facts or observations, which are unorganized and uninterpreted are called data (singular datum).

▷ According to Probst/Raub/Romhardt [PRR97]



▷ **Example 9.1.2.** The height of Mt. Everest (8.848 meters) is a datum.

Definition 9.1.3. A database is an organized collection of data, stored and accessed electronically from a computer system.

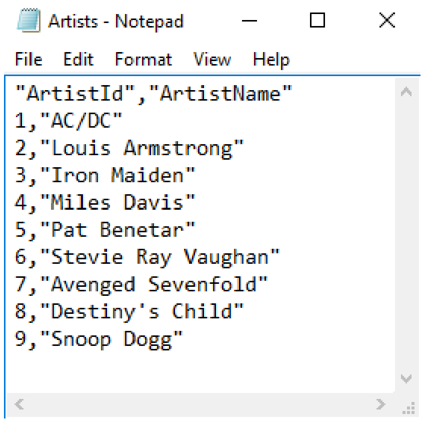
▷

To get an intuition about the possibilities of storing data, we look at some common ways – some of which we have already seen – and characterize them by some practical dimensions.

Storing Data Electronically

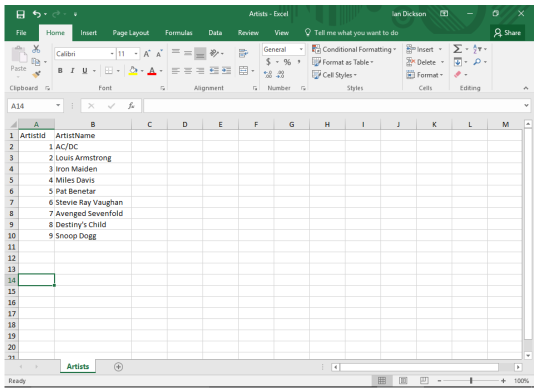
▷ Four conventional ways of storing data: (mileage varies)

- ▹ In the computer's memory (RAM) (very fast (+), random access (+), but not persistent (-))
- ▹ In a text file (persistent (+), fast (+), sequential access ( ), unstructured ( ))



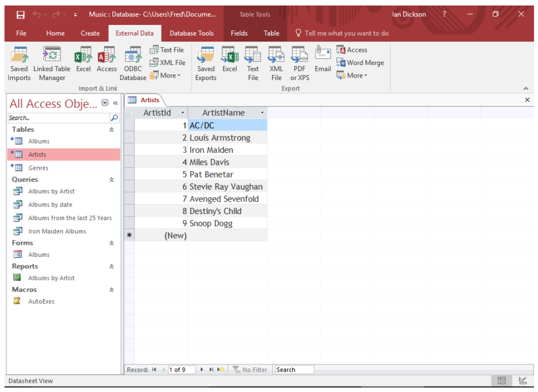
```
"ArtistId","ArtistName"
1,"AC/DC"
2,"Louis Armstrong"
3,"Iron Maiden"
4,"Miles Davis"
5,"Pat Benetar"
6,"Stevie Ray Vaughan"
7,"Avenged Sevenfold"
8,"Destiny's Child"
9,"Snoop Dogg"
```

- ▹ In a spreadsheet (persistent (+), 2D-structured (+-), relations (+), slow (-))


Artists - Excel
Ian Dickson
File Home Insert Page Layout Formulas Data Review View Tell me what you want to do Share
Artists
Ready

ArtistId	ArtistName
1	AC/DC
2	Louis Armstrong
3	Iron Maiden
4	Miles Davis
5	Pat Benetar
6	Stevie Ray Vaughan
7	Avenged Sevenfold
8	Destiny's Child
9	Snoop Dogg

- ▹ In a database (persistent (+), scalable (+), relations(+), managed (+), slow (-))


Music : Database- C:\Users\Fred\Docume...
Table Tools
Ian Dickson
File Home Create External Data Database Tools Fields Table Tell me what you want to do
All Access Obje...
Tables
Albums
Artists
Genres
Queries
Albums by Artist
Albums by date
Albums from the last 25 Years
Iron Maiden Albums
Forms
Albums
Reports
Albums by Artist
Macros
AutoExec
Record: 1 of 9 No Filter Search
Datasheet View

ArtistId	ArtistName
1	AC/DC
2	Louis Armstrong
3	Iron Maiden
4	Miles Davis
5	Pat Benetar
6	Stevie Ray Vaughan
7	Avenged Sevenfold
8	Destiny's Child
9	Snoop Dogg
(New)	

▷ Databases constitute the most scalable, persistent solution.

We will study the practical aspects of one particularly important class of database systems: relational database management systems.

9.2 Relational Databases

We will now study a particular kind of database: relational database, as these are the most widely used and structured ones.[3] EdN:3

(Relational) Database Management Systems

▷ **Definition 9.2.1.** A database management system (DBMS) is program that interacts with end users, applications, and a database to capture and analyze the data and provides facilities to administer the database.

▷ There are different types of DBMS, we will concentrate on relational ones.

▷ **Definition 9.2.2.** In a relational database management system (RDBMS), data are represented as tables: every datum is represented by a row (also called database record), which has a value for all columns (also called an column attribute) or field). A null value is a special "value" used to denote a missing value.

▷ **Remark:** Mathematically, each row is an n tuple of values, and thus a table an n-ary relation. (useful for standardizing RDBMS operations)

▷ **Example 9.2.3 (Bibliographic Data).**

LastN	FirstN	YOB	YOD	Title	YOP	Publisher	City
Twain	Mark	1835	1910	Huckleberry Finn	1986	Penguin USA	NY
Twain	Mark	1835	1910	Tom Sawyer	1987	Viking	NY
Cather	Willa	1873	1947	My Antonia	1995	Library of America	NY
Hemingway	Ernest	1899	1961	The Sun Also Rises	1995	Scribner	NY
Wolfe	Thomas	1900	1938	Look Homeward, Angel	1995	Scribner	NY
Faulkner	William	1897	1962	The Sound and the Fury	1990	Random House	NY

▷ **Definition 9.2.4.** Tables are identified by table name and individual components of records by column name.

As RDBMS constitute the backbone of of modern information technology, there are many many implementations, commercial ones and open source ones as well. For our purposes, open-source systems are completely sufficient, so we list the most important ones here.

Open-Source Relational Database Management Systems

[3] EdNote: MK: In the last years, NoSQL databases and JSON have gained prominaence. Intro them at the end and reference them here.

- **Definition 9.2.5.** MySQL is an open source RDBMS. For simple data sets and web applications MySQL is a fast and stable multi user system featuring an SQL database server that can be accessed by multiple clients.
- **Definition 9.2.6.** PostgreSQL is an open source RDBMS with an emphasis on extensibility, standards compliance, and scalability.
- **Definition 9.2.7.** SQLite is an embeddable RDBMS. Instead of a database server, SQLite uses a single database file, therefore no server configuration is necessary.
- **Remark:** At the level we use SQL in IWGS, all are equivalent.
- We will use SQLite in IWGS, since it is easiest to install and configure.

Now that we have made our first steps in the SQL language and with RDBMS in general, let us pick a concrete RDBMS to experiment with.

Working with SQLite (via the SQLite shell)

- In IWGS we will use SQLite, since it is very lightweight, easy to install, but feature complete, and widely used.
- Download SQLite at `https://www.sqlite.org/download.html`,
  - e.g. sqlite—dll—win64—x64—3280000.zip for windows.
  - unzip it into a suitable location, start sqlite3.exe there
  - this opens a command line interpreter: the SQLite shell. (all DBs have one) test it with .help that tells you about more “dot commands”.

```
> sqlite3
SQLite version 3.24.0 2018—06—04 19:24:41
Enter ".help" for usage hints.
Connected to a transient in—memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite> .help
.archive ... Manage SQL archives: ".archive ——help" for details
.auth ON|OFF Show authorizer callbacks
[...]
```

  - If you have a database file books.db from ??, use that.

```
> sqlite3 books.db
SQLite version 3.24.0 2018—06—04 19:24:41
Enter ".help" for usage hints.
> .tables
Books
>select * from Books;
Twain|Mark|1835|1910|Huckleberry Finn|1986|Penguin USA|NY
Twain|Mark|1835|1910|Tom Sawyer|1987|Viking|NY
Cather|Willa|1873|1947|My Antonia|1995|Library of America|NY
Hemingway|Ernest|1899|1961|The Sun Also Rises|1995|Scribner|NY
```

```
Wolfe|Thomas|1900|1938|Look Homeward, Angel|1995|Scribner|NY
Faulkner|William|1897|1962|The Sound and the Furry|1990|Random House |NY
Tolkien|John Ronald Reuel|1892|1973|The Hobbit|1937|George Allen  Unwin|UK
```

▷ .tables shows the available tables

select ∗ from Books is SQL (see below); it shows all entries of the Books table.

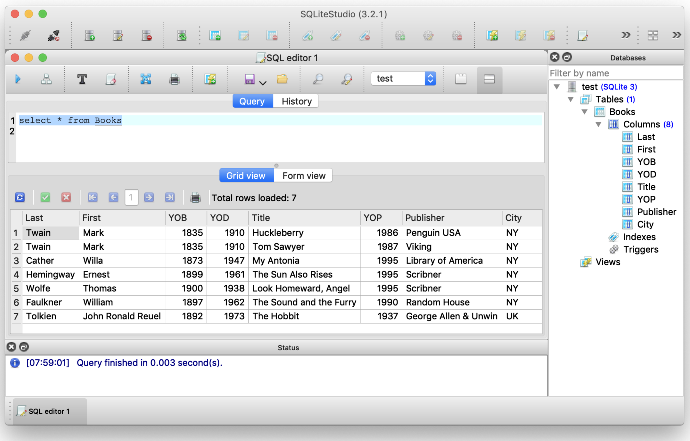
Interacting with SQLite via the database shell is nice, but can be quite tedious. Fortunately, there are better alternatives.

A Graphical User Interface for SQLite

▷ **Definition 9.2.8.** A database browser is a graphical user interface for a RDBMS that (typically) bundles an SQL instruction editor with displays for query results and the database schema in separate windows.

▷ I will sometimes use one for SQLite in the slides: SQLite Studio (lots of others)

▷ download from `https://sqlitestudio.pl`



```
select * from Books
```

Total rows loaded: 7

	Last	First	YOB	YOD	Title	YOP	Publisher	City
1	Twain	Mark	1835	1910	Huckleberry	1986	Penguin USA	NY
2	Twain	Mark	1835	1910	Tom Sawyer	1987	Viking	NY
3	Cather	Willa	1873	1947	My Antonia	1995	Library of America	NY
4	Hemingway	Ernest	1899	1961	The Sun Also Rises	1995	Scribner	NY
5	Wolfe	Thomas	1900	1938	Look Homeward, Angel	1995	Scribner	NY
6	Faulkner	William	1897	1962	The Sound and the Furry	1990	Random House	NY
7	Tolkien	John Ronald Reuel	1892	1973	The Hobbit	1937	George Allen & Unwin	UK

[07:59:01] Query finished in 0.003 second(s).

▷ Everything we can do with this, we can do with the database shell as well. (just looks nicer)

9.3 SQL – A Standardized Interface to RDBMS

Idea: To interact with RDBMSs, we need a language to describe tables to the system, so that they can be created, read, updated, and deleted. In fact while we are at it, we need a language for all RDBMS operations. The domain specific language SQL (pronounced like "sequel") fills this need. It is internationally standardized, so that it can be used as the lingua franca for all RDBMSs, insulating users and application programmers against system internals.

SQL: The Structured Query Language

▷ **Idea:** We need a language for describing all operations of a RDBMSs.

- ▷ **basics**: creating, reading, updating, deleting database components (CRUD)
- ▷ querying: selecting from and inserting into the database
- ▷ **access control**: who can do what in a database
- ▷ **transactions**: ensuring a consistent database state.

Definition 9.3.1. SQL, the structured query language is a domain-specific language for managing data held in a RDBMS. SQL instructions are directly executed by the RDBMS to change the database state or compute answers to SQL queries.

▷

We start off with a fragment of SQL that is concerned with setting up the database schema, which gives structure to the data in the database. This schema is used by the RDBMS to optimize database access.

DDL: Data Definition Language

▷ **Definition 9.3.2.** The data definition language (DDL) is a subset of SQL instructions that address the creation and deletion of database objects.

▷ **Definition 9.3.3.** The SQL statement **CREATE TABLE**⟪name⟫ (⟪coldefs⟫) creates a table with name ⟪name⟫. ⟪coldefs⟫ are column specifications that specify the columns: it is a comma-separated list of column names and SQL data type. The totality of all column specifications of all tables in a database is called the database schema.

▷ **Example 9.3.4 (Creating a Table).** The following SQL statement creates the table from ??

```
CREATE TABLE Books (
    LastN varchar(128), FirstN varchar(128),
    YOB int, YOD int, Title varchar(255), YOP int,
    Publisher varchar(128), City varchar(128)
  );
```

▷ Other **CREATE** statements exist, e.g. **CREATE** DATABASE ⟪name⟫.

▷ **Definition 9.3.5.** The SQL statement **DROP** ⟪obj⟫ ⟪name⟫ deletes the database object of class ⟪obj⟫ with name ⟪name⟫.

We have seen above that the database schema needs a data type for every column. We give an overview over the most important ones here.

SQL Data Types (for Column Specifications)

▷ **Definition 9.3.6.** SQL specifies data type for values including:

- ▷ VARCHAR (⟪length⟫): character strings, including Unicode, of a variable length is up to the maximum length of ⟪length⟫.

- ▷ BOOL truth values: **true**, **false** and case variants.
- ▷ INT: Integers
- ▷ FLOAT: floating point numbers
- ▷ DATE: dates, e.g. DATE '1999—01—01' or DATE '2000—2—2'
- ▷ TIME: time points in ISO format, e.g. TIME '00:00:00' or time '23:59:59.99'
- ▷ TIMESTAMP: a combination of DATE and TIME (separated by a blank).
- ▷ CLOB (⟪length⟫) (character large object) up to (typically) 2GiB
- ▷ BLOB (⟪length⟫) (binary large object) up to (typically) 2GiB

We now come to the SQL commands for inserting content into the database tables we have created above. This is quite straight-forward.

SQL: Adding Records to Tables

▷ **Definition 9.3.7.** SQL provides the **INSERT INTO** command for inserting records into a table. This comes in two forms:

1. **INSERT INTO** ⟪table⟫ **VALUES** (⟪vals⟫); where ⟪vals⟫ is a comma-separated list of values given in the order the columns were declared in the **CREATE TABLE** instruction.
2. **INSERT INTO** ⟪table⟫ (⟪cols⟫) **VALUES** (⟪vals⟫) where ⟪vals⟫ is a comma-separated list of values given in the order of ⟪cols⟫ (a subset of columns) all other fields are filled with **NULL**

▷ **Example 9.3.8 (Inserting into the Books Table).** The given the table Books from ?? we can add a record with

```
INSERT INTO Books
VALUES ('Tolkien', 'John Ronald Reuel', 1892, 1973, 'The Hobbit', 1937,
        'George Allen  Unwin', 'UK');
```

▷ **Example 9.3.9 (Inserting Partial Data).** Using the second form of the **INSERT** instruction, we can insert partial data. (all we have)

```
INSERT INTO Books (FirstN, LastN, YOB, title, YOP)
VALUES ('Michael', 'Kohlhase', '1964', 'IWGS Course Notes', '2018');
```

With an insert facility, we need to be able to delete records as well, again it is straight-forward, with the exception that we have to identify which records to delete.

SQL: Deleting Records from Tables

▷ **Definition 9.3.10.** The SQL delete statement allows to change existing records.

```
DELETE FROM ⟪table⟫ WHERE ⟪condition⟫;
```

▷ **Example 9.3.11.** Deleting the record for "Huckleberry Finn".

DELETE FROM Works **WHERE** Title = 'Huckleberry Finn'

▷ ⚠ If we leave out the **WHERE** clause, all rows are deleted.

▷ **Note:** There is much more to the **WHERE** clause, we will get to that when we come to SQL querying. (see ??)

And now we come to a variant of database insertion: record update. In principle, this could be achieved by deleting the record and then re-inserting the changed one, but the update instruction presented here is more efficient.

SQL: Updating Records in Tables

▷ **Definition 9.3.12.** The SQL update statement allows to change existing records.

UPDATE ⟪table⟫
SET $\langle\langle\text{column}\rangle\rangle_1 = \langle\langle\text{value}\rangle\rangle_1$, $\langle\langle\text{column}\rangle\rangle_2 = \langle\langle\text{value}\rangle\rangle_2, \ldots$
WHERE ⟪condition⟫;

▷ **Example 9.3.13.** Updating the publisher in "Huckleberry Finn".

UPDATE Books
SET Publisher = 'Chatto/Windus', YOP = 1884, City = 'London'
WHERE Title = 'Huckleberry Finn'

▷ ⚠ If we leave out the **WHERE** clause, all rows are updated.

9.4 ER-Diagrams and Complex Database Schemata

We now come to a very important aspect of structured databases: designing the database schema and with this determining the data efficiency and computational performance of the database itself. We get glimpse of the standard tool: entity relationship diagrams here.

Avoiding Redundancy in Databases

▷ Recall the books table from ??:

LastN	FirstN	YOB	YOD	Title	YOP	Publisher	City
Twain	Mark	1835	1910	Huckleberry Finn	1986	Penguin USA	NY
Twain	Mark	1835	1910	Tom Sawyer	1987	Viking	NY
Cather	Willa	1873	1947	My Antonia	1995	Library of America	NY
Hemingway	Ernest	1899	1961	The Sun Also Rises	1995	Scribner	NY
Wolfe	Thomas	1900	1938	Look Homeward, Angel	1995	Scribner	NY
Faulkner	William	1897	1962	The Sound and the Fury	1990	Random House	NY

▷ **Observation:** Some of the fields appear multiple times, e.g. "Mark Twain".

▷ ⚠ When the database grows this can lead to scalability problems:

- ▷ in querying: e.g. if we look for all works by Mark Twain
- ▷ in maintenance: e.g. if we want to replace the pen name "Mark Twain" by the real name "Samuel Langhorne Clemens".

▷ **Idea:** Separate concerns (here Authors, Works, and Publishers) into separate entities, mark their relations.

- ▷ Develop a graphical notation for planning
- ▷ Implement that into the database

After this discussion on why we need to design an efficient database schema to the entity relationship diagram themselves.

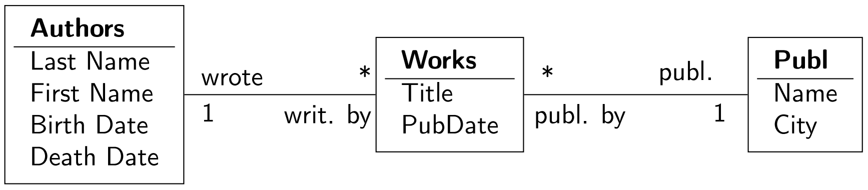
Entity Relationship Diagrams

▷ **Definition 9.4.1.** An entity relationship diagram (ERD) illustrates the logical structure of a database. It consists of entities that characterize (sets of) objects by their attributes and relations between them.

▷ **Example 9.4.2 (An ERD for Books).** Recall the Books table from ??:

LastN	FirstN	YOB	YOD	Title	YOP	Publisher	City
Twain	Mark	1835	1910	Huckleberry Finn	1986	Penguin USA	NY
Twain	Mark	1835	1910	Tom Sawyer	1987	Viking	NY
Cather	Willa	1873	1947	My Antonia	1995	Library of America	NY
Hemingway	Ernest	1899	1961	The Sun Also Rises	1995	Scribner	NY
Wolfe	Thomas	1900	1938	Look Homeward, Angel	1995	Scribner	NY
Faulkner	William	1897	1962	The Sound and the Fury	1990	Random House	NY

- ▷ **Problem**: We have duplicate information in the authors and publishers
- ▷ **Idea**: Spread the Books information over multiple tables.



Generally, a good database design is almost always worth the effort, since it makes the code and maintenance of the applications based on this database much simpler and intuitive.
We are fully aware, that this little example completely under-sells entity relationship diagrams and does not do this important topic justice. Fortunately, the DH students at FAU have the mandatory course "Konzeptuelle Modellierung" which does.
We now come to the implementation of the ideas from the entity relationship diagrams. The key idea is to have references between tables. These are mediated by special database columns types, which we now introduce.

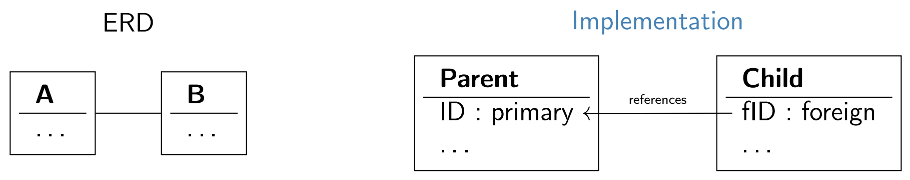
Linking Tables via Primary and Foreign Keys

▷ **Definition 9.4.3.** A column in a table can be designated as a primary key, if its values are non-null and unique i.e. all distinct.

▷ In DDL, we just add the keyword **PRIMARY KEY** to the column specification.

▷ **Definition 9.4.4.** A foreign key is a column (or collection of columns) in one table (called the child table) that refers to the primary key in another table (called the reference table or parent table).

▷ **Intuition:** Together primary keys and foreign keys can be used to link tables or (dually) to spread information over multiple tables.

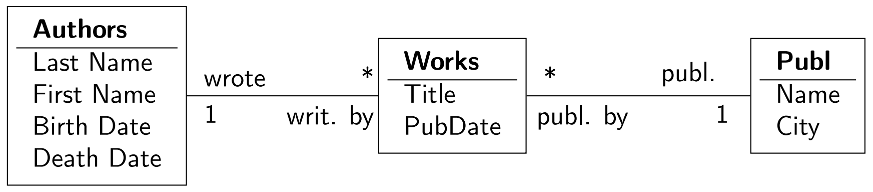


▷ **BTW:** Primary keys are great for identification in the **WHERE** clauses of SQL instructions.

We now fortify our intuition on primary and foreign keys by taking up ?? again.

Linking Tables via Primary and Foreign Keys (Example)

▷ **Example 9.4.5.** Continuing ??, we now implement



by introducing primary keys in the Authors and Publishers tables and referencing them by foreign keys in the Works table.

```
CREATE TABLE Authors (AuthorID int PRIMARY KEY,
    LastN varchar(128), FirstN varchar(128), YOB int, YOD int);

CREATE TABLE Publishers (PublisherID int PRIMARY KEY,
    Name varchar(128), City varchar(128));

CREATE TABLE Works (
    Title varchar(255), YOP int, AuthorID int, PublisherID int,
    FOREIGN KEY(AuthorID) REFERENCES Authors(AuthorID),
    FOREIGN KEY(PublisherID) REFERENCES Publishers(PublisherID));
```

Linking Tables via Primary and Foreign Keys (continued)

▷ **Example 9.4.6 (Inserting into the Works Table).** The given the tables Works Authors, and Publishers from **??** we can add a record with

```
INSERT INTO Authors VALUES (1, 'Twain', 'Mark', 1835, 1910);
INSERT INTO Publishers VALUES (1, 'Penguin USA', 'NY');
INSERT INTO Works VALUES ('Huckleberry Finn', 1986, 1, 1);

INSERT INTO Publishers VALUES (2,'Viking', 'NY');
INSERT INTO Works VALUES ('Tom Sawyer', 1987, 1, 2);
```

Note: We have introduced new integer-typed columns for the primary key in the Authors and Pubishers tables. In principle, we could have designated any existing column as a primary key instead, if we were sure that the entries are unique – in our case an unreasonable assumption, even for the publishers.

We have also chosen not to introduce a primary key in the Works table, which is probably a design mistake in the long run, because this would be very important to have for deletions and updates.

9.5 RDBMS in Python

Let us now see how we can interact with SQLite programmatically from Python instead of from the SQLite shell or the database browser.

Using SQLite from Python

▷ We will use the PySQLite package

- ▹ install it locally with pip install pysqlite for Python 3.
- ▹ use **import** sqlite3 to import the library in your programs.

▷ Typical Python program with sqlite3:

```
import sqlite3
# Open database connection
db = sqlite3.connect(⟪DBname⟫)
# prepare a cursor object using cursor() method
cursor = db.cursor()
# execute SQL commands using the execute() method.
cursor.execute("⟪SQL⟫")
⟪dataprocessingcode⟫
# make sure data reaches disk
db.commit()
# disconnect from server
db.close()
```

We will assume this as a wrapper for all code examples below.

The script schema shows the normal way of setting up the interaction with a database using sqlite3:

1. We connect to the database by specifying the database file in which the data is kept. Normally, this will be a file on the local file system, but we can also use the string :memory: which results

in an in-memory database (no persistent storage). The result of the sqlite3.connect function is a database object db.

2. Then we create a cursor object cursor (cf. slide 243 for more details) by using the cursor method of the datebase object db.

3. Then we execute SQL instructions via cursor.execute and do the data processing we need for our application.

4. To make sure that the changes we made to the database are actually reflected on disk in the database file ⟨⟨DBName⟩⟩, we commit the changes to disk via db.commit().

5. Finally, we close the database connection via the db.close method to make sure that all our changes have reached the database file.

We will now put this schema to use using **??** as a basis.

Creating Tables in Python

▷ **Example 9.5.1.** Creating the table of **??**

```
import sqlite3
# our database file
database = "C:\\sqlite\db\books.db"
# a string with the SQL instruction to create a table
create = """CREATE TABLE Books (
            LastN varchar(128), FirstN varchar(128), YOB int, YOD int,
            Title varchar(255), YOP int,Publisher varchar(128), City varchar(128));"""
insert1 = """INSERT INTO Books
              VALUES ('Twain', 'Mark', '1835', '1910', 'Huckleberry Finn', '1986',
                      'Penguin USA', 'NY');"""
insert2 = """INSERT INTO Books
              VALUES ('Twain', 'Mark', '1835', '1910', 'Tom Sawyer', '1987',
                      'Viking', 'NY');"""
# connect to the SQLIte DB and make a cursor
db = sqlite3.connect(database)
cursor = db.cursor()
# create Books table by executing the cursor
cursor.execute("DROP TABLE Books;")
cursor.execute(create)
cursor.execute(insert1)
cursor.execute(insert2)
db.commit() # commit to disk
db.close() # clean up by closing
```

In this example we first create an SQL instruction as a string, so that we can give them as arguments to the cursor.execute method conveniently.
Note that cursor.execute only executes a single SQL instructions (for safety reasons; see slide 246 – why does this help there?).
Note that we drop the Books table before (re)creating it, to be sure that we have the right structure and avoiding errors, when we run the Python script above twice. An alternative would have been to use CREATE TABLE IF NOT EXISTS, which only creates the table if there is none. But in our example here, where we directly fill the table, dropping any old tables with the name Books seems the right thing to do.
There is an issue that sometimes baffles beginners: I have created a table, inserted lots of data into it, closed the database, and the next time I connect to the database, it is empty ⇝ very annoying.

To understand this phenomenon, we have to understand a bit more how databases like SQLite work and the tradeoffs face when working working with such systems.

To commit or not to commit?

- ▷ **Recall:** SQLite computes with tables in memory but uses files for persistence.
- ▷ **Also Recall:** Memory access is 100-10.000 times as fast as file access.
- ▷ **Idea 1:** Keep tables in memory, write to file only when necessary.
- ▷ **Idea 2:** Give the user/programmer control over when to write to file
  - ▷ db = sqlite3.connect(⟪file⟫) connects to ⟪file⟫, but computes in memory,
  - ▷ db.commit() writes in-memory changes to ⟪file⟫.
- ▷ **Problem:** We can have multiple database connections to the same database file in parallel, there may be race conditions and conflicts.
- ▷ **Our Solution:** Commit often enough! (your responsibility/fault)
- ▷ **General Solution:** RDBMS offer database transactions. (not covered in IWGS)
- ▷ **Lazy Solution:** Set the connection to autocommit mode: (system decides) sqlite3.connect(⟪file⟫,isolation_level = None)

Excursion: The general solution to the problem of accessing a database from multiple programs or processes in parallel is solved by a complex technology called database transactions, which allow users' to define a sensible unit of work (via begin/end bracketing) called a transaction and makes sure that the process

- behaves as if the user's process has sole access to the database system for the duration of the transaction (isolation)
- any changes made during the transaction can be rolled back if an error occurs during processing (integrity).

Transactions are an essential, but complex technology that is beyond the scope of the IWGS course. For our understanding, db.commit is essentially just the end bracket of a transaction.

9.6 Excursion: Programming with Exceptions in Python

Before we go on, we discuss how we can deal with errors in Python flexibly, so that our web application will not drop into the Python level and present the user with a stack trace.

We first introduce what errors really are in the Python context and how they are raised and handled. Then we look at what this means for our handling of database connections.

How to deal with Errors in Python

- ▷ **Theorem 9.6.1 (Kohlhase's Law).** *I can be an idiot, and I do make mistakes!*
- ▷ **Corollary 9.6.2.** *Programming languages need a good way to deal with all kinds of errors!*
- ▷ **Definition 9.6.3.** An exception is a special Python object. Raising an exception e terminates computation and passes e to the next higher level.

▷ **Example 9.6.4 (Division by Zero).** The Python interpreter reports unhandled exceptions.

```
>>> −3 / 0
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
Zerodivisionerror: division by zero
```

▷ Exceptions are first-class citizens in Python, in particular they

- ▷ are classified by their classes in a hierarchy.
- ▷ exception classes can be defined by the user (they inherit from the Exception class)

```
class DivByZero (Exception)
    pass
```

- ▷ can be raised when an abnormal condition appears

```
if denominator == 0 :
    raise DivByZero
else
    ⟪computation⟫
```

- ▷ can be handled in a **try**/**except** block (there can be multiple)

```
try:
    ⟪tentativecomputation⟫
except : ⟪err⟫1, . . ., ⟪err⟫n :
    ⟪errorhandling⟫
finally :
    ⟪cleanup⟫
```

Let us now apply Python exceptions to our situation. Here the most important source of errors is the database connection step, where a database file might be missing or a remote host with the database file offline.

Playing it Safe with Databases

▷ **Observation 9.6.5.** *Things can go wrong when connecting to a database! (e.g. missing file)*

▷ **Idea:** Raise exceptions and handle them.

▷ **Example 9.6.6.** We encapsulate a **try**/**except** block into a function for convenience

```
import sqlite3
from sqlite3 import Error
def sql_connection():
    try:
        db = sqlite3.connect(':memory:')
        print("Connection is established: Database is created in memory")
    except Error :
        print(Error)
    finally:
        db.close()
```

The sqlite3 package provides its own exceptions, which we import separately. Other errors can be handled in additional **except** clauses.

9.7 Querying and Views in SQL

So far we have created, filled, and possibly updated databases, but we have not done anything useful with them. That is the realm of querying in SQL, which we will now come to.
We will first cover SQL querying from a single table. There are many variants of the **SELECT**/**FROM**/**WHERE** instruction. We explain the most commonly used ones.

SQL Querying: The SELECT Statement

- ▷ SQL uses the **SELECT** instruction for retrieving data from a database.
- ▷ **SELECT** ⟪columns⟫ **FROM** ⟪table⟫ returns all records from ⟪table⟫ restricted to the fields from ⟪columns⟫.
- ▷ **Definition 9.7.1.** The select instruction is the basic SQL query.
- ▷ **Example 9.7.2.** **SELECT** Title, YOP **FROM** Books;

```
Huckleberry Finn|1986
Tom Sawyer|1987
My Antonia|1995
The Sun Also Rises|1995
Look Homeward, Angel|1995
The Sound and the Furry|1990
The Hobbit|1937
```

- ▷ **SELECT DISTINCT** removes duplicate values
- ▷ **SELECT** $*$ **FROM** ⟪table⟫ returns all records from ⟪table⟫.
- ▷ **SELECT** ⟪columns⟫ **FROM** ⟪table⟫ **WHERE** ⟪cond⟫ returns all records that match condition ⟪cond⟫
- ▷ **Example 9.7.3.** **SELECT** FirstN, LastN **FROM** Books **WHERE** YOP = 1995;

```
Willa|Cather
Ernest|Hemingway
Thomas|Wolfe
```

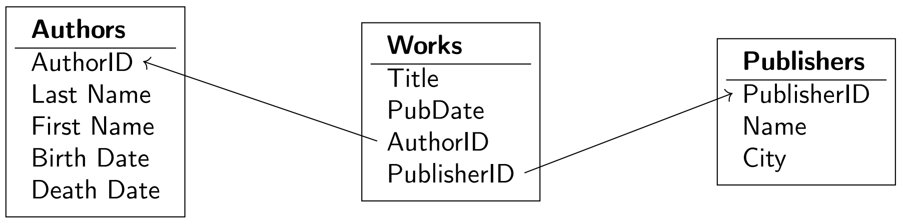
- ▷ **SELECT** ⟪columns⟫ **FROM** ⟪table⟫ **ORDER BY** ⟪colums⟫ orders the results by ⟪columns⟫
- ▷ **Example 9.7.4.** Ordering can be ascending (**ASC**) or descending (**DESC**)
  SELECT FirstN, LastN **FROM** Books **ORDER BY** LastN **ASC**, YOP **DESC**;

There are some more variants, for instance we can add a **GROUP BY** clause, which allows to group the result table according to various conditions.
We now generalize SQL queries by combining multiple tables into a virtual table from which we aggregate the results. Snsjoin over that combine multiple tables in queries are the technique that

allows to split data into multiple tables in the first place: we can re recreate the “original big table” via a query.
We will restrict ourselves to the simplest kind of table join: the inner join below. There are quite a few variants of joins; we refer the reader to the literature on them.

Joining Tables in Queries

- ▷ **Problem:** We can query single tables, how cross-table queries? E.g. in



- ▷ **Idea:** Virtually join tables for the query! (as if we had the large books table)
- ▷ **Definition 9.7.5.** A table join (or simply join) is a means for combining columns from one (self join) or more tables by using values common to each.
- ▷ **Example 9.7.6.** Joining all three tables from ??.

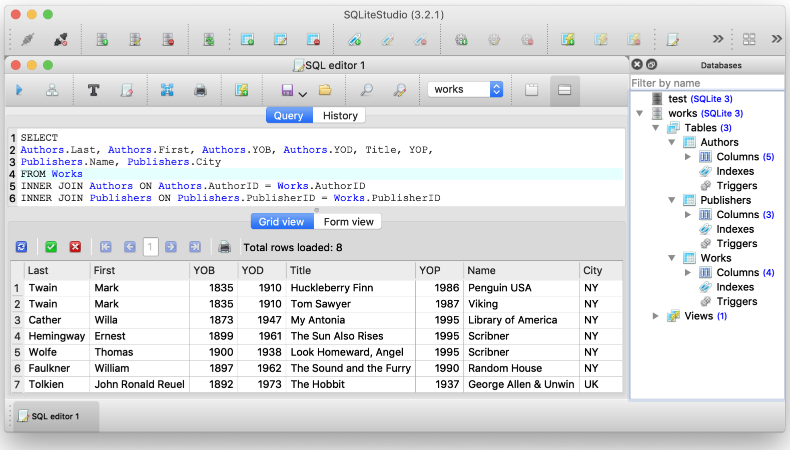
```
SELECT
  Authors.LastN, Authors.FirstN, Authors.YOB, Authors.YOD,
  Title, YOP, Publishers.Name, Publishers.City
FROM
  Works
INNER JOIN Authors ON Authors.AuthorID = Works.AuthorID
INNER JOIN Publishers ON Publishers.PublisherID = Works.PublisherID
```

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 238 2025-02-06

The key idea in the query in ?? are the **join** statements in the last two lines. They do two things: first the tell SQL to extend the Works table with data from the two tables Authors and Publishers, and second they tell SQL how the extension should work: by making sure that in the extension the records in the Works table are extended with the (unique!) record in the Authors table, that has the same AuthorID, and analogously for the records from the Publishers table. Thus the two joins implement the two arrows in the ER diagram at the top of the slide. The result of this query is displayed on the next slide.

Joining Tables in Queries (Result)

- ▷ **Example 9.7.7.**



```
SELECT
Authors.Last, Authors.First, Authors.YOB, Authors.YOD, Title, YOP,
Publishers.Name, Publishers.City
FROM Works
INNER JOIN Authors ON Authors.AuthorID = Works.AuthorID
INNER JOIN Publishers ON Publishers.PublisherID = Works.PublisherID
```

Total rows loaded: 8

	Last	First	YOB	YOD	Title	YOP	Name	City
1	Twain	Mark	1835	1910	Huckleberry Finn	1986	Penguin USA	NY
2	Twain	Mark	1835	1910	Tom Sawyer	1987	Viking	NY
3	Cather	Willa	1873	1947	My Antonia	1995	Library of America	NY
4	Hemingway	Ernest	1899	1961	The Sun Also Rises	1995	Scribner	NY
5	Wolfe	Thomas	1900	1938	Look Homeward, Angel	1995	Scribner	NY
6	Faulkner	William	1897	1962	The Sound and the Furry	1990	Random House	NY
7	Tolkien	John Ronald Reuel	1892	1973	The Hobbit	1937	George Allen & Unwin	UK

Note that the result of the query from ?? shown in ?? exactly recreates the original big Books table from ??. So we see that we have "lost nothing" by separating the data into three more efficient and less redundant – tables.

We have seen above that we can join physical database tables to larger virtual ones whenever we need it in a SQL query. This is good, but it can be made even better. RDBMS allow to persist virtual table in the database schema itself as views.

Database Views: Persisting Queries

- ▷ **Observation:** Via the join in ??, the Works table queries like the original Books table.
- ▷ **Wouldn't it be nice** If we could also insert/update into that?
- ▷ **Definition 9.7.8.** A database view (or simply view) is a virtual table based on the result set of a query. A view contains rows and columns, just like a real table. The field in a view are fields from one or more real tables in the database.
- ▷ *Remark 9.7.9.* In many RDBMS we can even insert, delete, and update records in a view, just as in any other table of the database.

  The RDBMS achieves this by automatically translating any change to the view into a set of changes to the underlying physical tables.
- ▷ ⚠ but not in SQLite. (this is an omission due to simplicity)

Remark: With views we can "have our cake and eat it too": We can make our database schema space efficient by removing redundancies using "small tables" and still have our "big tables" that make our life convenient e.g. when inserting records. Consider our Books example again: we can give the query from ?? a name and let the RDBMS treat it as a (virtual) table.

Database Views: Persisting Queries (Books Example)

▷ **Example 9.7.10.** Use the query from ?? to define a view

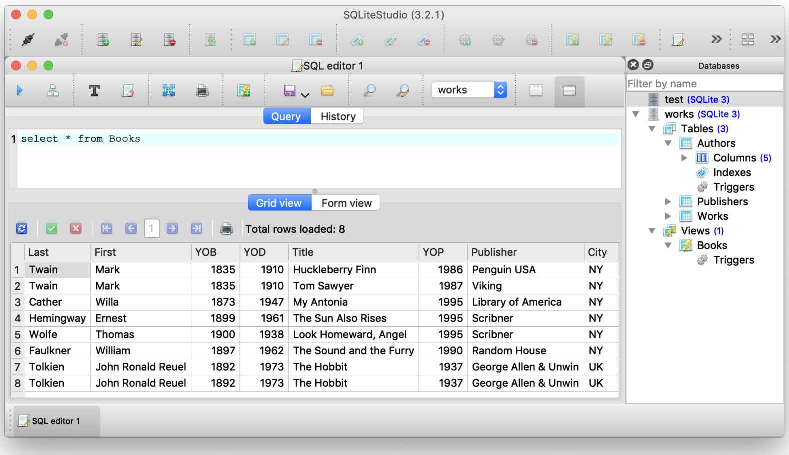
```
CREATE VIEW Books AS
SELECT
  Authors.LastN AS LastN, Authors.FirstN AS FirstN,
  Authors.YOB AS YOB, Authors.YOD AS YOD,
  Title, YOP,
  Publishers.Name AS Publisher, Publishers.City AS City
FROM
  Works
INNER JOIN Authors ON Authors.AuthorID = Works.AuthorID
INNER JOIN Publishers ON Publishers.PublisherID = Works.PublisherID
```

Use AS clauses in SELECT to specify column names.

The proof is in the pudding. We see that Books view behaves exactly like the big (unstructured) books table from above. On the right of the database browser window we can see that it is actually a view.

Database Views: Persisting Queries (Books Example)

▷ **Example 9.7.11.**



```
select * from Books
```

Total rows loaded: 8

	Last	First	YOB	YOD	Title	YOP	Publisher	City
1	Twain	Mark	1835	1910	Huckleberry Finn	1986	Penguin USA	NY
2	Twain	Mark	1835	1910	Tom Sawyer	1987	Viking	NY
3	Cather	Willa	1873	1947	My Antonia	1995	Library of America	NY
4	Hemingway	Ernest	1899	1961	The Sun Also Rises	1995	Scribner	NY
5	Wolfe	Thomas	1900	1938	Look Homeward, Angel	1995	Scribner	NY
6	Faulkner	William	1897	1962	The Sound and the Furry	1990	Random House	NY
7	Tolkien	John Ronald Reuel	1892	1973	The Hobbit	1937	George Allen & Unwin	UK
8	Tolkien	John Ronald Reuel	1892	1973	The Hobbit	1937	George Allen & Unwin	UK

9.8 Querying via Python

Now it is time to turn to understanding querying programmatically in Python. The main concept to grasp is that of a cursor.

Working with Cursors

▷ **Definition 9.8.1.** A cursor is a named object that encapsulates a set of query results in a (virtual) database table.

- ▷ To work with a cursor in sqlite3,
  - ▷ create a cursor object via the cursor method of your database object.
  - ▷ Open the cursor to establish the result set via its **execute** method
  - ▷ Fetch the data into local variables as needed from the cursor.
- ▷ The cursor class in sqlite3 provides additional methods:
  - ▷ fetchone(): return one row as an array/list
  - ▷ fetchall(): return all rows a list of lists.
  - ▷ fetchsome(⟨⟨n⟩⟩): return ⟨⟨n⟩⟩ rows a list of lists.
  - ▷ rowcount(): the number of rows in the cursor
- ▷ **Intuition:** Cursors allow programmers to repeatedly use a database query.

Again, we fortify our intuitions by making a little example: we pretty-print the some of the information by looping over result of fetching all the records from a given cursor.[4]

Extended Example: Listing Authors from the Books Table

▷ **Example 9.8.2.**

```
sql = 'SELECT FirstN, LastN, YOB FROM Books WHERE YOD < 1950;'
cursor.execute(sql)
print ('There are ',cursor.rowcount,' books, whose authors died before 1950:\n')
for row in cursor.fetchall() :
    print (row[0],' ',row[1], '; born ',row[3],'\n')
print('That is all; if you want more, add more to the database!')
```

If we have a large number of uniform SQL instructions, then we can bundle them, by iterating over a list of parameters. In the example below, we explicitly write down the list, but in applications, the list would be e.g. read from a metadata file.

Inserting Multiple Records (Example)

▷ The cursor.executemany method takes an SQL instruction with parameters and a list of suitable tuples and executes them.

▷ **Example 9.8.3.** So the final form of insertion in ?? would be to define variable with a list of book tuples:

```
booklist = [
    ('Twain', 'Mark', 1835, 1910, 'Huckleberry Finn', 1986, 'Penguin USA', 'NY'),
    ('Twain', 'Mark', 1835, 1910, 'Tom Sawyer', 1987, 'Viking', 'NY'),
    ('Cather', 'Willa', 1873, 1947, 'My Antonia', 1995, 'Library of America', 'NY'),
    ('Hemingway', 'Ernest', 1899, 1961, 'The Sun Also Rises', 1995, 'Scribner', 'NY'),
    ('Wolfe', 'Thomas', 1900, 1938, 'Look Homeward, Angel', 1995, 'Scribner', 'NY'),
    ('Faulkner', 'William', 1897, 1962, 'The Sound and the Furry', 1990, 'Random House ', 'NY'),
    ('Tolkien', 'John Ronald Reuel', 1892, 1973, 'The Hobbit', 1937,'George Allen  Unwin', 'UK')]
```

[4]EdNote: MK: show the results

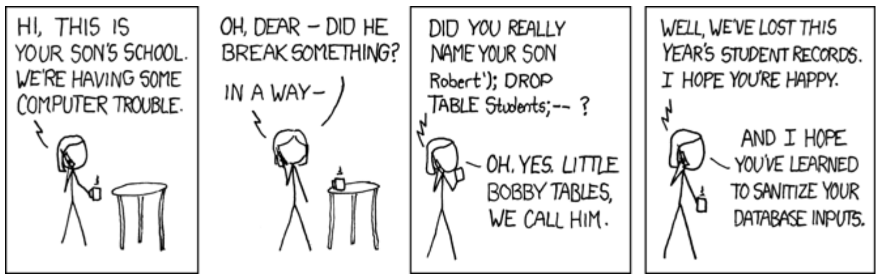
and then insert it via a call of cursor.executemany:

```
cursor.executemany('INSERT INTO Books VALUES (?,?,?,?,?,?,?,?)',booklist)
```

Now that we understand how to deal with databases programmatically, we can come to a real-world menace: SQL injection attacks. A large portion of the "hacking" events, where a database is taken over by malicious agents are based – at least in part – on such a technique. Therefore it is important to understand the basic principles involved, if only to understand how we can safeguard against them – see e.g. slide 247 below.

Beware of the Python/SQLite Interaction

▷ **What have we learned?:** At least you now understand the following web comic: (https://xkcd.com/327/)


HI, THIS IS YOUR SON'S SCHOOL. WE'RE HAVING SOME COMPUTER TROUBLE.
OH, DEAR – DID HE BREAK SOMETHING?
IN A WAY–
DID YOU REALLY NAME YOUR SON Robert'); DROP TABLE Students;-- ?
OH. YES. LITTLE BOBBY TABLES, WE CALL HIM.
WELL, WE'VE LOST THIS YEAR'S STUDENT RECORDS. I HOPE YOU'RE HAPPY.
AND I HOPE YOU'VE LEARNED TO SANITIZE YOUR DATABASE INPUTS.

▷ **Definition 9.8.4.** We call this an SQL injection attack.

▷ **Hint:** Imagine a web application where you add student names for enrolment.

```
name = input("Please enter student name: ")
cursor.execute(f"INSERT INTO Students VALUES (... ,{Name}, ...);")
```

For the input Robert'); DROP TABLE Students; this has a Python line generates and executes the SQL instructions

```
INSERT INTO Students VALUES (..., 'Robert'); DROP TABLE Students;
```

Now we can understand why the restriction of cursor.**execute** to only one SQL instruction enhances security of the code: The hypothetical cursor.**execute**('INSERT ...') command expects one instruction, but with the parameter substitution in the f string gets two. This would have raised an exception and saved the school administration.
Finally we come back to the topic of preventing SQL injection attacks. We had seen that these occur when we build the argument string for a cursor.execute call. While the single-instruction-restriction of is some help, it is not enough. We essentially have to remove all the SQL instructions from any input string we substitute with. Fortunately, SQL is standardized, so we can implement that once and for all.

SQLite3 Parameter Substitution

▷ **Observation 9.8.5.** *We often need variables as parameters in* cursor.execute.

▷ **Example 9.8.6.** In ?? we can ask the user for a year.

▷ **The python way** would be to use f strings

```
year = input('Books, whose author died before what year?')
sql = f'SELECT FirstN, LastN, YOB FROM Books WHERE YOD < {year}'
cursor.execute(sql) # ⚠ never use f−strings here −−> insecure
```

But this leads to vulnerability by SQL injection attacks. (⇝ Bobby Tables)

▷ **Definition 9.8.7.** sqlite3 supplies a parameter substitution that SQL sanitizes parameters (removes problematic SQL instructions).

▷ **The sqlite3 way** uses parameter substitution (multiple ? possible ⇝ tuple)

```
year = input('Books, whose author died before')
select = 'SELECT Title FROM Books WHERE YOD < ?'
cursor.execute(select,(year,))
```

or in the "named style" ⇝ order-independent (argument is a dictionary)

```
century = input('Century of the books?')
select = 'SELECT Title, YOP FROM Books WHERE YOP <= :start AND YOP > :end'
datadict = {'start': (century − 1) * 100, 'end': century * 100}
cursor.execute(select,datadict)
```

9.9 Real-Life Input/Output: XML and JSON

We have seen how we can use Python programs to fill database tables programmatically. But the treatment above did not map the dominant data management practices. In practice, databases are filled from various information sources, usually CSV, XML, and JSON files. Conversely, the data from a database is often exported to the same file formats for backup and/or communication.

To show the practices, we will see how to import data from an XML file into a data base, and how to export data as JSON in Python; the latter is an important technique for modern web applications.

Filling a DB from via XML (Specification)

▷ **Idea:** We want to make a database based web application for NYC museums.

▷ **Recall** the public catalog from ??, the XML file is online at
`https://data.cityofnewyork.us/download/kcrm-j9hh/application/xml`

```
<?xml version="1.0" encoding="UTF−8"?>
<museums>
  <museum>
    <name>American Folk Art Museum</name>
    <phone>212−265−1040</phone>
    <address>45 W. 53rd St. (at Fifth Ave.)</address>
    <closing>Closed: Monday</closing>
    <rates>admission: $9; seniors/students, $7; under 12, free</rates>
    <specials>
      Pay−what−you−wish: Friday after 5:30pm;
      refreshments and music available
    </specials>
  </museum>
  <museum>
```

```
    <name>American Museum of Natural History</name>
    <phone>212—769—5200</phone>
    <address>Central Park West (at W. 79th St.)</address>
    <closing>Closed: Thanksgiving Day and Christmas Day</closing>
```

▷ **Idea:** We need Python program that

- ▷ provides a SQLite database with a table 'museum' with columns 'name', 'phone', . . . , 'specials' of appropriate type
- ▷ reads the XML file from the URL above and fills the table.

▷ **Possible Enhancement:** Encapsulate the functionality into a function, then we could run this program each night and keep the database up to date.

Let us actually implement this idea – after all, we have already seen all the building blocks already. The program itself is relatively straightforward; we go through the process step by step.

Filling a DB from via XML (Implementation)

▷ **Libraries:** urllib [UL] to retrieve the file and lxml [LXMLa] to parse it.

```
from lxml import etree
from urllib.request import urlopen
url = 'https://data.cityofnewyork.us/download/kcrm—j9hh/application/xml'
document = urlopen(url).read()
tree = etree.fromstring(document)
```

We now have a (large) XML tree in tree!

▷ Collect all the XML tags in all the museums (for the column names)

```
tags = []
for museum in tree:
    for info in museum:
        if info.tag not in tags:
            tags.append(info.tag)
```

▷ We create the SQLite database as discussed in slide 233.

▷ Then we assemble a table specification in a string columns:

```
columns = ""
for cn in tags:
    # All columns have their name and type TEXT
    columns += f", {cn} TEXT"
```

▷ Create the Museums table from the specification in columns

```
cursor.execute("DROP TABLE IF EXISTS Museums;")
cursor.execute(f"""CREATE TABLE Museums
                   (Id INTEGER PRIMARY KEY {columns});""")
```

▷ Now the most important part: We fill the database

```
for museum in tree:
    # Find and sanitise the contents of all child nodes of this museum.
    values = []
    for tag in tags:
        if museum.find(tag) != None:
            values.append(str(museum.find(tag).text).strip())
        else:
            values.append("−")

    # Insert the data for this museum into the database.
    cols = str(tuple(tags))

    # We need a tuple of one ? for each column.
    vals = "(" + ("?, " ∗ len(tags))[:−2] + ")"

    insert = f"INSERT INTO Museums {cols} VALUES {vals}"
    cursor.execute(insert, tuple(values))
```

▷ We finalize the transaction as discussed in slide 233.

If we want to get a sense of the size and complexity of the complete program, we can look at it in one pice here:

The complete code in one block – a mere 51 lines

```
import sqlite3
from lxml import etree
from urllib.request import urlopen

# Download the XML file and Parse it
url = 'https://data.cityofnewyork.us/download/kcrm−j9hh/application/xml'
document = urlopen(url).read()
tree = etree.fromstring(document)

# First run−through of the XML: Collect the info types there,
tags = []
for museum in tree:
    for info in museum:
        if info.tag not in tags:
            tags.append(info.tag)

# Next, create database accordingly. First assemble a columns string.
columns = ""
for cn in tags:
    # All columns have their name and type TEXT
    columns += f", {cn} TEXT"

# Then, make the Museums table using that.
db = sqlite3.connect("./museums.sqlite")
cursor = db.cursor()
cursor.execute("DROP TABLE IF EXISTS Museums;")
cursor.execute(f"""CREATE TABLE Museums
                  (Id INTEGER PRIMARY KEY {columns});""")

# Lastly, fill database.
for museum in tree:
    # Find and sanitise the contents of all child nodes of this museum.
    values = []
    for tag in tags:
        if museum.find(tag) != None:
            values.append(str(museum.find(tag).text).strip())
        else:
            values.append("−")

    # Insert the data for this museum into the database.
    cols = str(tuple(tags))
    # We need a tuple of one ? for each column.
    vals = "(" + ("?, " * len(tags))[:−2] + ")"

    insert = f"INSERT INTO Museums {cols} VALUES {vals}"
    cursor.execute(insert, tuple(values))

# Finalise Transaction
db.commit()
db.close()
```

We will use the output direction of the envisioned museums web application to introduce another standard data representation format: JSON the preferred data interchange format for web applications.

JSON — JavaScript Object Notation

▷ **Definition 9.9.1.** JSON (JavaScript Object Notation) is an open standard file format for interchange of structured data. JSON uses human readable text to store and transmit data objects consisting of attribute–value pairs and sequences.

▷ ⚠ JSON is very flexible, there need not be a regularizing schema.

▷ **Intuition:** JSON is for JavaScript as (nested) dictionaries are for Python.

- ▷ The browser can directly read JSON and use it via JavaScript.
- ▷ ⇝ AJAX $\widehat{=}$ JavaScript can query the backend for JSON data to update parts of the DOM. (lightweight interaction)

▷ **Consequence:**
JSON is the dominant interchange format for web applications.

▷ **Another Intuition:** JSON objects are like database records, but less rigid.

▷ **Idea:** Build a special JSON database. (JSON I/O; efficient storage)

▷ **Definition 9.9.2.** mongoDB is the most popular NoSQL database system. (no SQL inside)

As always, we will now look at how we can deal with with the newly introduced concept in Python. As always there is a special library that does nearly all the work; here it is (obviously named) json library. It smoothes over the syntactic differences between Python dictionaries and JSON objects.

Dealing with JSON in Python

▷ ⚠ Even though JSON concepts and syntax are similar to Python dictionaries, there are (subtle) differences.

▷ **Concretely:** Python allows more data types in dictionaries, e.g.

Python	JSON equivalent
True	true
False	false
float	Number
int	Number
None	null
dict	Object
list	Array
tuple	Array

- But these differences are systematic and can be overcome via the json library [JS].
  - json.dumps(⟨⟨dict⟩⟩) takes a Python dictionary ⟨⟨dict⟩⟩, produces a JSON string.
  - json.loads(⟨⟨json⟩⟩) takes a JSON string ⟨⟨json⟩⟩, produces a Python dictionary.

There are many ways to control the output (pretty-printing), see [JS].

We now give an JSON export program for the NYC Museums database for reference. All the technologies in this program have been covered above, so we just show it for self-study.

JSON Output for the NYC Museums DB

- **Libraries:** json for JSON [JS] and sqlite3 for the database.

```
import json
import sqlite3
```

- Connect to the SQLite database as usual and query the database for everything

```
db = sqlite3.connect("./museums.sqlite")
cursor = db.cursor()
cursor.execute("SELECT * FROM Museums;")
```

- Initialize a dictionary and the list of Museums column names

```
data = {}
data['museums'] = []
columns = ['name', 'phone', 'address', 'closing', 'rates', 'specials']
```

- For each of the rows in the Museums table build a row dictionary

```
for row in cursor.fetchall():
    # Generate a dictionary with columns as keys and entrys as values.
    rowdict = { columns[n] : row[n] for n in range(6) }

    # Add that dictionary to the JSON data structure.
    data['museums'].append(rowdict)
```

- Dump the data dictionary as JSON into a file

```
with open('museums.json', 'w') as outfile:
    json.dump(data, outfile)
```

- Close the database as usual.

We set the list variable columns manually for convenience. Of course, if the database schema should change, we have to adapt columns in or programs. Therefore a better way to handle this would be to get this information from the database itself, which we could do by the following command in SQLite:

```
PRAGMA table_info(Museums);
```

In our case, this gives us the following string, from which we can retrieve the information we need by a simple regular expression.

```
0|Id|INTEGER|0||1
1|name|TEXT|0||0
2|phone|TEXT|0||0
3|address|TEXT|0||0
4|closing|TEXT|0||0
5|rates|TEXT|0||0
6|specials|TEXT|0||0
```

But note that the PRAGMA instruction is specific to SQLite; other systems have other ways of getting to this information.

JSON Output for the NYC Museums DB

```
import json
import sqlite3

# Connect to database and query database for everything.
db = sqlite3.connect("./museums.sqlite")
cursor = db.cursor()
cursor.execute("SELECT * FROM Museums;")

# Setup soon-to-be-JSON dictionary and the necessary columns
data = {}
data['museums'] = []
columns = ['name', 'phone', 'address', 'closing', 'rates', 'specials']

# For every row in the result, do the following:
for row in cursor.fetchall():
    # Generate a dictionary with columns as keys and entrys as values.
    rowdict = { columns[n] : row[n] for n in range(6) }

    # Add that dictionary to the JSON data structure.
    data['museums'].append(rowdict)

# Write collected JSON data to file.
with open('museums.json', 'w') as outfile:
    json.dump(data, outfile)

# Close database
db.close()
```

And now we can see the result of this export – at least an initial fragment for space reasons.

JSON Example (NYC Museums)

▷ **Example 9.9.3.** The NYC museums data from ?? as JSON:
We represent the data as a "sequence" of (nested) "dictionaries"

```
[
   {"name": "American Folk Art Museum",
    "phone": "212-265-1040",
    "address": "45 W. 53rd St. (at Fifth Ave.)",
    "closing": "Closed: Monday",
    "rates": {
        "admission": "$9",
        "seniors/students": "$7",
        "under 12": "free",
    }
    "specials": "Pay—what—you—wish: Friday after 5:30pm;
                refreshments and music available"
   }
   {"name": "American Museum of Natural History",
    "phone": "212—769—5200",
    "address": "Central Park West (at W. 79th St.)"
    "closing": "Closed: Thanksgiving Day and Christmas Day"
    "rates": {
        "admission": "$16",
        "seniors/students": "$12",
        "kids 2—12": "$9",
        "under 2": "free"
    }
   }
   ...
   ]
```

9.10 Exercises

Problem 10.1 (Setting up the Database)

In this exercise we will set up our database tables. Start by cloning the KirmesDH repository[1]. The dataset consists of a directory img/, which contains images and a folder metadata/ containing CSV files. The other directories are not important for this assignment.

Familiarize yourself with the metadata format. As you can see most files employ the same columns, however some data may be missing. We will mirror the given column structure in our database.

1. In the given code skeleton, change the values of the variables metadataFolder and imageFolder at the top of the file according to your folder structure.

2. Establish a connection to the database. Use the databaseName variable.

3. Create a table with name Images in the database with the following column structure:

   - FileName, type TEXT
   - Title, type TEXT
   - Subtitle, type TEXT
   - Archive, type TEXT
   - Artist, type TEXT
   - Location, type TEXT
   - Date, type TEXT
   - Genre, type TEXT
   - Material, type TEXT
   - Url, type TEXT
   - Content, type BLOB

4. At the end of the file, commit all changes you made to the database and close it.

[1] `https://gitlab.cs.fau.de/iwgs-ss19/KirmesDH`

Run your script and open the resulting database file in the DB Browser for SQLite. Make sure that you see the Images table and that its layout is correct.

Problem 10.2 (Parsing the Input Data)

In this exercise we will parse the metadata files and extract all relevant data. Since the input data is not curated very carefully and some entries may be missing, we need to design our program as robustly as possible.

Amend the `parseMetadata` function in the given Python script for this assignment. The prepared code opens the `CSV` file and uses the `csv` library to parse it. Detailed information on the `csv.DictReader` can be found here: `https://docs.python.org/3/library/csv.html#csv.DictReader`.

In the loop do the following for each row of the file:

1. Use the `getValue` function to extract the relevant data.
2. Call the `addImage` function with the data.

Make sure that the data is parsed correctly by running your program and printing the extracted values. Assure that the program does not crash if certain data fields are not available.

Problem 10.3 (Inserting Data into the Database)

In this last exercise we fill our database with the parsed data. Before starting with this task, assure that the previous two assignments work correctly.

Complete the `addImage` function.

1. Check whether in the img/ folder a file with the specified file name exists. If yes, open and read it and store the content in the `imageData` variable.
2. Insert all data fields into the database by issuing the correct SQL command.

Run your script. Make sure it does not crash and check your database in the *DB Browser*. All values should be in the correct column. Some rows should have values in the Content column. In the *DB Browser* you can see the image when you click on the table cell.

We will now start establishing a web server, using the `bottle` framework we introduced last semester. We are building on top of the code above, so you may either continue with your own code or use the sample solution from last week as a starting point for this exercise.

For the web server we again prepared a code skeleton for you (Server_Skeleton.py and Index_Skeleton.tpl).

Problem 10.4 (Adding a Primary Key to our Table)

Our table Images from last week supports nearly all functionality we need. However currently it lacks the ability to uniquely identify a single entry, since all properties could be featured in multiple entries.

We therefore introduce primary keys. To this end, amend your Images table by adding a field Id of type INTEGER. Mark it as a primary key. When inserting into your database, you don't actually have to provide a value for the Id, since SQLite will simply use the next free number.

Problem 10.5 (Setting up our Web Server)

We will now set up a simple web server using the bottle framework. As a starting point you can use the Server_Skeleton.py and Index_Skeleton.tpl we provide you.

You might need to install the bottle package first. In your command prompt (terminal) issue the following command:

```
pip install bottle
```

You should now be able to run the provided code. Make sure you adapt the value of the variable databasename to match your database file.

After starting you can access your website by visiting the URL `http://localhost:8080/` in your browser. The content of this page is for you to implement.

We provide a route /imageraw in the getImage function. Follow the instructions in the code to try out the function and see how it works. For all operations which need to display images from the database on your website you should use this route.

Your job is to implement the index function, which is called when the home page is visited. In the end this page should display a large table where all entries of your database are listed.

1. Start by querying your database for the data you want to display. Select at least the Id, Title, Subtitle, Artist, Material and Archive of each entry. Issuing the appropriate SQL command should provide you a large list of entries. Make sure that this works before continuing.

2. Last semester we created websites in bottle by creating HTML code from python. This does not scale well to larger projects. We will therefore use bottle's own template engine, which allows you to write normal HTML documents, which you can augment with snippets of python code. You can read about the templating in the bottle documentation: `https://bottlepy.org/docs/dev/tutorial.html#templates`.

   From the index function, pass the data you queried from the database to the template function. In the Index_Skeleton.tpl file, create a HTML table. This should employ columns for each data field you queried (Title, Subtitle, etc).

   Inject python code with the appropriate syntax, which loops over the queried data and fills the table. The Archive field should be a link, which leads you the archive's website. Run your server, visit its URL and check if everything works.

3. Augment your HTML table by adding one more column called Thumbnail. This should display a small version of the digital image stored in each data entry. For this refer to the following tutorial: `https://www.w3schools.com/tags/tag_img.asp`.

   Set the thumbnail to an appropriate size (e.g. 200 pixels). As source use the /imageraw route described above. Make sure you specify the correct id for each entry.

   Test your website and enjoy it!

Now we will augment our web server by another route, which displays detailed information for a single image entry. As a reminder: The code skeleton is available on StudOn together with this assignment sheet or in the Kirmes repository. Just pull the latest version of the repo!

Problem 10.6 (Details Page)

Our overview table is nice, but we would like the user to be able to inspect a certain entry more closely. We will therefore create a new route, which displays information for a single image on its own page.

1. In your Server.py file, create a new route /details/<id:int>. Given an Id as argument, the function should query the database for this entry. If no entry with the Id can be found, use bottle's abort function to display an error with the code 404: `https://bottlepy.org/docs/dev/tutorial.html#http-errors-and-redirects`.

2. Create a new template file Details.tpl. From your python code, call the template with the information you queried from the database. In the template, write HTML code which displays the given information in a nice and easy-to-read way.

   Some information might not be available (NULL/None). Handle this case!

   Test your page by navigating to the details URL for some example image, e.g. `http://localhost:8080/details/27`. Make sure, that all data is displayed correctly.

3. On the details page, also display the image in full size. You may again use the /imageraw/id route from last week as source.

4. Amend your Index.tpl from last week in the following way: Each image thumbnail in the table should be a link (<a href=...>), which leads to the details page of this respective entry, i.e. by clicking on the thumbnail of image 27 your website should navigate to the URL `http://localhost:8080/details/27`.

Problem 10.7 (New Entries and Editing)

The next step to creating a useful web application is to allow the user to insert new entries and edit existing ones.

We have prepared the code for adding new entries for you in this week's Server.py skeleton. If you want to continue with your own code, you can copy the functions new, submitNew and getValue from the skeleton to your own file. Also copy the file New.tpl to your directory. In your Index.tpl, add a link at the top of the page, which leads to the /new route.

Familiarize yourself with the given code. Understand how it works and how the data flows.

Editing entries is similar to adding new ones. Both require a form to insert data, which is then sent to a routine to handle the database calls. For the form the only difference is that some data is already filled out. For now we will only allow editing of the metadata, not the digital image itself. Your edit form does not need to allow changing the digital image.

1. Create a new file Edit.tpl. Take the given New.tpl as a starting point. Since we do not want to allow changing the image for now, you can omit the Image input field.

2. In your python code, create a new route /edit/<id:int>. In the function, query the database for the entry with the given id. Since this is the same operation as in the /details/ route, you can reuse this code. Call the Edit.tpl template with your queried data.

3. For fields, which are already filled out, the form should display the current value. To this end, refer to the value attribute of the <input> fields. Test your page by navigating to the URL of an example entry, e.g. `http://localhost:8080/edit/27`. Make sure the available data is displayed correctly.

4. The key difference to the New.tpl form is, that we already have an entry, i.e. an Id. This must be passed via the form to the function, which handles the database update.

   HTML forms allow hidden fields, which look like this:

   ```
   <input type='hidden' name='id' value='{{id}}'>
   ```

   Since the field is set to hidden, it will not show up on the web page. Nevertheless, its value (the id) will be sent with the rest of the filled out form data. Use the above code to add the Id to the form.

5. Create another route /submit_edit of type POST. Refer to the given /submit_new route for details. Obtain all data from the input form. Afterwards, issue an SQL UPDATE command to update the entry with the given Id and provide the values from the form.

   In the end, use bottle's redirect functionality to navigate to the details page of the edited entry. Again, refer to the submitNew function for details.

6. In the Edit.tpl file, make sure that the form action is set to the correct route.

7. On the details page, create a link Edit, which leads to your /edit/<id> route.

Chapter 10

Project: A Web GUI for a Books Database

In this chapter we will pull together the technologies we have learned into a simple web application project. We will do so in multiple setups. We first make a bare-bones application (see ??) and then step by step extend it with new features:

- Ajax techniques for selectively loading page fragments: ??
- Access control and management: ??
- Deploying Python applications as programs: ??

Bricolage Programming: With this project we want to demonstrate a common practice of moderen programming: pulling together program fragments or solution ideas from various sources (e.g. the IWGS lecture notes or various tutorials or even answers from stack overflow `https://stackoverflow.com`, a question and answer site for professional and enthusiast programmers) and then adapting them to the current project and fitting them together into a coherent program that works as expected.

This approach to programming is often called "bricoleur style" [Tur95] or bricolage programming because it relies on handicraft-like tinkering with pieces of existing materials.

Contrary to what many classical programming courses still insinuate they seem to say that you have to know everything before you can start with a project – the advent of the internet with its multitude of high-quality programming related resources has made bricoleur style programming effective and efficient.

Actually, bricolage is a technique that should be leaned and adopted as a tool, especially for part-time programmers as practitioners in the digital humanities tend to be.

The web application project in this chapter is a bricolage project, only that we have almost all the ideas in the IWGS course notes already and we do not have to google for them on the web.

10.1 A Basic Web Application

We bring together all we have learned into a basic web application that allows to list all the books in a database, as well as add, edit, and delete book records.

We use our running example of the books table as a basis, and add a web application layer via the bottle WSGI server-side scripting framework in Python.

We have intentionally kept the application very simple, so that it can serve as the basis of other projects. The full source is available at `https://gl.mathhub.info/courses/FAU/IWGS/blob/master/source/databases/code/books-app.py`. The respective template files are siblings.

Building a full Web Application with Database Backend

▷ **Observation 10.1.1.** *With the technology in ?? and ?? we can build a full web application in less than*

- ▷ *100 lines of Python code and* *(back-end/routes)*
- ▷ *less than 70 lines of HTML template files.* *(front end)*

▷ **Functionality:** Manage a database of books, in particular: (e.g. your library at home)

- ▷ add a new book to the database
- ▷ delete a book from the database
- ▷ update (i.e. change) an existing book

▷ The source is at `https://gl.mathhub.info/courses/FAU/IWGS/blob/master/source/booksapp/code/books-app.py`.

Now, if you download the file books−app.py and all the sibling template files ∗.tpl at `https://gl.mathhub.info/courses/FAU/IWGS/blob/master/source/booksapp/code/`, you can start the application from the terminal by typing python3 books−app.py. This will yield something like

```
> python3 books−app.py
Bottle v0.12.18 server starting up (using WSGIRefServer())...
Listening on http://localhost:8080/
Hit Ctrl−C to quit.
```

So enter the url `http://localhost:8080/` into the URL bar of your browser, and test the setup. But let us return to the implementation of the web application.
We do the usual things to set up the web application: we load the libraries, connect to the data base, and so on.

The Books Application: Setup

▷ We have already seen how to set up the database in slide 245.

```
import sqlite3
from sqlite3 import Error
from bottle import route, run, debug, template, request, get, post

# our database file
database = "books.db"
db = sqlite3.connect(database)
```

▷ But we want to receive result rows as dictionaries, not as tuples, so we add

```
db.row_factory = sqlite3.Row
```

▷ We give ourselves a cursor to work with

```
cursor = db.cursor()
```

▷ We start the bottle server

```
run(host='localhost', port=8080, debug=True)
```

▷ And of course, we eventually commit and close the database in the end

```
db.commit()
db.close()
```

The next step is to create a table for the books. This is a completely standard SQL CREATE statement which we execute in the cursor we have established during setup.

The Books Application: Backend

▷ We specify the database schema and create the Books table

```
bookstable = """
CREATE TABLE IF NOT EXISTS Books (
    Last varchar(128), First varchar(128),
    YOB int, YOD int, Title varchar(255), YOP int,
    Publisher varchar(128), City varchar(128)
  );
  """
cursor.execute(bookstable)
```

The next step is strictly optional. But it is so annoying to have to start with an empty database when the web application first comes up. So we provide a list of seven books. But, if we are not careful, these books will be inserted into the database every time we start up the application. Recall that we did not drop the Books table in the code above.

The Books Application: Books to Play With

▷ Data about books as a Python list of 8-tuples:

```
initialbooklist = [
    ('Twain', 'Mark', 1835, 1910, 'Huckleberry Finn', 1986, 'Penguin USA', 'NY'),
    ('Twain', 'Mark', 1835, 1910, 'Tom Sawyer', 1987, 'Viking', 'NY'),
    ('Cather', 'Willa', 1873, 1947, 'My Antonia', 1995, 'Library of America', 'NY'),
    ('Hemingway', 'Ernest', 1899, 1961, 'The Sun Also Rises', 1995, 'Scribner', 'NY'),
    ('Wolfe', 'Thomas', 1900, 1938, 'Look Homeward, Angel', 1995, 'Scribner', 'NY'),
    ('Faulkner', 'William', 1897, 1962, 'The Sound and the Fury', 1990, 'Random House ', 'NY'),
    ('Tolkien', 'John Ronald Reuel', 1892, 1973, 'The Hobbit', 1937,'George Allen  Unwin', 'UK')]
```

▷ If the Books table is empty, we fill it with the tuples in initialbooklist:

```
row = cursor.execute('SELECT * FROM Books LIMIT 1').fetchall()
if not row:
    cursor.executemany('INSERT INTO Books VALUES (?,?,?,?,?,?,?,?)',initialbooklist)
```

▷ **Idea:** To find out if the table is empty (surprisingly clumsy)

- ▷ we fetch a list with at most one row (LIMIT 1);
- ▷ if Books is empty, row is the empty list which evaluates to false in a conditional.

In a more complete version of the books application we would probably have used a keyword argument like ––initbooks to the program. We will cover command line parsing – the technology that enables behavior modifiers – ??. The next task is to create a route for the main page of the application, i.e. the page booksapp.py serves at `http://localhost:8080/`. We want a listing of all the books in the database in a table.

The Books Application Routes: The Application Root

▷ We only need to add the bottle routes for the various sub pages.

▷ **The main page:** Listing the book records in the database

```
@route('/')
def books():
    query = 'SELECT rowid,Last,First,YOB,YOD,Title,YOP,Publisher,City FROM Books'
    cursor.execute(query)
    booklist = cursor.fetchall()
    return template('books',books=booklist,num=len(booklist),cols=cols)
```

▷ This uses the following templates: the first generates a table of books from the template file books.tpl

```
<p>There are {{num}} books in the database</p>
<table>
  % include('th.tpl', cols=cols)
  % for book in books : include('book.tpl',**book,cols=cols) end
  <tr><th><a href="/add"><button>add a book</button></a></th></tr>
</table>
```

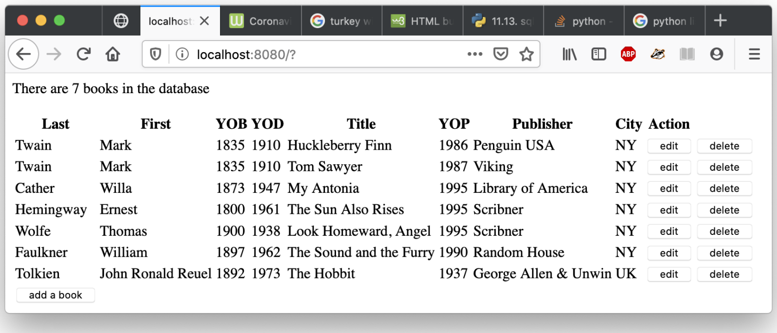
The backend of this is very simple: we fire up a simple SQL query that selects all the records from the Books table. As we configured the database connection to return database records as Python dictionaries, the variable booklist variable is a list of data dictionaries, which we can feed to the STPL template books.tpl, which creates the return page for `http://localhost:8080/`. This page consists of a paragraph which reports on the number of books in the database and then a table which is built up from

1. a table header which is simply imported from a template file th.tpl
2. a body, which is created by iterating over booklist, feeding each row – a Python dictionary – to the template book.tpl as keyword arguments via the double star operator, and
3. a table row with a link to the add route for adding new books.

Before we show the nested templates, let us inspect the result:

The Books Application Root: Result

▷ Here is the page of the books application in its initial state.



There are 7 books in the database

Last	First	YOB	YOD	Title	YOP	Publisher	City	Action	
Twain	Mark	1835	1910	Huckleberry Finn	1986	Penguin USA	NY	edit	delete
Twain	Mark	1835	1910	Tom Sawyer	1987	Viking	NY	edit	delete
Cather	Willa	1873	1947	My Antonia	1995	Library of America	NY	edit	delete
Hemingway	Ernest	1800	1961	The Sun Also Rises	1995	Scribner	NY	edit	delete
Wolfe	Thomas	1900	1938	Look Homeward, Angel	1995	Scribner	NY	edit	delete
Faulkner	William	1897	1962	The Sound and the Furry	1990	Random House	NY	edit	delete
Tolkien	John Ronald Reuel	1892	1973	The Hobbit	1937	George Allen & Unwin	UK	edit	delete
add a book									

Indeed we have the report on the number of books and a table which ends in an "add a book" link. The table header and rows contain the seven data cells and two more for possible actions on the database records. The next two templates are responsible for that; they are called in the books template above.

The Books Application Root: More Templates

▷ **Recall:** The books.tpl template file

```
<p>There are {{num}} books in the database</p>
<table>
  % include('th.tpl', cols=cols)
  % for book in books : include('book.tpl',**book,cols=cols) end
  <tr><th><a href="/add"><button>add a book</button></a></th></tr>
</table>
```

that generates this result via the following two templates:

▷ It inserts the table header via th.tpl:

```
% for col in cols:
  <th>{{col}}</th>
% end
<th rowspan="2">Action</th>
```

▷ and iterates over the list of books, using the template file book.tpl:

```
<tr>
  <td>{{Last}}</td><td>{{First}}</td><td>{{YOB}}</td><td>{{YOD}}</td>
  <td>{{Title}}</td><td>{{YOP}}</td><td>{{Publisher}}</td><td>{{City}}</td>
  <td><a href="/edit/{{rowid}}"><button>edit</button></a></td>
  <td><a href="/delete/{{rowid}}"><button >delete</button></a></td>
</tr>
```

▷ **Row Id Trick:** Note the slightly subtle use of the rowid column in this template. It is (only) used in the two action buttons to specify which book to add/edit.

The template th.tpl is completely elementary, book.tpl is called with keyword arguments whose values substituted for the {{⟨⟨key⟩⟩}} template variables. The last two columns in the table are the action links that point to the add and delete routes we present next.
The "add a book" functionality is distributed over two routes: a GET route for the path /add/

and a POST route for the same path. The first is responsible for showing the input form, whereas the second parses the POST request generated by the first one and fills the database with the results. Let us look at the implementation in detail.

The Books Application Routes: Adding Book Records

▷ We add a route for adding a books record (for the add button)

```
@get('/add')
def add():
    return template('add',cols=cols)
```

Note that this is the route for the GET method on the path /add.

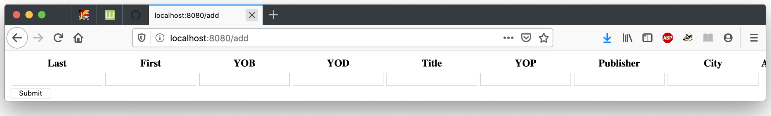
▷ This uses the template file add.tpl:

```
<form action="/add" method="post">
  <table>
    % include('th.tpl', cols=cols)
    <tr>
        % for td in cols:
        <td><input type="text" name="{{td}}"/></td>
        % end
    </tr>
  </table>
  <input type="submit" value="Submit"/>
</form>
```

The implementation is a rather straightforward application of a template that provides a HTML form. The only interesting thing is that we can reuse the template th.tpl from above for the table header. This not only saves effort, but also makes the user experience consistent over the various parts of the application.

The Books Application Routes: Adding Book Records

▷ The result is



▷ The action in the HTML form is to POST to the path /add. Thus we need POST route for /add as well:

```
@post('/add')
def addResponse():
    data = parseResponse()
    ins = '''INSERT INTO Books VALUES
             (:Last,:First,:YOB,:YOD,:Title,:YOP,:Publisher,:City)'''
    cursor.execute(ins,data)
    return template('response', data = data, cols=cols,
                                  rowid = cursor.lastrowid,
                                  text = 'New book record received')
```

Note the use of sqlite3 parameter substitution in addResponse!

The addResponse function that answers the POST route for the path /add/ just inserts a new database record in to the Books table. Note the use of the SQLite3 parameter substitution here. We substitute the parameters :⟪key⟫ in the string ins with the corresponding values in the Python dictionary data which we obtain as the result of the parseResponse function, which we will look at next.

The Books Application Routes: Adding Book Records

▷ This uses the function parseResponse, which we will reuse later.

```
def parseResponse ():
    data = {'Last': request.forms.get('Last'),
            'First': request.forms.get('First'),
            'YOB': request.forms.get('YOB'),
            'YOD': request.forms.get('YOD'),
            'Title': request.forms.get('Title'),
            'YOP': request.forms.get('YOP'),
             'Publisher': request.forms.get('Publisher'),
            'City': request.forms.get('City')}
    return data
```

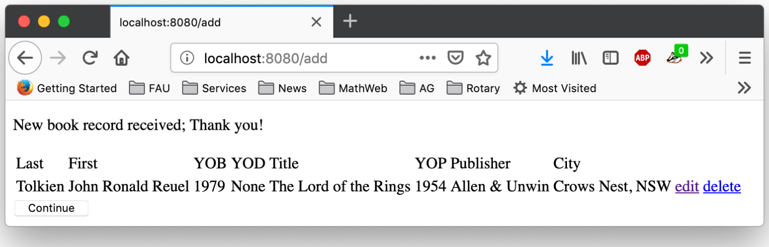
▷ and the template repsonse.tpl:

```
<form action='/'>
  <p>{{text}}; Thank you!</p>
  <table>
    % include('th.tpl',cols=cols)
    % include('book.tpl',**data,cols=cols)
  </table>
  <input type="submit" value="Continue"/>
</form>
```

The parseResponse function is almost trivial, it just queries the response object that comes from the POST request for the various components via the forms.get method and packages the results in a Python dictionary that feeds the response.tpl template. The latter creates a HTML form without text input fields we only use it to trigger a GET request to the path / (the application root that displays the updated book list). Note that we re-use the templates th.tpl and books.tpl from above again.

The Books Application Routes: Adding Book Records

▷ Here is the result after filling in Tolkien's "*Lord of the Rings*":


localhost:8080/add
localhost:8080/add
Getting Started FAU Services News MathWeb AG Rotary Most Visited

New book record received; Thank you!

Last	First	YOB	YOD	Title	YOP	Publisher	City		
Tolkien	John Ronald Reuel	1979	None	The Lord of the Rings	1954	Allen & Unwin	Crows Nest, NSW	edit	delete

Continue

The next relevant rout is the "delete a book" functionality. Here we use another new feature: when creating a database table in SQLite3, the system creates an additional primary key column rowid. In particular we have a rowid column in the Books table, which we make use of.

The Books Application Routes: Deleting Book Records

▷ We add a route for deleting book records (for the delete button)

```
@get('/delete/<id:int>')
def delete(id):
    cursor.execute('DELETE FROM Books WHERE rowid = ?',(id,))
    return template('delete')
```

Note that we have a dynamic route here: We use the named wildcard <id:int> to obtain the rowid of the record to be deleted.

▷ The template file delete.tpl does the obvious:

```
<form action='/'>
  <p>Book record deleted; Thank you!</p>
  <input type="submit" value="Continue"/>
</form>
```

Note that the link on the "delete" buttons in the books table root (see template book.tpl above) has the form <button href="/edit/{{rowid}}">edit</button>, i.e. it references the rowid column. This is picked up in the GET route for /delete/<id:int> path via the named wildcard <id:int>. This makes sure the right database record is deleted.
The routes for editing book records combine techniques from the ones for adding and deleting. From the former we use the layout into a GET and POST route, from the latter, we use the dynamic route

The Books Application Routes: Editing Book Records

▷ **Idea:** Combine techniques from the add and delete routes

```
@get('/edit/<id:int>')
def edit(id):
    cursor.execute('SELECT * FROM Books WHERE rowid = ?',(id,))
    return template('edit',cursor.fetchone(),id = id,cols=cols)

@post('/edit/<id:int>')
def editResponse(id):
    data = parseResponse()
    up = """UPDATE Books
            SET Last = :Last, First = :First, YOB = :YOB, YOD = :YOD,
                Title = :Title, YOP = :YOP, Publisher = :Publisher,
                City = :City
            WHERE rowid = :rowid"""
    data.update({'rowid': id})
    cursor.execute(up,data)
    return template('response',data=data,text='Updated book record',cols=cols)
```

In this case we have a small subtlety: the update instruction and the template edit.tpl need a rowid key/value pair. We solve this by updating the data dictionary suitably. Now we only have to give the template edit.tpl, which is rather straightforward.

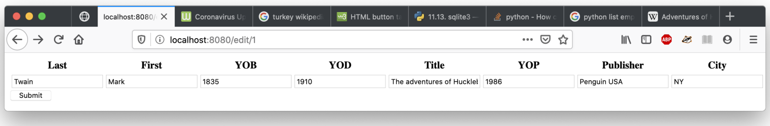
Books Application Routes: Editing Book Records (cont.)

▷ The template file edit.tpl is similar to add.tpl above, but pre-fills the input fields with the database record values.

```
<form action="/edit/{{id}}" method="post">
  <table>
    % include('th.tpl', cols=cols)
    <tr>
      <td><input type="text" name="Last" value="{{Last}}"/></td>
      <td><input type="text" name="First" value="{{First}}"/></td>
      <td><input type="text" name="YOB" value="{{YOB}}"/></td>
      <td><input type="text" name="YOD" value="{{YOD}}"/></td>
      <td><input type="text" name="Title" value="{{Title}}"/></td>
      <td><input type="text" name="YOP" value="{{YOP}}"/></td>
      <td><input type="text" name="Publisher" value="{{Publisher}}"/></td>
      <td><input type="text" name="City" value="{{City}}"/></td>
      <td><input type="submit" value="Submit"/></td>
    </tr>
  </table>
</form>
```

Books Application Routes: Editing Book Records (cont.)

▷ The result is



Last	First	YOB	YOD	Title	YOP	Publisher	City
Twain	Mark	1835	1910	The adventures of Hucklel	1986	Penguin USA	NY

Submit

▷ Again, we use the template response.tpl, which we fill with a different message.

The main message to take home from this experiment is that we can build a simple but complete web application with less than 100 lines of Python code and less than 70 lines of HTML template files.

10.2 Asynchronous Loading in Modern Web Apps

The web applications we have seen up to now have been relatively conventional, based mostly on server-side scripting together with some client-side computation via JavaScript. This is a powerful setup with one problem. Whenever the user needs new data from the server, the browser has to request a new web page – even if only a small fragment of the original page needs to be changed.

The solution to this problem is to use JavaScript itself to load the new information and directly integrate the result into the DOM, using a technology called Ajax. In this section we will introduce Ajax by extending the database from ?? with a lightweight front-end web application.

Before we get into the example, we introduce Ajax as a technology itself and recap the idea of client-side computation using the DOM. The code in this section will be considerably more complex than what we have seen before. But it shows many of the characteristical ideas of modern web application development in a nutshell. That should make it worthwile to study, even if that may take more than one attempt.

AJAX for more responsive Web Pages

▷ **Definition 10.2.1.** Ajax, (also AJAX; short for "Asynchronous JavaScript and XML") is a set of client side techniques for creating asynchronous web applications.

▷ **Definition 10.2.2.** A process p is called asynchronous, iff the parent process (i.e. the one that spawned p) continues processing without waiting for p to terminate.

▷ **Intuition:** With Ajax, web applications can send and retrieve data from a server without interfering with the display and behaviour of the existing page.

▷ **Application:** By decoupling the data interchange layer from the presentation layer, Ajax allows web pages and, by extension, web applications, to change content dynamically without the need to reload the entire page.

▷ **Observation:** Almost all modern web application extensively utilize Ajax.

▷ **Note:** In practice, modern implementations commonly use JSON instead of XML.

Recall the HTML rendering pipeline in browsers around the DOM we introduced for client-side computation.

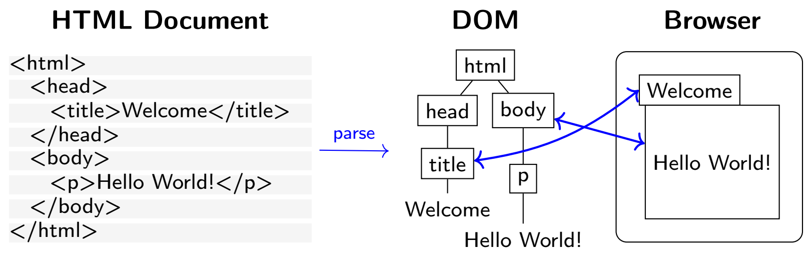
Background: Rendering Pipeline in browsers

▷ **Observation:** The nested markup codes turn HTML documents into trees.

▷ **Definition 10.2.3.** The document object model (DOM) is a data structure for the HTML document tree together with a standardized set of access methods.

▷ **Rendering Pipeline:** Rendering a web page proceeds in three steps

1. the browser receives a HTML document,
2. parses it into an internal data structure, the DOM,
3. which is then painted to the screen. (repaint whenever DOM changes)



```
<html>
  <head>
    <title>Welcome</title>
  </head>
  <body>
    <p>Hello World!</p>
  </body>
</html>
```

The DOM is notified of any user events. (resizing, clicks, hover,. . .)

The most important concept to grasp here is the tight synchronization between the DOM and the screen. The DOM is first established by parsing (i.e. interpreting) the input, and is synchronized with the browser UI and document viewport. As the DOM is persistent and synchronized, any change in the DOM is directly mirrored in the browser viewpoint, as a consequence we only need to change the DOM to change its presentation in the browser. This exactly is the purpose of the client side scripting language, which we will go into next.

We will put the abstract ideas about Ajax and JSON introduced above to practical use. This will make our understanding much more concrete.
The first step in the development of a Ajax based front end for the books application – as in any software project – is to specify the intended behaviour of the front-end and plan the implementations.

Example: Details on Request via AJAX

▷ **Idea:** Use Ajax in a web application for the books application

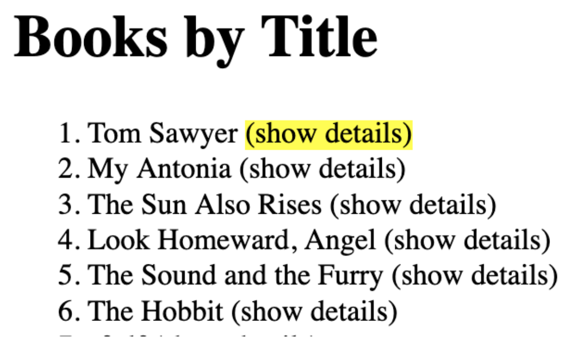
- ▷ The start page just has a list of book titles, and
- ▷ details are fetched by an Ajax request and presented in line.

▷ **Planning the Program:** We need a bottle server with

1. a dynamic route that returns JSON-encoded data for a given book,
2. a route for the main page that lists the book titles,
3. stpl template files for list items with an Ajax request, and
4. a JavaScript function that reads the JSON and inserts it into the DOM.

Here we see a mockup of what the result will look like:

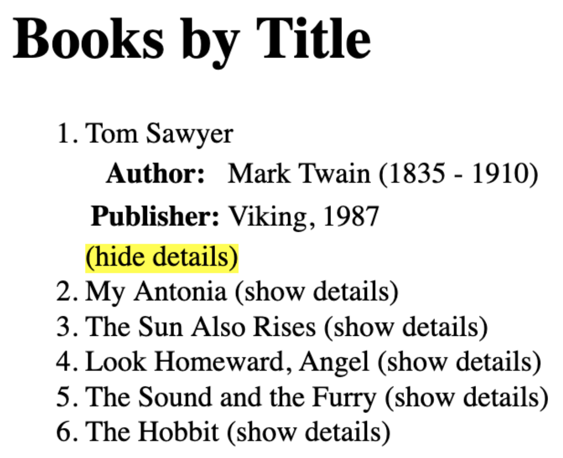
The finished product (initial state)



Books by Title

1. Tom Sawyer (show details)
2. My Antonia (show details)
3. The Sun Also Rises (show details)
4. Look Homeward, Angel (show details)
5. The Sound and the Furry (show details)
6. The Hobbit (show details)

The finished product (with details loaded)



Books by Title

1. Tom Sawyer

   Author: Mark Twain (1835 - 1910)

   Publisher: Viking, 1987

   (hide details)
2. My Antonia (show details)
3. The Sun Also Rises (show details)
4. Look Homeward, Angel (show details)
5. The Sound and the Furry (show details)
6. The Hobbit (show details)

Now we are ready to begin with the implementation. Fortunately, the first step – serving the main page and the JSON data for a given book is very simple, indeed that is exactly what bottle was created for, since it is a routine task for building modern web applications.

The Routes (Serving HTML and JSON)

▷ After setting up the database and co, we have a standard route:

```
@route('/')
def books():
    cursor.execute('SELECT rowid, Title, YoP FROM Books')
    rv = cursor.fetchall()
    return template('titles', books=rv)
```

▷ JSON routes and APIs are very easy in bottle: we just return a dictionary.

```
@route('/json/<id:int>')
def book(id):
    cursor.execute(f'SELECT * FROM Books WHERE rowid={id}')
    row = cursor.fetchone() # Only one result, rowid is a primary key.
    return dict(zip(row.keys(), row)) # Pair up column names with values.
```

▷ **Dictionaries and JSON in Bottle:** Bottle automatically transforms Python dictionaries into JSON strings; sets the Content Type header to application/json.

The Basic Templates

▷ The template titles.tpl is also standard

```
<html>
% include('bookshead.tpl')
<body>
  <h1>Books by Title</h1>
  <ol>
    % for bk in books: include('title.tpl',Id=bk[0], title=bk[1]) end
  <ol>
</body>
</html>
```

▷ The template title.tpl presents a single book title

```
<li>
  <span class="booktitle">{{title}}</span>
  <span id="content{{Id}}"></span>
  <span class="interact" id="interact{{Id}}"
              onclick="load_details({{Id}})">(show details)</span>
</li>
```

The empty span will be filled by an Ajax call later!

▷ The interesting things happen in bookshead.tpl (up next)

But now it becomes more tricky. We set up a couple of scripts in head of bookshead.tpl, which we will now take a more detailed look at.

The Script load_details

▷ bookshead.tpl starts supplying jQuery and a jQuery templating library:

```
<script type="application/javascript"
    src="http://ajax.googleapis.com/ajax/libs/jquery/3.6.0/jquery.min.js"></script>
<script type="application/javascript"
    src="https://cdn.jsdelivr.net/gh/codepb/jquery−template@1.5.10/dist/jquery.loadTemplate.min.js"></script>
```

▷ The main contribution of bookshead.tpl is the jQuery function load_details

```
async function load_details (numb) {
  /∗ Request Info via JSON, feed it to template, update "show details" span ∗/
  await $.getJSON("/json/" + numb,
      function (data) {$("#content" + numb).loadTemplate($("#open"), data)});
```

which uses the jQuery Ajax call $.getJSON. This takes two arguments:

1. the URL for the HTTP GET request
2. a JavaScript function that is called if the GET request was successful.

The function (in argument 2) is then used to extend the result of `$("#content"+numb)`, i.e. that element in the DOM whose `id` attribute is content$i$ where i is the value of the `numb` variable.

The Script load_details Continued

▷ We also use jQuery to change the onlick behaviour of the span element (from load_details to toggle_details, explained below) and the text contained therein.

```
  interact = $("#interact" + numb)

  /* change click behaviour of interaction span from show to toggle */
  interact.removeAttr('onclick');
  interact.attr('onClick', 'toggle_details(' + numb + ');');

  /* also change included text appropriately */
  interact.html("(hide details)");
}
```

▷ Recall the structure of title.tpl: For every book we have a title, a content element that starts out empty and gets filled when load_details is called, and a clickable interaction element that triggers load_details.

```
<li>
  <span class="booktitle">{{title}}</span>
  <span id="content{{Id}}"></span>
  <span class="interact" id="interact{{Id}}"
             onclick="load_details({{Id}})">(show details)</span>
</li>
```

▷ The toggle_details-function used above does nothing but setting the content element to hidden or visible and changing the text of the interaction element.

```
function toggle_details (numb) {
  /* hide or show appropriate content element */

  content = $("#content" + numb);
  interact = $("#interact" + numb);

  if(content.css('display') == 'none') {
    content.show();
    interact.html("(hide details)");
  } else {
    content.hide();
    interact.html("(show details)");
  }
}
```

Now let us look at this process in more detail. Apart from the fact that we are using jQuery template processing and the syntax is different, this works exactly like bottle template processing, which we have extensively practiced above. So just buckle up and enjoy the ride.

jQuery Template Processing

▷ **Recall:** We are still trying to understand
$("#content" + numb).loadTemplate($("#open''),data)
It extends the empty <span id=''content$i$> in title.tpl with a details table:

▷ The loadTemplate method takes two arguments

1. a template; here the result of $(#open),i.e. the element in bookshead.tpl whose id attribute is open (note the type attribute that makes it HTML)

```
<script type="text/html" id="open">
  <table>
     <tr>
       <th>Author:</th>
       <td>
         <span data−content="First"></span> <span data−content="Last"></span>
          (<span data−content="YOB">−</span> − <span data−content="YOD"></span>)
       </td>
     </tr>
     <tr>
        <th>Publisher:</th>
        <td><span data−content="Publisher"></span>, <span data−content="YOP"></span></td>
     </tr>
  </table>
</script>
```

2. a JavaScript data object: here the argument of the success function: the JSON record provided by the server under route /json/i

```
{"Last": 'Twain',
 "First": 'Mark',
 "YoB": 1835,
  "YoD": 1910,
  "Title": 'Huckleberry Finn',
  "YoP": 1986,
  "Publisher": 'Penguin USA',
  "City": 'NY'}
```

▷ The jQuery template processing places the value of the data−content attribute into the <span>. The resulting table constitutes the generated "detail view":

```
<table>
  <tr>
    <th>Author:</th>
    <td>
      <span>Mark</span> <span>Twain</span>
      (<span>1835</span>−<span>1910</span>)
    </td>
  </tr>
  <tr>
    <th>Publisher:</th>
    <td><span>Penguin USA</span>, <span>NY</span></td>
  </tr>
</table>
```

▷ **Note:** Both the JavaScript object in step 2. as well as the result of the template processing show afterwards are virtual objects that exist only in memory. In particular, we do not have to write them explicitly.

Now, we will show you the code in its entirety, it is less than 100 lines. So with the right tools, a modern web page| with Ajax is not that difficult (once you wrap your head around it).

Code: An AJAX-based Frontend for the Books App

▷ booksapp—ajax.py: the web server with two routes

```
import sqlite3
from bottle import route, run, template, static_file

# Connect to database
db = sqlite3.connect("./books.db")
# Row factory so we can have column names as keys.
db.row_factory = sqlite3.Row
cursor = db.cursor()

@route('/')
def books():
    cursor.execute('SELECT rowid, Title, YoP FROM Books')
    rv = cursor.fetchall()
    return template('titles', books=rv)

# JSON interfaces are very easy in bottle, just return a dictionary
@route('/json/<id:int>')
def book(id):
    cursor.execute(f'SELECT * FROM Books WHERE rowid={id}')
    row = cursor.fetchone() # Only one result, rowid is a primary key.
    return dict(zip(row.keys(), row)) # Pair up column names with values.

run(host='0.0.0.0', port=32500, debug=True)
# Close database
db.close()
```

▷ titles.tpl styles the list of book titles

```
<html>
% include('bookshead.tpl')
<body>
  <h1>Books by Title</h1>
  <ol>
    % for bk in books: include('title.tpl',Id=bk[0], title=bk[1]) end
  <ol>
</body>
</html>
```

▷ title.tpl styles a single book

```
<li>
  <span class="booktitle">{{title}}</span>
  <span id="content{{Id}}"></span>
  <span class="interact" id="interact{{Id}}"
             onclick="load_details({{Id}})">(show details)</span>
</li>
```

▷ bookshead.tpl provides the whole head of the main page.

```
<head>
  <title>Books with Ajax Details</title>
  <meta charset="utf-8">
  <style>.interact:hover { background-color: yellow; }</style>

<script type="application/javascript"
    src="http://ajax.googleapis.com/ajax/libs/jquery/3.6.0/jquery.min.js"></script>
<script type="application/javascript"
    src="https://cdn.jsdelivr.net/gh/codepb/jquery-template@1.5.10/dist/jquery.loadTemplate.min.js"></script>

<script type="text/html" id="open">
  <table>
```

```
      <tr>
        <th>Author:</th>
        <td>
          <span data-content="First"></span> <span data-content="Last"></span>
           (<span data-content="YOB">-</span> - <span data-content="YOD"></span>)
        </td>
      </tr>
      <tr>
         <th>Publisher:</th>
         <td><span data-content="Publisher"></span>, <span data-content="YOP"></span></td>
     </tr>
  </table>
</script>

<script type="text/javascript">
/* async because we're waiting for the template magic to finish before appending */
async function load_details (numb) {
  /* Request Info via JSON, feed it to template, update "show details" span */
  await $.getJSON("/json/" + numb,
      function (data) {$("#content" + numb).loadTemplate($("#open"), data)});

  interact = $("#interact" + numb)

  /* change click behaviour of interaction span from show to toggle */
  interact.removeAttr('onclick');
  interact.attr('onClick', 'toggle_details(' + numb + ');');

  /* also change included text appropriately */
  interact.html("(hide details)");
}

function toggle_details (numb) {
  /* hide or show appropriate content element */

  content = $("#content" + numb);
  interact = $("#interact" + numb);

  if(content.css('display') == 'none') {
    content.show();
    interact.html("(hide details)");
  } else {
    content.hide();
    interact.html("(show details)");
  }
}
</script>
</head>
```

10.3 Deploying the Books Application as a Program

Now we address the fact that a web appplication is usually deployed on a unix server, by sysadmins who are accustomed the unix way of handling – configuring, starting, etc. – applications. We will first introduce a way to make Python scripts as shell commands and give them arguments optional and mandatory ones.

Deploying The Books Application as a Program

▷ **Note:** Having a Python script booksapp.py you start with python3 booksapp.py is sufficient for development.

▷ If you want to deploy it on a web server, you want more: The sysadmin you deliver your web application to wants to start and manage it like any other UNIX command.

▷ **After all,** your web server will most likely be a UNIX (e.g. linux) computer.

▷ In particular behavioural variants should be available via command line options, i.e. strings starting with – after the command.

▷ **Example 10.3.1.** To run the books application without output (−q or −−quiet) and initialized with the seven book records we want to run

```
booksapp −q −−initbooks
```

Deploying The Books Application as a Program

▷ **Example 10.3.2.** If we forget the options, we need help:

```
> booksapp −−help
Usage: <yourscript> [options]

Options:
  −h, −−help show this help message and exit
  −q, −−quiet don't print status messages to stdout
  −l FILE, −−log=FILE write log reports to FILE
  −−initbooks initialize with seven book records
```

▷ **Definition 10.3.3.** The command line option −−help or −h is traditionally used for the help option.

Deploying a Python Script as a Shell Command/Executable

▷ We can make our a Python script behave like a native shell command.

▷ The file extension .py is only used by convention, we can leave it out and simply call the file booksapp.

▷ Then we can add a special Python comments in the first line

```
#!/usr/bin/python3
```

which the shell interprets as "call the program python3 on me".

▷ Finally, we make the file hello executable, i.e. tell the shell the file should behave like a shell command by issuing

```
chmod u+x booksapp
```

in the directory where the file booksapp is stored.

▷ We add the line

```
export PATH="./:${PATH}"
```

to the file .bashrc. This tells the shell where to look for programs (here the respective current directory called .)

Working with Options in Python

▷ We have the optparse library for dealing with command line options (install with pip3)

▷ **Example 10.3.4 (Options in the Books Application).**

```
from optparse import OptionParser
parser = OptionParser()
parser.add_option("-l", "--log", dest="logfile",
                  help="write logs to FILE", metavar="FILE")
parser.add_option("-q", "--quiet",
                  action="store_false", dest="verbose", default=True,
                  help="don't print status messages to stdout")
parser.add_option('--version',dest="version",default=1.0,type="float",
                  help="the version of the books application")

options, args = parser.parse_args()
# do something with the options and their args.
print ('VERSION :', options.version)
```

Chapter 11

Image Processing

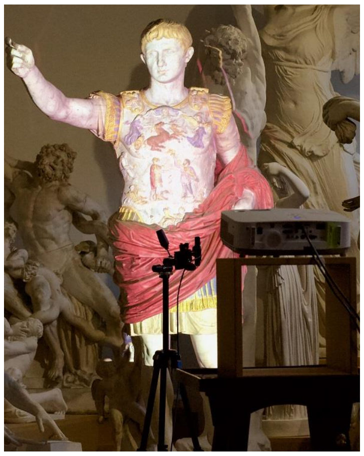
We will now begin a new topic on our way to a useful image database. In particular we will see how computer scientists think about images, how digital images are represented in computer memory and what we can do with them.

11.1 Basics of Image Processing

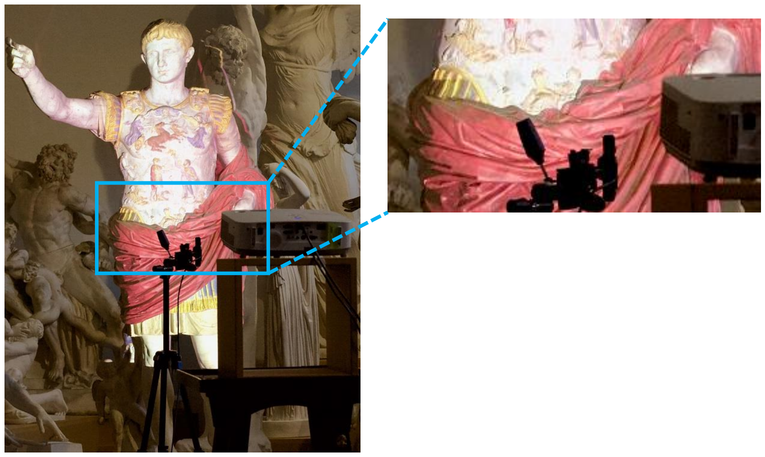
11.1.1 Image Representations

Images

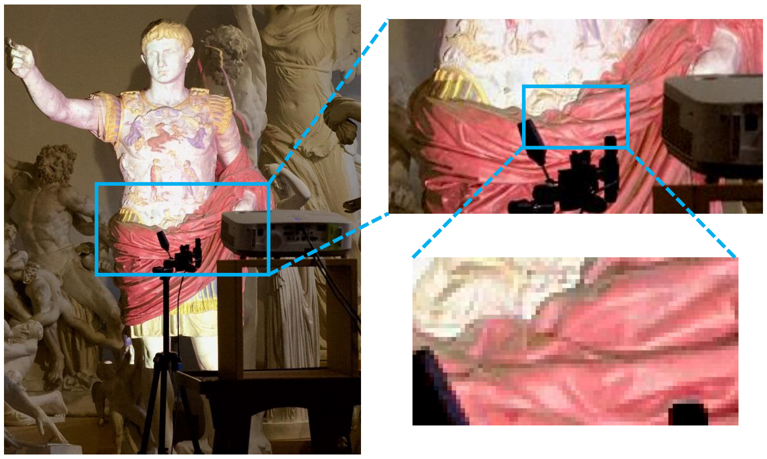
▷ **Example 11.1.1 (Zooming in on Augustus).** A digital image taken by a standard DSLR camera. Let's zoom in on it!



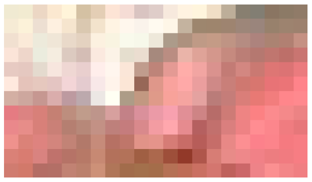
And a bit more



When zooming in on an image, we start to see blocks of colors, which are organized in a regular grid.



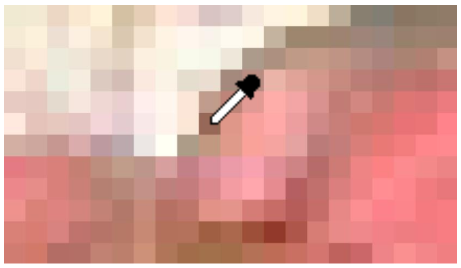
Images as Rasters of Pixels

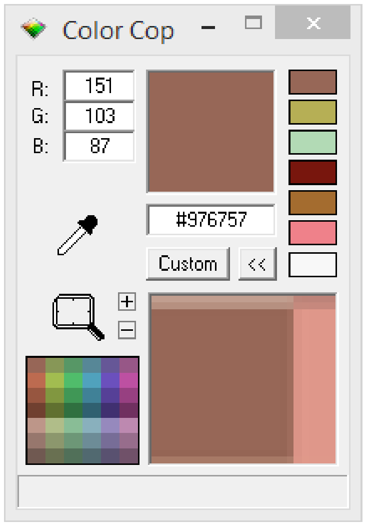


- ▷ If we zoom in quite a bit more, we see
- ▷ **Observation:** The colors are arranged in a two- dimensional grid (raster).

▷ **Definition 11.1.2.** We call the grid raster and each entry in it pixel (from "picture element").

Colors





▷ **Definition 11.1.3.** Colors are usually represented in RGB format, i.e. as triples $\langle R, G, B\rangle$ with three channels (also called bands).

▷ $R, G, B \in [0{,}255] \leadsto$ One Byte per channel per pixel.

▷ Images in this format can store $256 \cdot 256 = 256^3$ (about 16 million) colors.

Each pixel stores color information. We can obtain the values stored in digital images using a color picker. Image processing programs like Microsoft Paint or Adobe Photoshop provide color pickers (pipettes), but there also exist standalone applications. In this example we are using Color Cop [1].

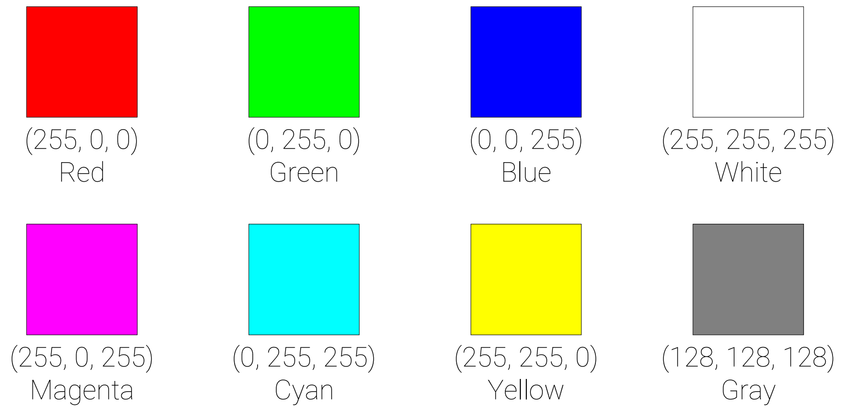
According to the color picker, our pixel stores the value (151, 103, 87). Colors are organized in the so-called RGB format, meaning a color is composed from a mixture of red (R), green (G) and blue (B). We call these components channels or bands.

The value in each of these channels typically ranges from 0 to 255. This is because a single Byte can store exactly this value range and a Byte was deemed enough for most applications. We can deduce that a pixel has $256 \times 256 \times 256$ distinct value combinations, which is just over 16 million colors an image in this format can display. You might have seen this number on product descriptions of computer monitors or cameras.

Color Examples

▷ **Example 11.1.4.** A color can be represented by three numbers.

[1] http://colorcop.net/



▷ **Definition 11.1.5.** A color is called grayscale, iff $R = G = B$

A channel value of 0 means no intensity in this channel, a value of 255 corresponds to full intensity. Thus, in order to create a pure red we set the R channel to 255 and the other two to 0 (no green or blue). Other colors are achieved in a similar fashion.

Secondary colors (e.g. magenta, cyan, yellow) are created by mixtures of red, green, and blue. For example, we create magenta by mixing red and blue.

Different shades of gray are obtained, when R=G=B. White is the brightest gray we can achieve, by setting all values to 255. Black on the other hand has all channels set to 0 (meaning no light/intensity).

When processing colors it is often beneficial to think about normalized colors. We normalize colors by dividing by 255 (the highest value). Resulting color values are now between 0 and 1.

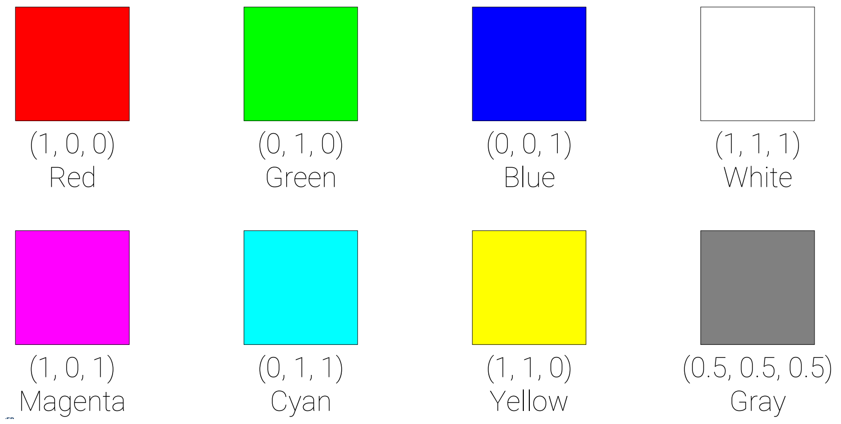
Normalized Color Values

▷ **Observation 11.1.6.** *For color representations, only the relative contribution of the band is imporant.*

▷ **Definition 11.1.7.** Normalized colors use pixel values between 0 and 1.

▷ **Idea:** Values are still stored as Bytes, but normalized before use: $v' = v/255$

▷ **Example 11.1.8.**



HTML Color Codes

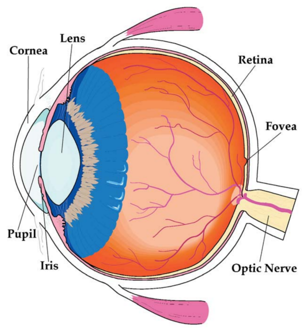
- HTML uses a shorthand notation for colors using hexadecimal numbers.
- **Example 11.1.9.**

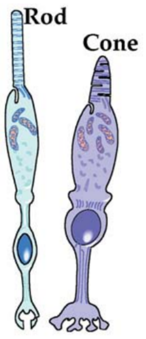
🟥	🟩	🟦	⬜
#FF0000 Red	#00FF00 Green	#0000FF Blue	#FFFFFF White
🟪	🟦	🟨	⬛
#FF00FF Magenta	#00FFFF Cyan	#FFFF00 Yellow	#808080 Gray

Recall from last semester: In HTML and CSS we often express colors in HTML color codes. This is the same principle as before, however the values are not expressed in decimal numbers but instead in hexadecimal.
Quick detour into the real world: Let's explore where the RGB format comes from.

The Human Eye

- **Definition 11.1.10 (The Human Eye).** Light from our surroundings enters our eye through the lens and then hits the retina on the back of our eye.





The retina has cones and rods, which are responsible for color and brightness vision, respectively.

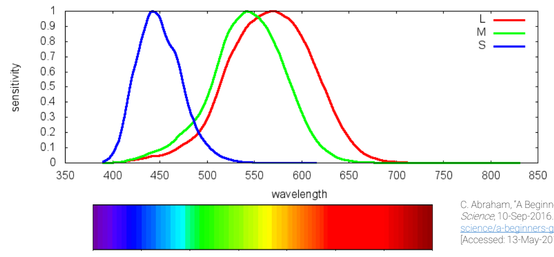
▷ Since we are interested in colors here, we will ignore the rods for the purpose of this course.

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 296 2025-02-06

Light is an electromagnetic radiation. Only a small part of this radiation is visible to the human visual system (wavelengths around 380 to 740 nanometers).

The Human Eye – Three Types of Cones

▷ **Sensitivity of the Three Cones:**



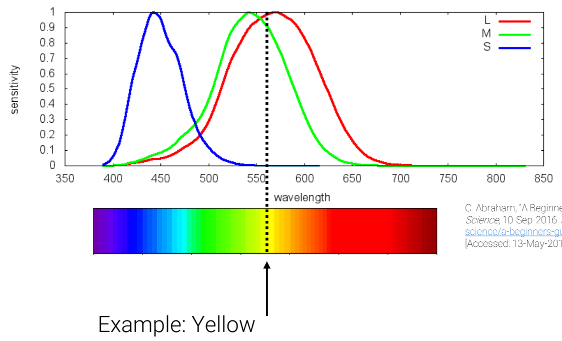
C. Abraham, "A Beginner's Guide to (CIE) Colorimetry," *Hipster Color Science*, 10-Sep-2016. Available: https://medium.com/hipster-color-science/a-beginners-guide-to-colorimetry-401f1830b65a. [Accessed: 13-May-2019].

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 297 2025-02-06

There are three types of cones, which react to different areas in this spectrum. They roughly correspond to the wavelengths, which we perceive as red, green, and blue (or rather long, middle, and short wavelengths).

The Human Eye – Three Types of Cones

▷ **Example 11.1.11 (We see Yellow).**



C. Abraham, "A Beginner's Guide to (CIE) Colorimetry," *Hipster Color Science*, 10-Sep-2016. Available: https://medium.com/hipster-color-science/a-beginners-guide-to-colorimetry-401f1830b65a. [Accessed: 13-May-2019].

Example: Yellow
Both "red" and "green" cone are stimulated.

▷ **Observation 11.1.12.** *We can create all (human-visible) colors as a mixture of red, green, and blue light.*

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 298 2025-02-06

When we now see yellow light for example, the two cones responsible for long and medium length wavelengths are stimulated. Our brain converts this stimulus to yellow.

However, let's imagine we perceive a mixture from red and green light. In this case these two

cones will be stimulated, too! Our brain is incapable of distinguishing between these two scenarios, since the physical stimulus on our eye is the exact same!
Monitors take advantage of this, since they usually also have pixels.

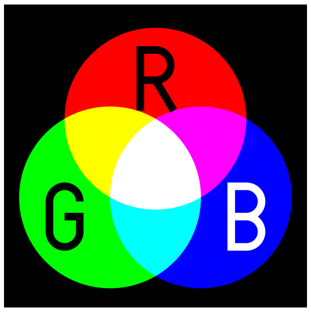
Monitors

▷ **Definition 11.1.13.** A computer monitor (or just monitor)is an output device for visual information.

▷ Monitors (usually) have pixels, too!

▷ **Definition 11.1.14.** In color monitors, pixels typically consist not of a single light source, but three distinct subpixels.

▷ If these subpixels are small enough and close together, our eye cannot see that the light actually comes from different points and thus perceives the mixture color.



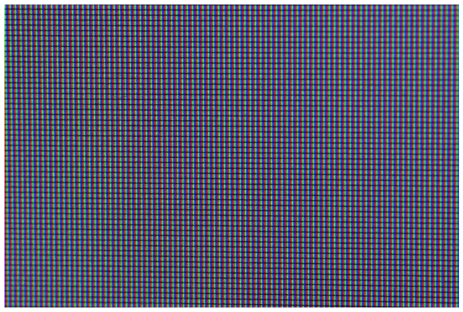
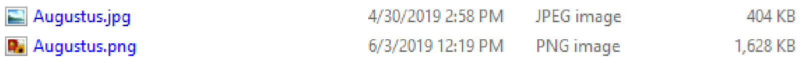


Image Size

▷ **Example 11.1.15 (Augustus again).**

Image: 1440×746 pixels
Expected file size:
Width · Height · *Channels*
$1440 \cdot 746 \cdot 3 = 3,222,720\text{B} \approx 3\text{MiB}$

▷ But if we look onto our disk we see somthing completely different:



Augustus.jpg	4/30/2019 2:58 PM	JPEG image	404 KB
Augustus.png	6/3/2019 12:19 PM	PNG image	1,628 KB

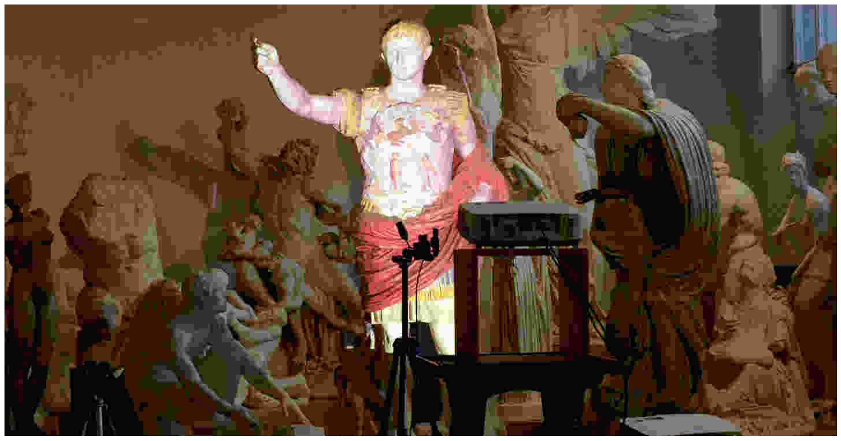
▷ On disk, images are usually compressed (JPEG, PNG, GIF,WebP etc). JPEG file size is smaller than PNG, but image quality is lost.

This is because images on disc are usually compressed and stored in an image file format like JPEG or PNG. Be careful with JPEG compression! JPEG sacrifices image quality in order to achieve smaller file sizes!

JPEG Compression Artefacts

▷ **Example 11.1.16 (Augustus again).** Here, the Augustus image is saved with a very high jpeg compression. The file size is tiny (27 KB, compare to 440 KB on previous slide). However, the image quality suffers.

JPEG creates blocks of pixels, and approximates the colors in this block with as few bits as possible (according to compression ratio).



AugustusCompressed.jpg 6/7/2019 9:11 AM JPEG image 27 KB

In this example we turned the JPEG compression very high, which leads to a tiny file size but strong artefacts in the image quality.

11.1.2 Basic Image Processing in Python

When processing digital images programatically, we have to load them from disc and then perform operations on them. In IWGS we will use Pillow library for this task. The example shows howimages are loaded from disc.

The Pillow Library for Image Processing in Python

▷ We will use the Pillow library in IWGS.

▷ **Definition 11.1.17.** Pillow is a fork (a version) of the old Python library PIL (Python Image Library). (hence the name)

▷ Details at `https://pillow.readthedocs.io/slides/stable/`

▷ **Install:** pip install Pillow

▷ **Example 11.1.18.** Determine the color of a particular pixel

```
from PIL import Image
# load image
im = Image.open('image.jpg')
```

```
im.show()
# access color at pixel (x, y)
x = 15
y = 300
r, g, b = im.getpixel((x, y))
```

▷ **Example 11.1.19.** Directly use the image object in jupyter notebooks:

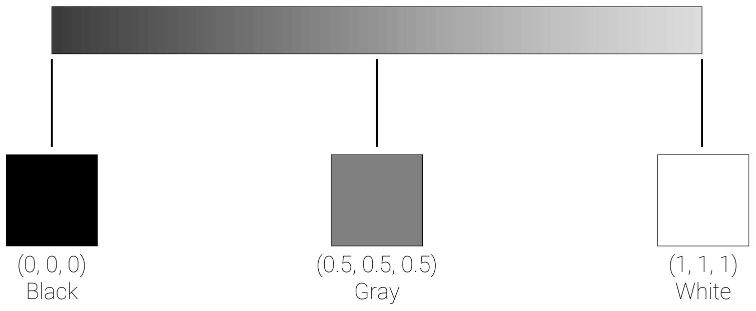
```
from PIL import Image
# load image
im = Image.open('image.jpg')
im # in Jupyter Notebooks, we can directly use the variable
```

The notebooks shows the image in a new cell.

Loading here means that the file is read, and that the compression is reversed, i.e. the digital image is decompressed. This means that the image which was before stored in JPEG compression is now present in memory. You can think about the loaded image as a long Python list of pixel values, i.e. one pixel after the other.

Grayscale Images

▷ **Recall:** A color is grayscale, iff R=G=B.



▷ **Idea:** If all channels have the same value, why store all three?

▷ Grayscale images usually have only one channel.

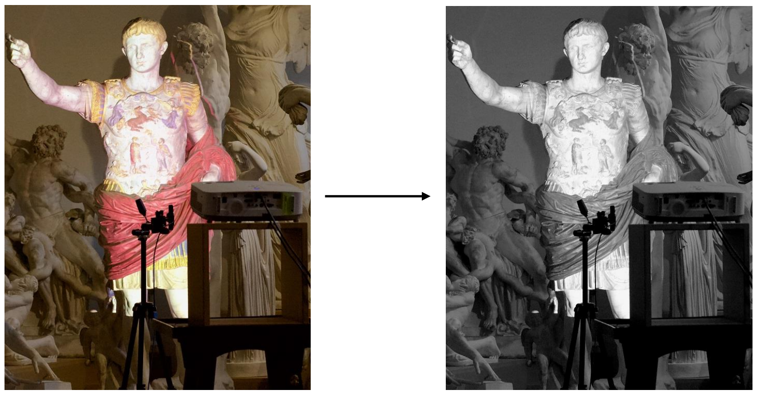
Since it is pointless to store each value three times, grayscale images usually only store one value per pixel, which is then tripled before display.
Conversion from color to grayscale images is a common operation, which most image processing tools (Photoshop etc.) support. It serves as a first example of what we can do with digital images.

Grayscale Conversion

▷ **Observation 11.1.20.** *Humans are very sensitive to green, less to red, and least to blue.*

▷ **Definition 11.1.21.** To convert an image to an grayscale image (grayscale conversion), we compute $Gray = 0.21R + 0.71G + 0.08B$

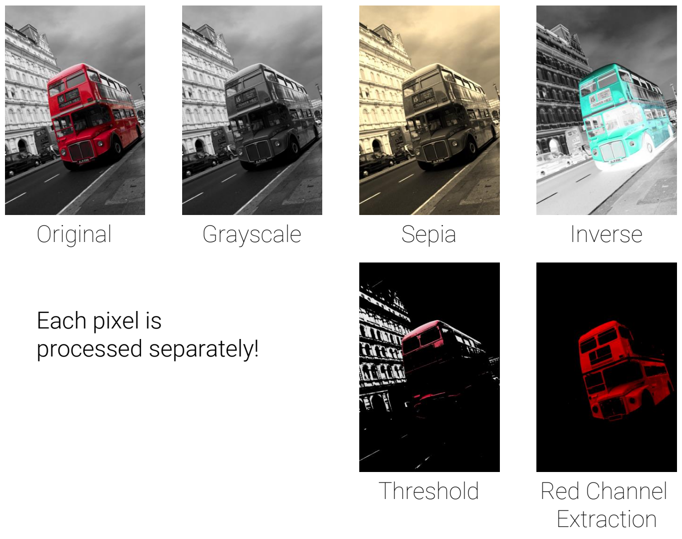
▷ **Example 11.1.22 (Grayscale Conversion).**



Grayscale conversion is a *weighted sum* of the three channel values. This means, each channel value is multiplied with a factor and then the values are summed up to form a single value. Since humans are very sensitive to green, the G channel has the highest weight.
We now show some more image operations.

More Image Operations

▷ **Example 11.1.23 (More Image Operations).**



▷ As for grayscale conversion of these process each pixel separately.

Implementation of these operations is very simple in Python. Since we store all our pixels in a large list in Pillow, we can simply create a for-loop over this list, do our calculation and store the result in a new image at the same pixel coordinate.

Image Operations in Pillow

- ▷ The pillow library supports many image operations out of the box.
- ▷ **Example 11.1.24 (Grayscale Conversion and Inversion in Pillow).**

```
from PIL import Image, ImageOps
im = Image.open ('image.jpg')
# convert to grayscale
gray = ImageOps.grayscale(im)
# invert image
inverse = ImageOps.invert(im)
```

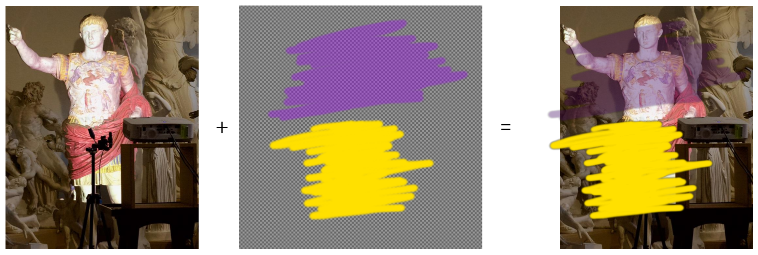
- ▷ Complete List: `https://pillow.readthedocs.io/en/stable/reference/ImageOps.html`

Transparency is an important operation. In this example we want to layer two digital images on top of each other. We thus need to store for each pixel a measure of how transparent it is.

We expand our RGB notion to RGBA, by introducing a fourth channel A. A stands for alpha and corresponds to the opacity of a pixel, i.e. a value of 0 means zero opacity (fully transparent), a value of 1 (normalized) means fully opaque (no transparency).

Transparency and Image Composition

- ▷ Sometimes we want to overlay images ⇝ layers.
- ▷ We need a notion of how transparent a pixel is.
- ▷ **Definition 11.1.25.** We introduce a fourth channel: A (for alpha). Alpha is the opacity (inverse of transparency). A pixel is now $\langle R, G, B, A\rangle$.
- ▷ **Example 11.1.26 (Combining Images).**

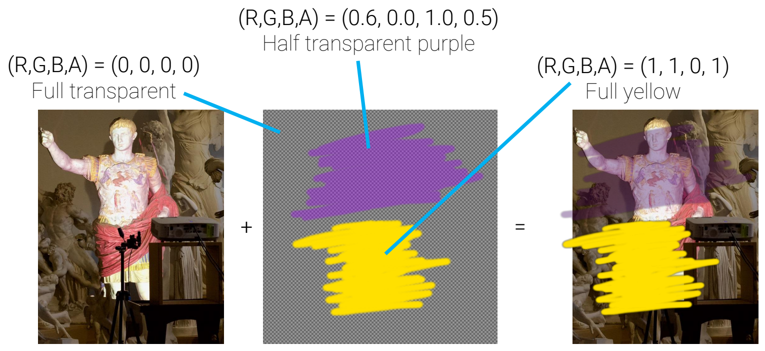


- ▷ **Note:** The order of layers is important here: The Augustus image is below the other image! The Augustus image has *no* transparency, the second image does!

See examples for the opacity here. Fully transparent regions (visualized by the checkerboard), have an alpha value of 0. Fully opaque regions have a value of 1. Intermediate values are possible which correspond to partial transparency.

Transparency (continued)

▷ **Example 11.1.27 (Combining Images).**



$R_{target} = (1\text{-}A) \times R_{augustus} + A \times R_{purple,yellow}$
$G_{target} = (1\text{-}A) \times G_{augustus} + A \times G_{purple,yellow}$
$B_{target} = (1\text{-}A) \times B_{augustus} + A \times B_{purple,yellow}$

The final image is then composed by deciding for each pixel how much color from each source image should contribute. Note that this is again a per-pixel operation, which can easily be implemented with a simple for-loop.

11.1.3 Edge Detection

We will now look at more interesting image operations. A typical example especially important for object recognition in digital images is to find features i.e. areas in the image, which are recognizable.

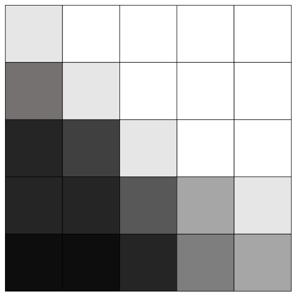
For example, let's say we want to find so-called edges in our image, i.e. areas where the color changes rapidly. Edges often correspond to object outlines. We will see an example later.

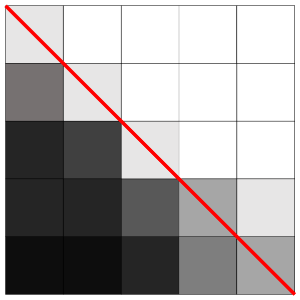
Edge Detection

▷ **Goal:** Find interesting parts of image (features).

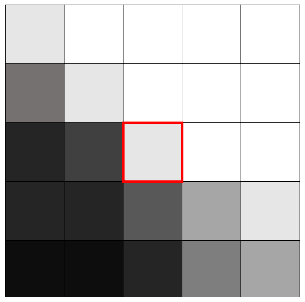
▷ **Definition 11.1.28.** Edge detection is the process of finding edges, i.e. image sections, where color changes rapidly.

▷ **Example 11.1.29 (Edge Detection).**

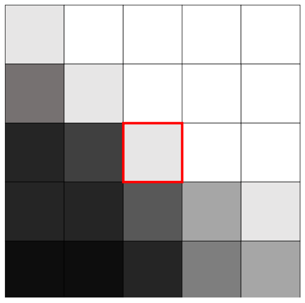




Clearly there is an edge in this image. How do we detect it automatically?



Decide for each pixel, whether it is on an edge. Here: Is marked pixel an edge pixel?



Inspect neighbor pixels.

▷ **Definition 11.1.30.** We call a pixel a horizontal edge pixel, iff

$$l_B - I_T + I_{BL} - I_{TL} + I_{BR} - I_{TR} > \tau$$

for some threshold τ and a vertical edge pixel, iff

$$l_R - I_L + I_{TR} - I_{TL} + I_{BR} - I_{BL} > \tau$$

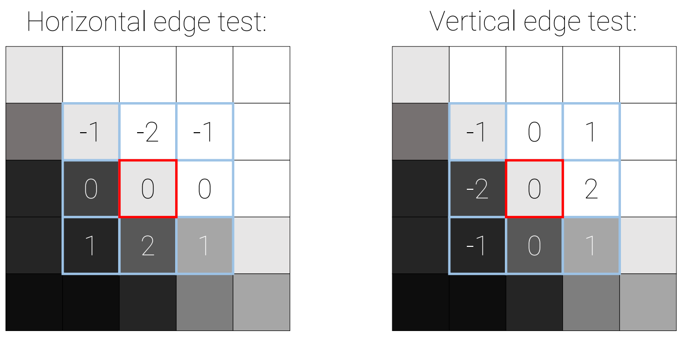
In this (admittedly simple) example image, we can clearly see, that there is an edge present, where the color shifts fast from dark to light. We will now explore, how we can detect such an edge automatically.
The idea is to decide for each pixel if it is part of an edge or not (binary decision, yes or no). Let's take the marked pixel as example, but rememb er that the following operations are performed on each pixel in the image.
The idea for this edge detection algorithm is to compare the pixel column left to our marked pixel to the column to the right. If the difference between the two columns is large, we know that we are observing a vertical edge.

Analogous we can do the same for horizontal edges, by comparing the row above to the row below our marked pixel.

We could perform this operation using only the pixels marked by L, R, B, and T, so only the direct neighbors. By taking the diagonal pixels into consideration, too, we make sure we only detect larger features.

Algorithm: Sobel Filter

- ▷ **Idea:** There is a general algorithm that computes this.
- ▷ **Definition 11.1.31.** Given a $3{\times}3$ matrix M, the Sobel filter computes a new pixel value by getting the pixel value of each neighbor in 3x3 window, multiply with the components in M and adding everything up.
- ▷ **Observation 11.1.32.** *Given a suitable matrix M, the Sobel filter computes the quantities from ??.*
- ▷ **Example 11.1.33 (Edge Tests via Sobel Filters).**



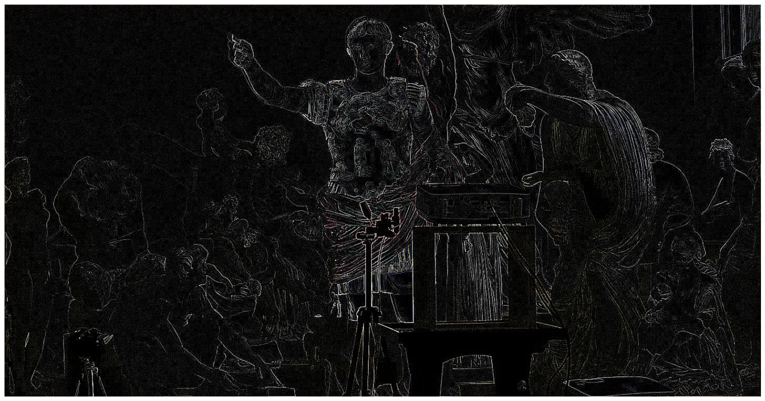
The operation we described here is called Sobel filter, named after Irwin Sobel.

Usually the direct neighbors are deemed more important than the diagonal neighbors. The pixel values of the neighbor pixels are thus weighted, such that the direct neighbors contribute more.
Here we see an example of edge detection. White pixels in the right image are pixels, which were classified as edge pixels, i.e. pixels where large changes in color are present. Black pixels are no

edges.

Edge-Detection in Pillow

▷ **Example 11.1.34 (Augustus and his Edges).**



▷ **Example 11.1.35 (Edge Detection in Pillow).**

```
from PIL import Image, ImageFilter
im = Image.open('augustus.jpg')
edges = im.filter(ImageFilter.FIND_EDGES)
edges.show() # or just edges in Jupyter
```

SOME RIGHTS RESERVED

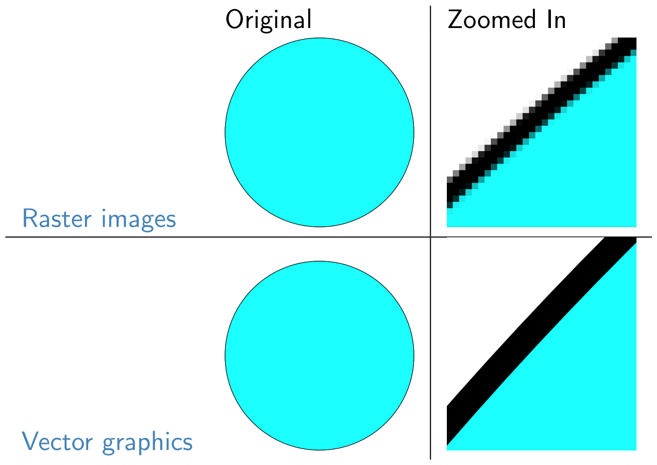
11.1.4 Scalable Vector Graphics

The digital images we talked about so far store colors in a large grid of pixels (a raster). A common problem with these types of images is that we cannot zoom in on them as far as we want, without loosing quality. At a certain point we start to see the individual pixels.

Vector graphics are an alternative way of storing digital images, which solve this problem.

Vector Graphics

- ▷ **Problem:** Raster images store colors in pixel grid. Quality deteriorates when image is zoomed into.
- ▷ Vector graphics solve this problem!



The idea of vector graphics is fundamentally different than the idea of raster images. Instead of storing pixels, we now store shape information!

For example, for a circle we don't store a color for each pixel, but we rather just store where the circle is, along with its radius, color, etc.

Vector Graphics (Definition)

- ▷ **Definition 11.1.36.** Image representation formats that store shape information instead of individual pixels, are refered to as vector graphics.
- ▷ **Example 11.1.37.** For a circle, just store
  - ▷ center
  - ▷ radius
  - ▷ line width
  - ▷ line color
  - ▷ fill color
- ▷ **Example 11.1.38.** For a line, store
  - ▷ start and end point
  - ▷ line width
  - ▷ line color

Note that most monitors cannot display vector graphics. There are vector monitors, but they are not common.

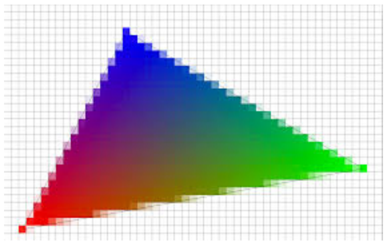
Vector Graphics Display

▷ There are devices that directly display vector graphics.

▷ **Example 11.1.39.**

▷ **Definition 11.1.40.** For monitors, vector graphics must be rasterized – i.e. converted into a raster image before display.

▷ **Example 11.1.41.**



The monitor displayed in ?? here does not have pixels. It instead moves a laser and traces a polygon (the asteroids and spaceship). The laser stimulates a phosphor layer, which then glows. Common monitors work with pixels. Vector graphics are thus rasterized (i.e. turned into raster images) just before being displayed. The rasterizer decides for each pixel, whether it is inside or outside the shape and thus what RGB value to display.
On the edges of ??, we see pixels whose barycenter is outside the triangle but that are colored in a very light variant of the adjoining pixels. This technique is called anti-aliasing and is used to make the jagged lines created by rasterization less noticeable to the human eye.
We now introduce a concrete representation format for vector graphics.

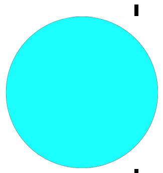
SVG is one image format for vector graphics. Since it is XML based we are able to read it. As described above, we can create circles by specifying a position, radius, and style (color etc).

Scalable Vector Graphics (SVG)

▷ **Definition 11.1.42.** Scalable Vector Graphics (SVG) is an XML-based markup format for vector graphics.

▷ **Example 11.1.43.**

```
<svg xmlns="http://www.w3.org/2000/svg"
     width="100" height="100" >
  <circle cx="50" cy="50" r="50"
      style="fill:#1cffff; stroke:#000000; stroke−width:0.1" />
</svg>
```



- ▷ The <svg> tag starts the SVG document, width, height declare its size.
- ▷ The <circle> tag starts a circle. cx, cy is the center point, r is the radius. style describes how the circle looks.

As the SVG size is 100x100 and the circle is at (50,50) with radius 50, it is centered and fills the whole region.

More SVG Primitives

▷ **Example 11.1.44 (Rectangle).**

```
<rect x="..." y="..." width="..." height="..." style="..." />
```

▷ **Example 11.1.45 (Ellipse).**

```
<ellipse cx="..." cy="..." rx="..." ry="..." style="..." />
```

▷ **Example 11.1.46 (Line).**

```
<line x1="..." y1="..." x2="..." y2="..." style="..." />
```

▷ **Example 11.1.47 (Text).**

```
<text x="..." y="..." style="...">This is my text!</text>
```

▷ **Example 11.1.48 (Image).**

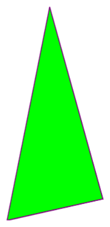
```
<image xlink:href="..." x="..." y="..." width="..." height="..." />
```

We can draw arbitrary polygons by specifying a list of coordinates.

SVG Polygons

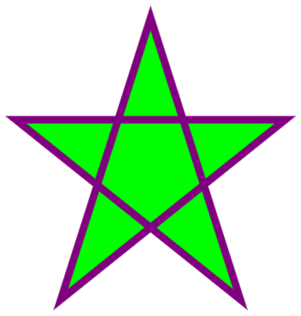
▷ **Example 11.1.49 (An SVG Triangle).**

```
<svg height="210" width="500" xmlns="http://www.w3.org/2000/svg">
  <polygon points="200,10 250,190 160,210"
           style="fill:lime;stroke:purple;stroke−width:1"/>
</svg>
```



▷ **Example 11.1.50 (An SVG Pentagram).**

```
<svg height="210" width="210" xmlns="http://www.w3.org/2000/svg">
  <polygon points="100,10 40,198 190,78 10,78 160,198"
           style="fill:lime;stroke:purple;stroke−width:5;fill−rule:nonzero;"/>
</svg>
```



SVG can directly be embedded in HTML!

SVG in HTML

▷ SVG can be used in dedicated files and referenced in a <**img**> tag. (file ending .svg)

▷ It can however also be written directly in HTML files.

▷ **Example 11.1.51.** Triangle from ?? embedded in HTML file

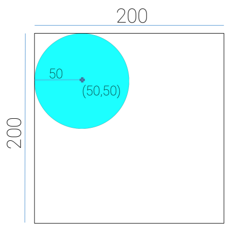
```
<html>
  <body>
    <svg height="210" width="500" xmlns="http://www.w3.org/2000/svg">
      <polygon points="200,10 250,190 160,210"
          style="fill:lime;stroke:purple;stroke−width:1" />
    </svg>
  </body>
</html>
```

We now explore a useful attribute of SVG called viewBox. We said that we can zoom in onto vector graphics as far as we want without loosing quality, so let's give ourselves the possibility to do so.

The SVG viewBox Attribute

- ▷ **Idea:** The SVG viewBox attribute allows us to zoom into an image.
- ▷ **Example 11.1.52.**

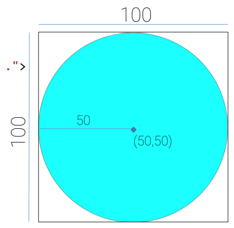
```
<svg width="200" height="200" xmlns="...">
  <circle cx="50" cy="50" r="50" style="..." />
</svg>
```



Here, the width and height are scaled by a factor of 2 to give us a little more room. Sometimes we want to specify a larger image, but only display a section of it.

- ▷ **Example 11.1.53.**

```
<svg width="200" height="200" xmlns="..."
      viewBox="0 0 100 100" >
  <circle cx="50" cy="50" r="50" style="..." />
</svg>
```



viewBox specifies a region inside our canvas. Only things inside that are drawn. The resulting image is then stretched to the canvas size (zoom effect).

The top example shows a 200 by 200 units large SVG canvas. In the top left quadrant we draw a circle.

The second code snippet employs the viewBox attribute, which specifies an area of the image we want to display. In this example we give it a region from (0,0) to (100,100), meaning we specify exactly this upper left quadrant.

viewBox now does two things: First, it only draws objects inside this region, i.e. it discards everything outside. Second, it stretches this region to the whole SVG canvas. This means, that our final image is still 200 by 200 units (pixels) in size, but we only see a region of our original image. This gives a zoom effect.

11.2 Project: An Image Annotation Tool

Project: Kirmes Image Annotation Tool

- ▷ **Problem:** Our Books-App project was a fully functional web application, but does not do anything useful for DigiHumS.
- ▷ **Idea:** Extend/Adapt it to a database for image annotation like LabelMe [LM].
- ▷ **Setting:** Prof. Peter Bell (formerly at FAU) conducts research on baroque paintings on parish fairs (Kirmes) and the iconography in these paintings. We want to build an annotation system for this research.
- ▷ **Project Goals:**

1. Collect kirmes images in a database and display them,
2. mark interesting areas and provide meta data,

3. display/edit/search annotated information.

1. is analogous to Books-App, for 2/3. we need to know more

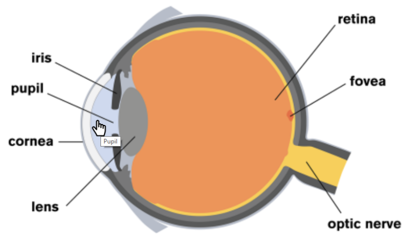
▷ **Plan:** Lern the necessary technologies in class, build the system in exercises

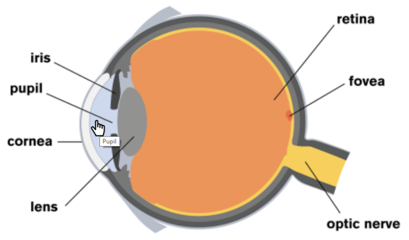
In our quest for an image annotation technology, we will first explore HTML image maps.

HTML Image Maps

▷ **Definition 11.2.1.** HTML image maps mark areas in an digital image and assign names and links to them.

▷ **Example 11.2.2.** An image map adds hover and on click behavior





Clicking on the pupil leads to: `https://en.wikipedia.org/wiki/Pupil`

Clicking on the vitreous body leads to: `https://en.wikipedia.org/wiki/Vitreous_body`

```
<html>
  <body>
    <img src="Human_Eye_Structures.png" usemap="#image−map"/>
    <map name="image−map">
      <area title="Pupil"
            href="https://en.wikipedia.org/wiki/Pupil"
            coords="102,117,143,219" shape="rect"/>
      <area title="Vitreous Body"
            href="https://en.wikipedia.org/wiki/Vitreous_body"
            coords="242,166,107" shape="circle"/>
    </map>
  </body>
</html>
```

▷ Easy creation of image maps: `https://www.image-map.net/`

Image maps provide a way to mark areas in an image. These areas act as links, i.e. clicking on them leads to different URLs. For example in this case there are two regions in the image (pupil and vitreous body), which - when clicked on - direct your browser to the respective Wikipedia articles.
<img> tag specifies image as always, but we no add a new attribute `usemap` that specifies the name of an image map to use (here image−map).

The map itself is defined by the **<map>** element (with the same name!). Inside the map we define our areas for the two parts of the eye we want to annotate. In this example we use a rectangle for the pupil and a circle for the vitreous body.

This is specified by the two **<area>** elements, which have a title attribute (shown on hover) and a link (href). The shapes are specified by the shape attribute with values rect, circle, poly, ... and some coordinates specified in the coords attribute.
Image maps are useful for certain tasks, but aren't quite what we want for our annotation tool. They are somewhat difficult to work with, especially if you want the areas to react to your mouse.

Problems of HTML Image Maps

- ▷ **Problem:** Image maps do not allow interaction:
  - ▷ the name attribute can only contain unstructured information.
  - ▷ no integrated highlight for image maps area,
  - ▷ no onclick or onmouseover attributes.

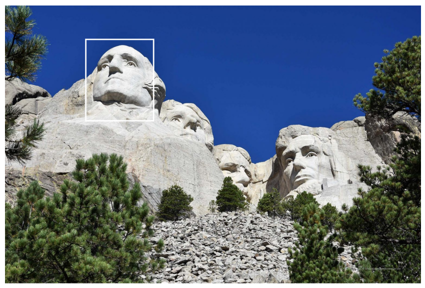
  But the whole point is to have (arbitrarily) complex metadata for image regions.
- ▷ **New Plan:** Use a newer technology: SVG and CSS.

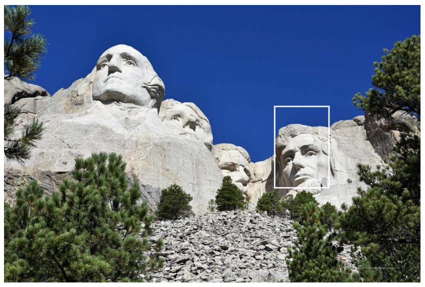
Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 322 2025-02-06

We therefore go a different route, by using SVG and CSS: The whole functionality of the annotation tool will be implemented in a single SVG image where CSS provides the interactivity.
First we implement the equivalent of an image map by including a raster graphic (our image) and four rectangles for the annotation areas. Coordinates of the rectangles can be read out from any image processing tool like Microsoft Paint or GIMP.

Handcrafting better Image Annotations with SVG and CSS

- ▷ **Idea:** Integrate the image and the areas into one SVG and make areas interactive via CSS.
- ▷ **Example 11.2.3 (Paper Prototype).** Highlight regions and display information on hover.


George Washington


Abraham Lincoln

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 323 2025-02-06

Displayed here is our goal behavior, which we will pursue on the following slides. As we have not implemented this, we could have created this in an image processing program, e.g. photoshop or GIMP. We call such a mockup for informing our design intuition a paper prototype.

The rectangles mark certain parts of our image and react to the mouse being moved over them. On the one hand the area is highlighted by the white rectangles. Additionally descriptive text is displayed below the image (in this case the name of the respective president).

SVG Annotation Implementation Areas

▷ **Implementing Areas as Rectangles:**

```
<svg xmlns="http://www.w3.org/2000/svg" width="1536" height="1024" >
  <!-- Image -->
  <image width="1536" height="1024" xlink:href="mount_rushmore.jpg" />
  <!-- Areas in image as rects. -->
  <rect x="300" y="125" width="250" height="300"/>
  <rect x="550" y="225" width="200" height="300"/>
  <rect x="750" y="375" width="200" height="300"/>
  <rect x="999" y="375" width="200"height="300"/>
</svg>
```

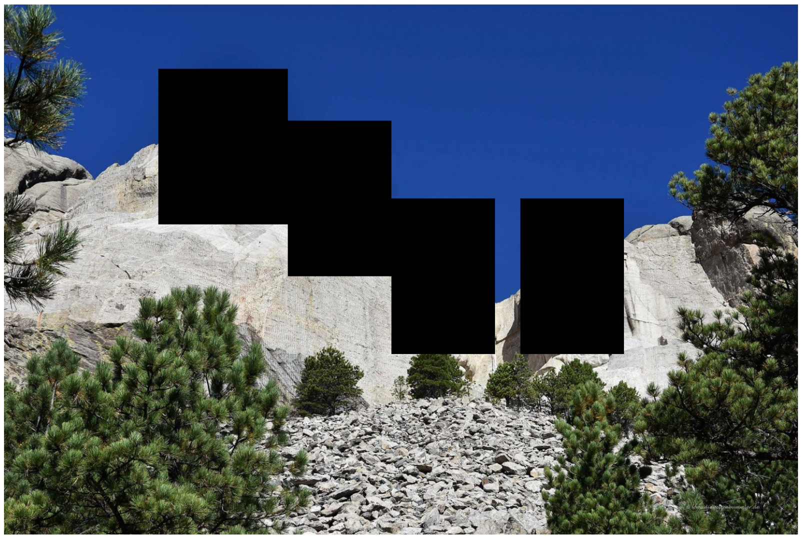
Add four <rect>s (one for each president).

Note again: The image is **not** a vector graphic. Even though it is embedded in a SVG environment, it will not have the benefits of a vector graphic, i.e. it will lose quality when zoomed in on.
Note furthermore: the order of elements in our SVG matters! Here the <rect> tags are specified *after* the image. SVG draws the elements from top to bottom. The rectangles are therefore drawn on top of the image.

Swapping this order would lead to the image being drawn on top of the rectangles. This means, that the rectangles would not be visible!

SVG Annotation Implementation Result

▷ **Areas as Rectangles – Result:** Now the rectangles are visible



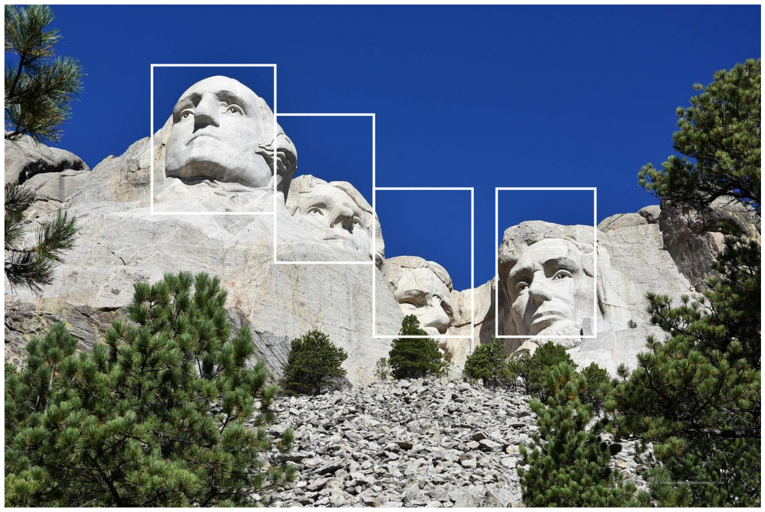
The rectangles are now visible in our SVG. Their color defaults to black, so let's fix this next, so that we can actually see our image again.

We add a CSS stylesheet to our site. This can either be defined in a separate file (like in this example), or be specified directly in the HTML inside of <**style**> tags.

Adding CSS for the Areas

▷ **Example 11.2.4 (Adding CSS).**

```
rect {fill—opacity:0; stroke:white; stroke—opacity:1; stroke—width:5px}
```



Our goal is to give the rectangles a solid white border, but no inner color. We thus change the stroke (border) parameters.

The fill opacity is set to zero, in order to make it completely transparent so we see the presidents' heads again. However, the rectangles are always visible and do not react to our mouse input. We will fix this next.

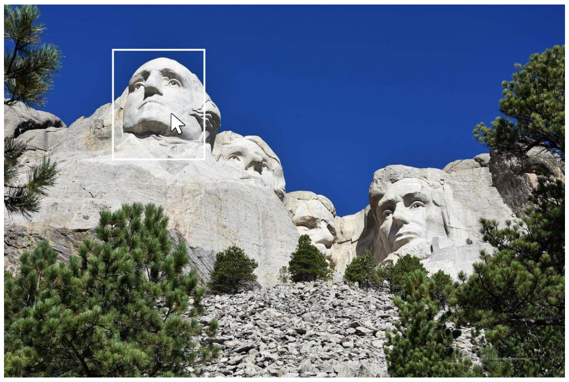
Selectively Highlighting Areas

▷ **Problem:** Now the rectangles are always visible.

▷ **Idea:** make the rectangles invisible by default only show them on hover.

▷ **CSS:** We set the stroke opacity to zero by default and add a hover selector.

```
rect {fill—opacity:0; stroke:white; stroke—opacity:0; stroke—width:5px}
rect:hover {stroke—opacity:1}
```



The hover selector of the rectangles specifies their style, whenever the mouse is over the element. This allows us to specialize the appearance for this case: we set the opacity back to one, meaning full opacity and thus visibility.

Net Effect: The rectangles are now invisible, expect when hovered over by the mouse.

We will now add the description text to each of our annotation areas. Since our text should appear below the image, let's start by giving ourselves a bit more room in the SVG canvas. We thus increase the SVG height by a bit. Note, that this does not impact the image (because it has an own height).

Adding Annotation Text

▷ **Adding Annotation Text** and making space for it.

```
<svg xmlns="http://www.w3.org/2000/svg" width="1536" height="1224" >
  <!-- Image -->
  <image width="1536" height="1024" xlink:href="mount_rushmore.jpg" />
  <!-- Areas in image as rects, text below -->
  <rect x="300" y="125" width="250" height="300" />
  <text x="100" y="1200">George Washington</text>
  <rect x="550" y="225" width="200" height="300" />
  <text x="100" y="1200">Thomas Jefferson</text>
  <rect x="750" y="375" width="200" height="300" />
  <text x="100" y="1200">Theodore Roosevelt</text>
  <rect x="999" y="375" width="200" height="300" />
  <text x="100" y="1200">Abraham Lincoln</text>
</svg>
```

and we add some CSS:

```
text {fill:black; opacity:1; font-size:100px}
```

We then add the text. Note, that all text elements have the exact same position below the image. They only differ in the text they display (the name of the president).

We write each text element directly below the corresponding rectangle tag, for reasons we will explain in a bit!

We also style the text: The text color is specified by the fill attribute. This is the default, so it's not really necessary to specify this. However, oftentimes it is advisable to be as verbose as possible with certain attributes, because this more clearly shows our intention.

Adding Annotation Text – Result

▷ **Adding Annotation Text – Result:**

[illegible]

The text is still unreadable, mainly because all texts are right above each other, but this is expected so far, since we specified all text tags to have the same position. Our main problem is, that the text does not react to our mouse input yet. Remember: Our goal is that each text element is only displayed, when the corresponding rectangle in the image is hovered by the mouse.
Our approach is analogous to the hovering of the rectangles we did previously. We text a default opacity of zero, and a hover opacity of one.

Remember though, that the hover selector always influences the element it is specified on, i.e. when writing `text:hover`, and then changing the opacity, this changes the opacity when we hover over the text, *not* when we hover the rectangle. We thus introduce the CSS sibling operator, +.

Selectively Showing Annotations

▷ **Problem:** Now the annotations are always visible.

▷ **Idea:** Add CSS hover effect for <rect>s, which effects the |<text>|.

▷ **Definition 11.2.5.** The CSS sibling operator + modifies a selector so that it (only) affects following sibling elements (same level).

▷ **Example 11.2.6.** In the CSS directive

```
rect:hover + text {<rules>}
```

Selector Sibling operator Target

the rules affect the SVG <text> directly after the <rect> element.

▷ **Again:** The order of elements in the HTML is important!

▷ **CSS:** We set the opacity to zero by default and add a hover selector for the following <text> sibling.

```
text {fill:black; opacity:0; font−size:100px}
rect:hover + text {opacity: 1}
```

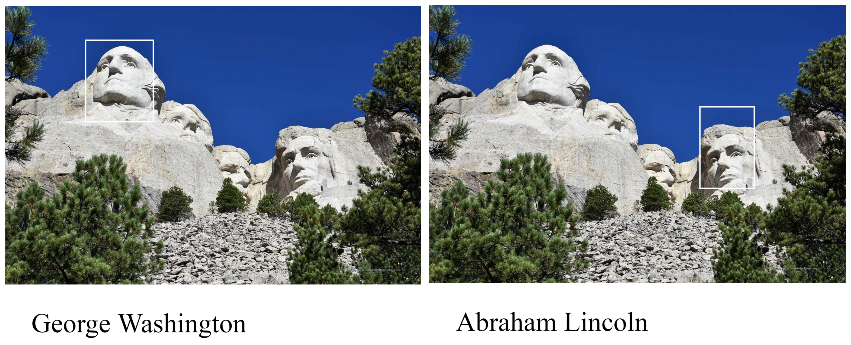
The sibling operator influences the next element of the specified type (in our case text) in the HTML/SVG. This is why earlier we put the text nodes always directly after the rectangle.

This way, when a rectangle is hover over, the next text node is always the corresponding description and will thus become visible.

Image Annotation Tool – Final Result

▷ Now our annotation tool works as expected!

▷ **Example 11.2.7 (Final Result).** Highlight regions and display information on hover.



11.3 Fun with Image Operations: CSS Filters

Let's explore more the capabilities CSS has to offer for applying operations to digital images. In this example we make an image gray, by specifying a grayscale filter attribute. The argument of the filter gives us the possibility to make the image only a litte gray. Since it is set to 100% in this example, the image is converted to perfect grayscale.

CSS Image Filters

▷ **Goal:** Apply image filters (grayscale etc.) directly in CSS.

▷ **Example 11.3.1 (Image Effects via inline CSS).**

```
<img style="filter: grayscale(100%)" src="augustus.jpg" alt="no image"/>
```

▷ **Disadvantage:** The original image is delivered to client. When user saves the image, they get the original!

One extremely important thing to keep in mind is that CSS is executed on the client (the user's browser). The original image or text is delivered to the client, where the filter is applied. You can try this out by right-clicking a filtered image on a web site and saving it to your hard drive. Note, that the original digital image is saved!

The implication here is, that for certain content it is best to perform the filter on the server and then deliver the filtered content to the user, so that he or she does not even have the possibility to get the original. This however also means more computation on the server, which might be expensive.

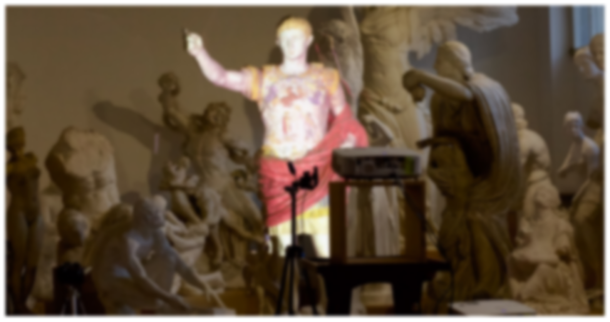
Rule of thumb: Perform as much as possible on the client (CSS and JavaScript) and as much as necessary on the server (for example python in bottle).

Here are more examples of image filters. The CSS selectors here start with dots. This makes them influence HTML elements of the respective class name, i.e. the selector `.shadow` gives the HTML element with class `shadow` a drop shadow.

Some more CSS Filters

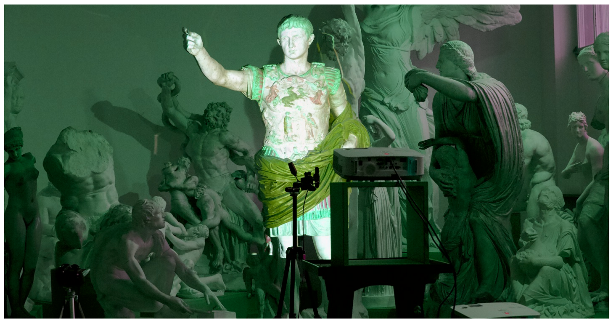
▷ **Example 11.3.2 (Image Effects via CSS Style sheets).**

```
<img style="filter: blur(4px" src="augustus.jpg" alt="no image"/>
```



```
<img style="filter: contrast(180%)" src="augustus.jpg" alt="no image"/>
```

```
<img style="filter: hue-rotate(90deg)" src="augustus.jpg" alt="no image"/>
```



Blurring: Blur is an image operation, which mixes each pixel's color with the colors of its neighbor. The operation is thus similar to our edge detection example from earlier, but with different weights per neighbor pixel.

Also, for blur it is possible to specify larger neighborhoods. In this case the radius of our neighborhood is 4 pixels, meaning that we mix the colors of a region with radius 4.

Contrast: Contrast makes dark colors darker and light colors lighter for arguments over 100%. This increases the range between the darkest and lightest pixel.

For arguments under 100%, the contrast shrinks.

Hue Rotation: The color wheel at the top might look familiar to you. It is a standard way of displaying colors. The outer ring is roughly equivalent with the colors of the rainbow (with some exceptions; purple for example is not a rainbow color).

The hue-rotate filter rotates this color wheel, such that each color lands in a different spot. In our example (90deg), red becomes green. This effect can be observed on Augustus' cloak.
Another useful thing is the combination of CSS filters. For example you can blur an image and then convert it to grayscale, as showcased in the example.

Combining CSS Filters

▷ **Idea:** We can also combine image filters flexibly. The easist way is when we define CSS classes for that.

▷ **Example 11.3.3 (Tie CSS Filters to Classes).**

```
<html>
  <head>
    <style type="text/css">
```

```
      .blur { filter: blur(4px); }
      .brightness { filter: brightness(0.30); }
      .contrast { filter: contrast(180%); }
      .grayscale { filter: grayscale(100%); }
      .huerotate { filter: hue-rotate(180deg); }
      .invert { filter: invert(100%); }
      .opacity { filter: opacity(50%); }
      .saturate { filter: saturate(7); }
      .sepia { filter: sepia(100%); }
      .shadow { filter: drop-shadow(8px 8px 10px green); }
    </style>
  </head>
  <body>
    <img class="blur invert saturate" alt="no image" src="augustus.jpg" />
  </body>
</html>
```

▷ **Note:** The order is important: Changing the order of filters yields different results.

Digital image are not the only HTML element which can be filtered. It turns out that you can apply filters to nearly everything in HTML, for example text. Note that here we are using the `blur` class from earlier.

Filtering Everyghing Else

▷ **Note:** CSS filters don't just apply to images! (Almost) everything can be filtered.

▷ **Example 11.3.4 (Filtering Text (Blurring)).**

```
<p style="filter: blur(3px)">A severely blurred Text</p>
```

A fun thing to play around with are CSS animations.

CSS Animations

▷ **Definition 11.3.5.** CSS animations change state of an object over time.

▷ **Example 11.3.6 (Inverting an image).**

```
img {animation: invertAnimation 1s forwards}

@keyframes invertAnimation {
    from {filter: none}
    to {filter: invert(100%)}
  }
```

In this case we define an animation called *invertAnimation* which applies an inversion-filter. The syntax specifies that at the beginning of the animation, no filter should be applied and in the end we want the image to be completely inverted.

We then apply the animation to all elements of tag `<img>`. We declare that the animation should run one second (1s), so the image is inverted after one second.

The last attribute specifies what should happen after the animation is completed. `forwards` means that the element should simply stay how it is, so it stays inverted after the one second.

SVG Filters

▷ **Note:** Unfortunately in SVG the filtering works differently from CSS.

▷ **Example 11.3.7 (Blurring Mt. Rushmore in SVG).**

```
<svg xmlns="http://www.w3.org/2000/svg" width="1536" height="1024">
  <style> image {filter: url(#myCustomFilter)}</style>
  <image width="1536" height="1024" xlink:href="mount_rushmore.jpg" />
  <!-- Image filter -->
  <filter id="myCustomFilter">
    <feGaussianBlur stdDeviation="5" />
  </filter>
</svg>
```

▷ **Example 11.3.8 (SVG Filters can be combined).**

```
  <filter id="myCustomFilter">
    <feGaussianBlur stdDeviation="5" />
    <feColorMatrix type="saturate" values="0.1" />
  </filter>
```

In the first example we define a filter at the bottom. We give it a name (*myCustomFilter*), which we can then reference in the CSS snippet above. With the `url` function we can apply a filter with the given name to all images.

The *Gaussian Blur* filter here is similar to the *blur* filter in CSS.
Similarly to HTML, we can combine filters in SVG as well. In the second example we apply a saturation filter after the blur. This is similar to a grayscale filter.

11.4 Exercises

Problem 4.1 (Basic Image Manipulation)

In this exercise we will explore Pillow's image processing capabilities. Create a new Python file ImageManip.py and import the Image and ImageOps modules like this:

```
from PIL import Image, ImageOps
```

Write a Python function transformImage, which takes as arguments a digital image and a string. The string describes, which transformation should be applied to the image. For example, if the value of the passed string is "gray", your function should convert the image to grayscale and return the resulting image.

You find a complete list of Pillow's image processing functions here: `https://pillow.readthedocs.io/mod/stable/reference/ImageOps.html`. Your function should at least support five of them.

You can freely choose the string value you want to assign each operation. For example, if you want to support the grayscale operation, you can choose whether the expected string is supposed to be "gray" or "grayscale" or something else, as long as it is sensible.

If the passed string does not match any operation, just return the original image.

Outside the function, load an image from your hard drive using Pillow's Image.open function. You may use one of the images in the Kirmes repository or use one of your own digital images.

Test your transformImage function by passing the image, along with some strings specifying the image operation. Display the transformed image using Pillow's show functionality.

Refer to the lecture notes for examples of the open and show methods.

Problem 4.2 (Watermarking Images)

In this exercise we will add functionality to apply a watermark to a digital image. We provide a watermark image (Watermark.png) together with this assignment (StudOn and Kirmes repository), but feel free to create one yourself.

Create a new Python function applyWatermarkToImage, which takes an image as argument. In the function, load the watermark image from your hard drive. Then use Pillow's alpha_composite function to overlay the watermark on top of the input image: `https://pillow.readthedocs.io/mod/stable/reference/Image.html#PIL.Image.Image.alpha_composite`

Note that there are two versions of alpha_composite in Pillow. The one we are using here directly modifies the original image and does not return a new one.

At the end of your function, convert the watermarked image back to RGB (analogous to above) and return the result.

Test your function and show the watermarked image! You can also use the save function to write the image to your hard drive:

```
im.save(``filename.jpg'', ``JPEG'')
```

Optional for the highly motivated: Check out the following tutorial, if you want to write arbitrary text as watermark: `https://pillow.readthedocs.io/mod/stable/reference/ImageDraw.html#example-draw-partial-opacity-text` Note: When they load a font (fnt = ImageFont.truetype(...)), just pass "arial.ttf" as argument (or another font which is installed on your PC).

Problem 4.3 (Putting Thumbnails in Database)

Our image database and front-end are taking shape. On the home page we currently show an overview of all entries including thumbnails.

These thumbnails are small (200 pixels wide), yet we always load the full size image from the database. This is not particularly efficient, since all these (potentially very large) digital images need to be transferred to the client. We will try to fix this in this exercise.

We provide two new Python files with this exercise (ImageManip.py and ImageHelper.py). The first provides some basic image processing techniques (from last week). The latter provides functionality to create Pillow images from binary data (and vice versa) or to load Pillow images from a URL.

Familiarize yourself with the two files. You do not need to understand everything in the Python code, but make sure that you read the comments and that you understand what kind of functionality is given.

Now perform the following tasks:

1. In the BuildDB.py script, import the two provided files and Pillow:

   ```
   import ImageHelper
   import ImageManip
   from PIL import Image
   ```

2. In the BuildDB.py script add one more column to the database called Thumbnail of type BLOB. This will store our thumbnail.

3. Adapt the addImage function, such that it creates a Pillow image from the imageData variable (look in the ImageHelper file for a function you can use for this task). Create the thumbnail (see file ImageManip). Then convert the image back to a binary blob and store it in the Thumbnail field of our database.

   See the comments in the BuildDB.py file for more details.

4. In the Server.py script add a new route /thumbnail/<id:int>. This should be exactly the same as the /imageraw/<id:int> route (which already exists), with one exception: It should return the Thumbnail instead of the Content field.

5. Lastly, in the Index.tpl make sure, that your new /thumbnail route is used instead of the /imageraw. On the details page the original sized image should stay of course.

Problem 4.4 (Displaying Annotations)

In this exercise we will finally give our database frontend the ability sto display annotations on top of our images. For now, these annotations come from files already provided in the Kirmes repository in the xml/ subfolder. Each of the files in this directory describes areas (rectangles) in a given image, along with a description text.

We have prepared the parsing of these files for you, so you don't need to change anything in the BuildDB.py script. Nevertheless, check the table creation near the end of the file (from line 246). In addition to the Images table we worked with for the last couple of weeks, we now have a second table in our database, called Annotations. This table stores the following information:

1. Id: The id of the annotation (analogous to the Id field in the Images table).

2. ImageId: The id of the annotated digital image.

3. Description: A text describing the annotations.

4. X, Y, Width, Height: The position and dimensions of the rectangle in the digital image.

The ImageId is a foreign key, which references the primary key Id attribute of the Images table. For example, an annotation entry with ImageId=27 defines an annotation for the image entry with Id=27. Note, that multiple annotations might reference the same digital image.

You don't need to do anything in this file, but make sure that you run it, so that your database is filled with the annotation data. Double check in the DB Browser, that the Annotation table is properly created and filled.

Now our frontend just needs to display the annotation information. To this end, amend the /details/ route in the Server.py script, such that for the given image id, it queries the database for annotations.

In the Details.tpl file, iterate over the annotations (if any exist), and create a <rect> and a <text> for each. Fill in the information from the annotation (position and size of the rectangle, description for the text). See the lecture notes for details, if you are unsure how this works.

Check if everything works as expected by visiting the /details/ page for an image, which has annotations. Not too many images actually have annotations, but some do. For example the image with id 146 should have a couple.

Make sure that by hovering the mouse over an annotation region, the rectangle highlights (gets brighter) and the description text is shown.

We will now give the user the ability to edit annotations directly in the browser. The idea is that changing the values of an annotation (position, size, text) is always easier in a graphical user interface than by typing in the values in an XML file.

The process requires two parts. First the user must be able to interactively change the values in the browser. Second, the changes they made must be saved back to the database.

In order to ensure a pleasant user experience the first part should be performed directly in the browser, so that not every mouse click must be sent to the server and back. Since this requires JavaScript, we have provided this part for you.

Run your server and visit a details page of any image, which has annotations, e.g. `http://localhost:8080/details/146`. At the bottom you should see a checkbox Edit Annotations. If this is checked, you should see a list of all annotations.

The currently selected element in this list is editable. You can change the annotation description in the text box. You can change the position and size of the annotation rectangle by dragging the marked (red) rectangle in the image. Note that you can both move and resize the rectangle.

New annotations can be added with the New Annotation button at the bottom and deleted by clicking the bin icon.

The changes you made are sent to the server, when the Save Changes button is clicked. Saving the changes in the database is for you to implement.

Right now clicking Save Changes should do nothing (even though the website displays a notification saying that the changes have been saved).

You can verify that saving is not working by making some changes. Then click Save Changes and refresh the page. All changes should be gone (because they are not stored in the database).

Problem 4.5 (Editing Annotations)

In the Server.py script you can find a new route /edit_annotations. Since this receives data (i.e. the changes you made to the annotations), it is marked as POST.

The function loops over a list of changes and gets the necessary data.

Implement the following: For each entry in the list of changes, issue the correct SQL command to update the values (hint: UPDATE ...). At the end of the function, commit your changes to the database (db.commit()).

Test your function! In the browser, edit one or multiple annotations and click Save Changes. Refresh the page. Your changes should still be there!

Problem 4.6 (Deleting Annotations)

Complete the /edit_annotations route by issuing a DELETE command for each entry passed to this function. Again, don't forget to commit your changes.

Test your code by deleting entries in the browser and refreshing the page!

Problem 4.7 (Adding Annotations)

Adding new annotations (/new_annotations) is slightly more complicated (but not much). Note that this function takes in the imageID as an argument.

In the loop, extract the individual fields from the annotation variable (similar to the way it's done in /edit_annotations). Since this is a **new** annotation, there is no annotationID this time.

Issue an INSERT command for each new annotation. Then get the id of the newly stored entry (cursor.lastrowid) and append this id to the newIds list. These new ids will be sent back to the client (browser) at the end of the function. This is already implemented.

Lastly, test your functionality! You should now be able to add new annotations in the browser, which will persist even if you refresh the page.

Chapter 12

Ontologies, Semantic Web for Cultural Heritage

In the last chapter IWGS, we will discuss a virtual research environment for cultural heritage. Before we present the system itself, we take a close look at the underlying technology: ontologies, semantic web technologies, and linked open data.

12.1 Documenting our Cultural Heritage

Before we even start talking about the WissKI system, we should become clear on the concepts involved. We start out with the notion of cultural heritage itself.

Documenting our Cultural Heritage

- ▷ **Definition 12.1.1.** Cultural heritage is the legacy of physical artifacts cultural artefacts and practices, representations, expressions, knowledge, or skills – intangible cultural heritage (ICH) of a group or society that is inherited from past generations.
- ▷ **Problem:** How can we understand, conserve, and learn from our cultural heritage?
- ▷ **Traditional Answer:** We collect cultural artefacts, study them carefully, relate them to other artefacts, discuss the findings, and publish the results. We display the artefacts in museums and galleries, and educate the next generation.
- ▷ **DigHumS Answer:** In "Digital Humanities and Social Sciences", we want to represent our cultural heritage digitally, and utilize computational tools to do so.
- ▷ **Practical Question:** What are the best representation formats and tools?

There is another context in which we want to understand the WissKI system: that of research data. We will introduce the basic concepts now.

Research Data in a Nutshell

- ▷ **Definition 12.1.2.** Research data is any information that has been collected, observed, generated or created to validate original research findings. Although usually digital, research data also includes non-digital formats such as laboratory notebooks and diaries.

- **Types of research data:**
  - documents, spreadsheets, laboratory notebooks, field notebooks, diaries,
  - questionnaires, transcripts, codebooks, test responses,
  - audiotapes, videotapes, photographs, films,
  - cultural artefacts, specimens, samples,
  - data files, database contents (video, audio, text, images), digital outputs,
  - models, algorithms, scripts,
  - contents of an application (input, output, logfiles, schemata),
  - methodologies and workflows, standard operating procedures, and protocols,
- **Non-digital Research Data** such as cultural artefacts, laboratory notebooks, ice-core samples, or sketchbooks is often unique. Materials could be digitized, but this may not be possible for all types of data.

SOME RIGHTS RESERVED

The very idea of research data is they are retained to justify the published research: in particular just publishing tables of results and experiment descriptions in journals is not enough.

In the past, this has led to the practice of keeping meticulous lab books in the experimental sciences, and in recent times to the practice of publishing original data together with the results, so that experiments can be replicated and derived results can be re-calculated. This being pushed through the scientific organizations in the last decades.

But publishing raw data is also insufficient: experiments can only be replicated and derivations can only be checked if the underlying data can be obtained in practice, are complete and correct, and can be interpreted by the reader. This led to substantial institutional attention and – consequently – to many new developments:

FAIR Research Data: The Next Big Thing

- **Principle:** Scientific experiments must be replicated, and derivations must be checkable to be trustworthy. (consensus of scientific community)
- **Intuition:** Research data must be retained for justification, shared for synergies!
- **Consequence:** Virtually all scientific funding agencies now require some kind of research data strategy in proposals. (tendency: getting stricter)
- **Problem:** Not all forms of data are actually useable in practice.
- **Definition 12.1.3 (Gold Standard Criteria).** Research data should be FAIR:
  - Findable: easy to identify and find for both humans and computers, e.g. with metadata that facilitate searching for specific datasets,
  - Accessible: stored for long term so that they can easily be accessed and/or downloaded with well-defined access conditions, whether at the level of metadata, or at the level of the actual data,
  - Interoperable: ready to be combined with other datasets by humans or computers, without ambiguities in the meanings of terms and values,

- ▷ Reusable: ready to be used for future research and to be further processed using computational methods.

Consensus in the research data community; for details see [FAIR18; Wil+16].

▷ **Open Question:** How can we achieve FAIR-ness in a discipline in practice?

After these general considerations about research data, let us come back our primary concern in IWGS: research data in the humanities and social sciences.
If we look at the categories of research data we can expect in the humanities and social sciences, then we can categorize them into four broad categories. And we can see that we have already learned about many of them in IWGS.

Categories of Data in DigiHumS and their Formats

▷ We distinguish four broad categories of data in DigiHumS.

▷ **Definition 12.1.4.** Concrete data: digital representations of artefacts in terms of simple data,

- ▷ e.g. raster images as pixel arrays in JPEG. (see ??)
- ▷ e.g. books identified by author/title/publisher/pubyear. (see ??)

▷ **Definition 12.1.5.** Narrative data: documents and text fragments used for communicating knowledge to humans.

- ▷ e.g. plain text and formatted text with markup code (see ??)

▷ **Definition 12.1.6.** Symbolic data: descriptions of object and facts in a formal language

- ▷ e.g. 3+5 in Python (see ??)

▷ **Definition 12.1.7.** Metadata: "data about data", e.g. who has created these facts, images, or documents, how do they relate to each other? (not covered yet)

▷ **Observation 12.1.8.** *Metadata are the resources, DigiHumS results are made of (⇝ support that)*
The other categories digitize artefacts and auxiliary data.

▷ **Observation 12.1.9.** *We will need all of these – and their combinations – to do DigiHumS.*

The last kind – metadata – is arguably the most important kind in the it concerns the relations between artefacts, which are usually digitized into concrete data.

WissKI: a Virtual Research Env. for Cultural Heritage

▷ **Definition 12.1.10.** WissKI is a virtual research environment (VRE) for managing scholarly data and documenting cultural heritage.

- ▷ **Requirements:** For a virtual research environment for cultural heritage, we need
  - ▷ scientific communication about and documentation of the cultural heritage
  - ▷ networking knowledge from different disciplines (transdisciplinarity)
  - ▷ high-quality data acquisition and analysis
  - ▷ safeguarding authorship, authenticity, persistence
  - ▷ support of scientific publication
- ▷ WissKI was developed by the research group of Prof. Günther Görtz at FAU Erlangen-Nürnberg and is now used in hundreds of DH projects across Germany.
- ▷ FAU supports cultural heritage research by providing hosted WissKI instances.
  - ▷ See `https://wisski.data.fau.de` for details
  - ▷ We will use an instance for the Kirmes paintings in the homework assignments.

This leads to the following plan for the rest of the chapter.

Documenting Cultural Heritage: Current State/Preview

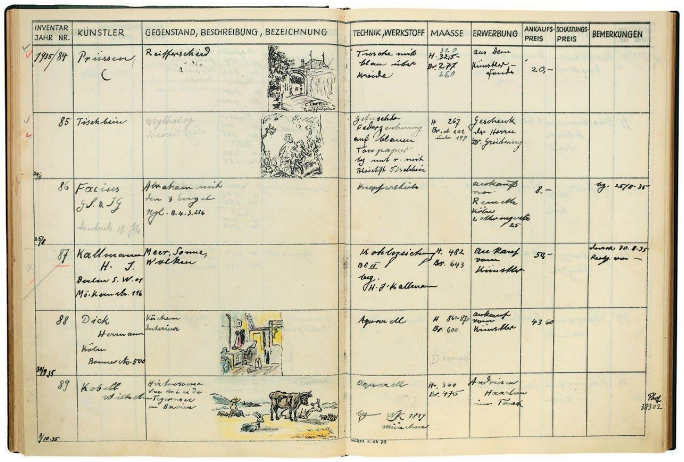
- ▷ Pre-DH State of cultural heritage documentation:
  - ▷ **scientific communication/documentation** by journal articles/books
  - ▷ **persistence**: paper records, file cards, databases (like our KirmesDB)
  - ▷ **Analysis**: manual examination of artefacts in museums/archives.
- ▷ **Idea:** Use more technology to do better.
- ▷ **Preview:** WissKI uses semantic web technologies to do just that. We will now
  - ▷ Motivate the semantic web (why do we need more than the WWW)
  - ▷ introduce ontologies, linked open data and their technology stacks
  - ▷ show off WissKI and offer a little project based on Kirmes corpus.

12.2 Systems for Documenting the Cultural Heritage

Let us now have a look at how we can use digital systems to document the cultural heritage. This is the backdrop against which we need to position the WissKI system.
The traditional methods of documenting cultural artefacts is in form of often handwritten – ledgers that inventory the collections of museums.

Documenting Cultural Artefacts: Inventory Books

- ▷ **Definition 12.2.1.** An inventory book is a ledger that identifies, describes, and records provenance of the artefacts in the collection of a museum.
- ▷ **Example 12.2.2 (An Inventory Book).**



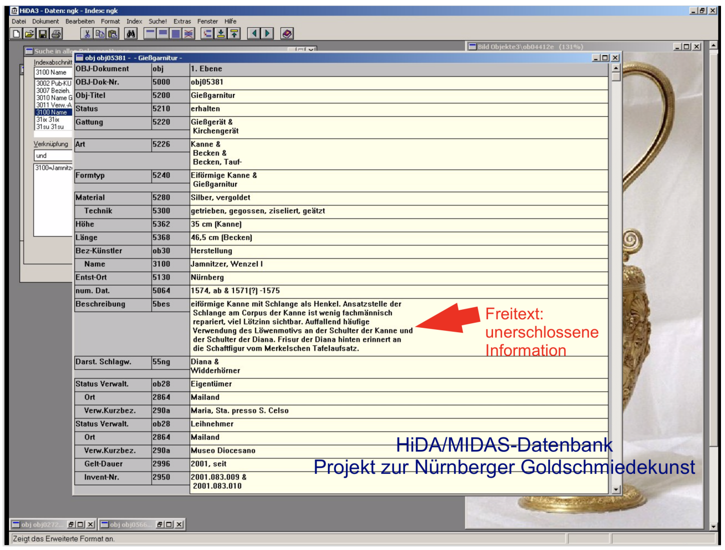
INVENTAR JAHR NR.	KÜNSTLER	GEGENSTAND, BESCHREIBUNG, BEZEICHNUNG	TECHNIK, WERKSTOFF	MAASSE	ERWERBUNG	ANKAUFS-PREIS	SCHÄTZUNGS-PREIS	BEMERKUNGEN
1935/84	Prüssen	Reifferscheid	Tusche mit blau über Kreide	31,0 H. 32,5 Br. 27,7 26,0	aus dem Künstlerfunde	20,–		
85	Tischbein	Mytholog. Scene	Getuschte Federzeichnung auf blauem Tonpapier bez. unt. r. mit Bleistift Tischbein	H 267 Br. ob. 202 unten 199	Geschenk des Herrn Dr. Greitering			
86	Facius G.F. & J.G.	Abraham mit den 3 Engel Ngl. B. 4. S. 216	Kupferstich		ankauf von Renard Köln Lotharingerstr. 25	8.–		bz. 25/8-35
87	Kallmann H. J. Berlin S.W. 61 Möckernstr. 116	Meer, Sonne, Wolken	Kohlezeichnung Bl. II bez. H. J. Kallmann	H. 482 Br. 643	Ankauf vom Künstler	50.–		durch 30. 8. 35 Rech. von –
88	Dick Hermann Köln Bonnerstr. 500	Küchen Interieur	Aquarell	H 860–870 Br. 600	ankauf vom Künstler	43 60		
89	Kobell Wilhelm	Hirtenscene Vue du Lac de Tegernsee en Bavière	Aquarell bez. WK 1819 München	H. 364 Br. 475	Andriesen Haarlem im Tausch			Phot. 38302

▷ **Problems:** non-digital, only single-user access, institution-local, no querying, . . .

If we want to improve on – or just digitize inventory books, the most obvious idea at least with what we have learned in IWGS – is to put the data into a database for persistence and use a web application for the user interface. Instead of surveying the multitude existing systems we want to improve on, let us briefly show an example.

Cultural Artefacts in Databases: Example

▷ **Example 12.2.3.** A typical database for cultural artefacts: (HiDa/MIDAS)


HiDA3 - Daten: ngk - Index: ngk
Datei Dokument Bearbeiten Format Index Suche! Extras Fenster Hilfe
Bild Objekte3\ob04412e (131%)
obj obj05381 - - Gießgarnitur -
Freitext: unerschlossene Information
HiDA/MIDAS-Datenbank
Projekt zur Nürnberger Goldschmiedekunst
Zeigt das Erweiterte Format an.

OBJ-Dokument	obj	1. Ebene
OBJ-Dok-Nr.	5000	obj05381
Obj-Titel	5200	Gießgarnitur
Status	5210	erhalten
Gattung	5220	Gießgerät & Kirchengerät
Art	5226	Kanne & Becken & Becken, Tauf-
Formtyp	5240	Eiförmige Kanne & Gießgarnitur
Material	5280	Silber, vergoldet
Technik	5300	getrieben, gegossen, ziseliert, geätzt
Höhe	5362	35 cm (Kanne)
Länge	5368	46,5 cm (Becken)
Bez-Künstler	ob30	Herstellung
Name	3100	Jamnitzer, Wenzel I
Entst-Ort	5130	Nürnberg
num. Dat.	5064	1574, ab & 1571(?) -1575
Beschreibung	5bes	eiförmige Kanne mit Schlange als Henkel. Ansatzstelle der Schlange am Corpus der Kanne ist wenig fachmännisch repariert, viel Lötzinn sichtbar. Auffallend häufige Verwendung des Löwenmotivs an der Schulter der Kanne und der Schulter der Diana. Frisur der Diana hinten erinnert an die Schaftfigur vom Merkelschen Tafelaufsatz.
Darst. Schlagw.	55ng	Diana & Widderhörner
Status Verwalt.	ob28	Eigentümer
Ort	2864	Mailand
Verw.Kurzbez.	290a	Maria, Sta. presso S. Celso
Status Verwalt.	ob28	Leihnehmer
Ort	2864	Mailand
Verw.Kurzbez.	290a	Museo Diocesano
Gelt-Dauer	2996	2001, seit
Invent-Nr.	2950	2001.083.009 & 2001.083.010

The system we see above is an instance of the HiDa/MIDAS system, which is in use in many museums for managing their collections. HiDa [HiDa] is a conventional (and commercial) relational database with a sophisticated user interface for data acquisition, reporing, exporting, and publication. Database schemata can be chosen from a set of options; here we see the MIDAS schema [BHK16].

The HiDa/MIDAS system is by no means the only one on the market, but the architecture is typical for the state of the art in most cultural institutions worldwide.

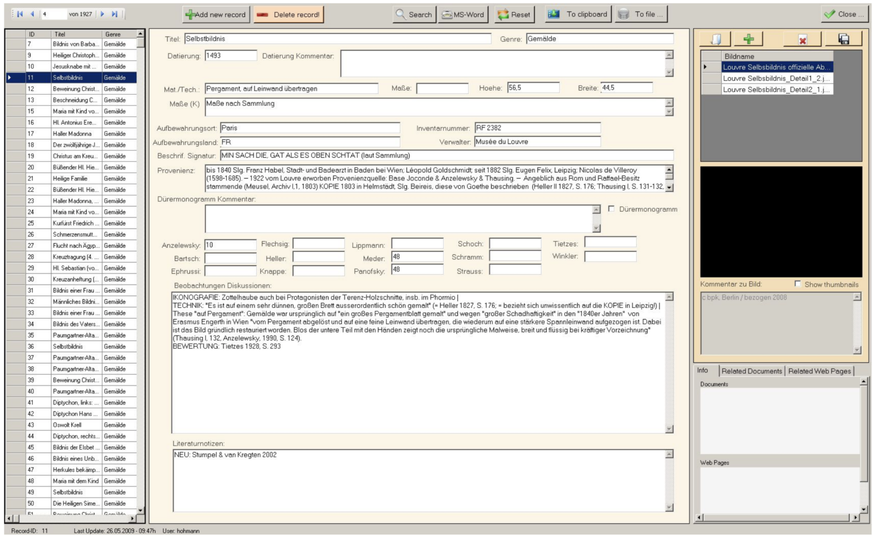
Cultural Artefacts in Databases: Pro/Con

- ▷ **Databases of Cultural Artefacts – Advantages:**
  - ▷ persistence, multi-user access, structured data,
  - ▷ web/catalog publication, standardized exports,
  - ▷ standardized performant query language.
- ▷ **Databases of Cultural Artefacts – Problems:**
  - ▷ identifiers are database local ⇝ no trans database relations,
  - ▷ database schemata are inflexible ⇜ we need extensions in practice,
  - ▷ free text as an un-structured, untapped resource.
- ▷ **Idea:** Relational databases impose structure, let's try something very unstructured: the world wide web. (up next)

Here is another example.

Cultural Artefacts in Databases II

- ▷ **Example 12.2.4.** Another database for cultural artefacts:



ID	Titel	Genre
7	Bildnis von Barba...	Gemälde
9	Heiliger Christoph...	Gemälde
10	Jesusknabe mit ...	Gemälde
11	Selbstbildnis	Gemälde
12	Beweinung Christ...	Gemälde
13	Beschneidung C...	Gemälde
15	Maria mit Kind vo...	Gemälde
16	Hl. Antonius Ere...	Gemälde
17	Haller Madonna	Gemälde
18	Der zwölfjährige J...	Gemälde
19	Christus am Kreu...	Gemälde
20	Büßender Hl. Hie...	Gemälde
21	Heilige Familie	Gemälde
22	Büßender Hl. Hie...	Gemälde
23	Haller Madonna, ...	Gemälde
24	Maria mit Kind vo...	Gemälde
25	Kurfürst Friedrich ...	Gemälde
26	Schmerzensmutt...	Gemälde
27	Flucht nach Ägyp...	Gemälde
28	Kreuztragung (4. ...	Gemälde
29	Hl. Sebastian (vo...	Gemälde
30	Kreuzanheftung (...	Gemälde
31	Bildnis einer Frau ...	Gemälde
32	Männliches Bildni...	Gemälde
33	Bildnis einer Frau ...	Gemälde
34	Bildnis des Vaters...	Gemälde
35	Paumgartner-Alta...	Gemälde
36	Selbstbildnis	Gemälde
37	Paumgartner-Alta...	Gemälde
38	Paumgartner-Alta...	Gemälde
39	Beweinung Christ...	Gemälde
40	Paumgartner-Alta...	Gemälde
41	Diptychon, links: ...	Gemälde
42	Diptychon Hans ...	Gemälde
43	Oswolt Krell	Gemälde
44	Diptychon, rechts...	Gemälde
45	Bildnis der Elsbet ...	Gemälde
46	Bildnis eines Unb...	Gemälde
47	Herkules bekämp...	Gemälde
48	Maria mit dem Kind	Gemälde
49	Selbstbildnis	Gemälde
50	Die Heiligen Sime...	Gemälde

Titel: Selbstbildnis Genre: Gemälde
Datierung: 1493 Datierung Kommentar:
Mat./Tech.: Pergament, auf Leinwand übertragen Maße: Hoehe: 56,5 Breite: 44,5
Maße (K) Maße nach Sammlung
Aufbewahrungsort: Paris Inventarnummer: RF 2382
Aufbewahrungsland: FR Verwalter: Musée du Louvre
Beschrif. Signatur: MIN SACH DIE, GAT ALS ES OBEN SCHTAT (laut Sammlung)
Provenienz: bis 1840 Slg. Franz Habel, Stadt- und Badearzt in Baden bei Wien; Léopold Goldschmidt; seit 1882 Slg. Eugen Felix, Leipzig; Nicolas de Villeroy (1598-1685). – 1922 vom Louvre erworben Provenienzquelle: Base Joconde & Anzelewsky & Thausing. – Angeblich aus Rom und Raffael-Besitz stammende (Meusel, Archiv I,1, 1803) KOPIE 1803 in Helmstädt, Slg. Beireis, diese von Goethe beschrieben (Heller II 1827, S. 176; Thausing I, S. 131-132,
Dürermonogramm Kommentar: ☐ Dürermonogramm
Anzelewsky: 10 Flechsig: Lippmann: Schoch: Tietzes:
Bartsch: Heller: Meder: 48 Schramm: Winkler:
Ephrussi: Knappe: Panofsky: 48 Strauss:
Beobachtungen Diskussionen:
IKONOGRAFIE: Zottelhaube auch bei Protagonisten der Terenz-Holzschnitte, insb. im Phormio |
TECHNIK: "Es ist auf einem sehr dünnen, großen Brett ausserordentlich schön gemalt" (= Heller 1827, S. 176; = bezieht sich unwissentlich auf die KOPIE in Leipzig!) |
These "auf Pergament": Gemälde war ursprünglich auf "ein großes Pergamentblatt gemalt" und wegen "großer Schadhaftigkeit" in den "1840er Jahren" von Erasmus Engerth in Wien "vom Pergament abgelöst und auf eine feine Leinwand übertragen, die wiederum auf eine stärkere Spannleinwand aufgezogen ist. Dabei ist das Bild gründlich restauriert worden. Blos der untere Teil mit den Händen zeigt noch die ursprüngliche Malweise, breit und flüssig bei kräftiger Vorzeichnung" (Thausing I, 132, Anzelewsky, 1990, S. 124).
BEWERTUNG: Tietzes 1928, S. 293
Literaturnotizen:
NEU: Stumpel & van Kregten 2002

Bildname
Louvre Selbsbildnis offizielle Ab...
Louvre Selbsbildnis_Detail1_2.j...
Louvre Selbsbildnis_Detail2_1.j...

Kommentar zu Bild: c bpk, Berlin / bezogen 2008

Record-ID: 11 Last Update: 26.05.2009 - 09:47h User: hohmann

Let us see whether this idea has merit.

Using the Web for the Cultural Heritage

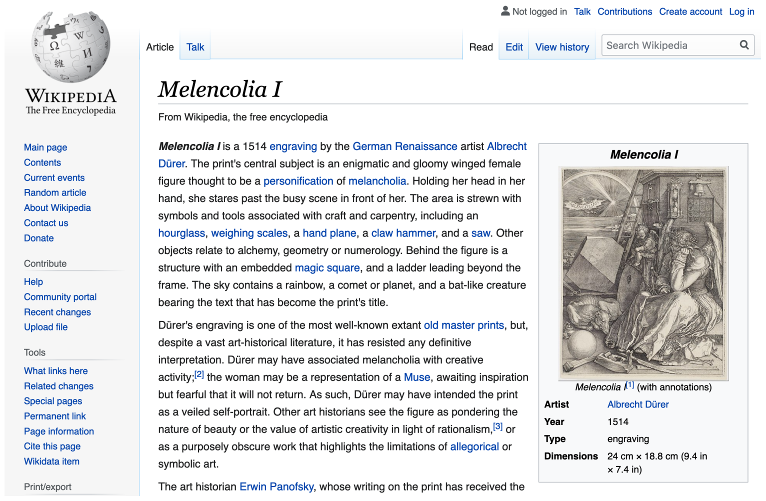
▷ **Idea:** Why not use the world wide web as a tool?

- ▷ it is inherently distributed and networked,
- ▷ the data formats HTML and XML are highly flexible,
- ▷ gives us instantaneous access to information/images/. . . ,
- ▷ allows collaboration and discussion. (wikis, fora, blogs)

Again, an example is in order to help understand the issues at hand.

Cultural Artefacts on the Web

▷ **Example 12.2.5.** A text about a cultural artefact (an etching by Dürer)



Melencolia I

From Wikipedia, the free encyclopedia

Melencolia I is a 1514 engraving by the German Renaissance artist Albrecht Dürer. The print's central subject is an enigmatic and gloomy winged female figure thought to be a personification of melancholia. Holding her head in her hand, she stares past the busy scene in front of her. The area is strewn with symbols and tools associated with craft and carpentry, including an hourglass, weighing scales, a hand plane, a claw hammer, and a saw. Other objects relate to alchemy, geometry or numerology. Behind the figure is a structure with an embedded magic square, and a ladder leading beyond the frame. The sky contains a rainbow, a comet or planet, and a bat-like creature bearing the text that has become the print's title.

Dürer's engraving is one of the most well-known extant old master prints, but, despite a vast art-historical literature, it has resisted any definitive interpretation. Dürer may have associated melancholia with creative activity;[2] the woman may be a representation of a Muse, awaiting inspiration but fearful that it will not return. As such, Dürer may have intended the print as a veiled self-portrait. Other art historians see the figure as pondering the nature of beauty or the value of artistic creativity in light of rationalism,[3] or as a purposely obscure work that highlights the limitations of allegorical or symbolic art.

The art historian Erwin Panofsky, whose writing on the print has received the

Melencolia I

Melencolia I[1] (with annotations)

Artist	Albrecht Dürer
Year	1514
Type	engraving
Dimensions	24 cm × 18.8 cm (9.4 in × 7.4 in)

▷ **Question:** Just how does the etching discussed here relate to Albrecht Dürer?

We collect the properties of the various approaches to documenting cultural artefacts to see how to proceed.

Using the Web for Cultural Heritage

▷ **Problems:** with using the Web as a resource

- ▷ Information is often of dubious quality (imprecise, typos, incomplete, . . .)
- ▷ Information is primarily written for human consumption
  - ▷ $\rightsquigarrow$ not machine-actionable, but full text search works (e.g. Google)
  - ▷ sometimes we can use established structures (e.g. Infobox in Wikipedia)

▷ **Evaluation:** The web is complementary to databases on the structure-vs-flexibility tradeoff scale for cultural heritage systems. (we need both)

▷ **Idea:** Use the semantic web for cultural heritage

- ▷ **Goal**: Make information accessible for humans and machines
- ▷ meaning capture by reference to real-world objects
- ▷ globally unique identifiers of cultural artefacts ($\widehat{=}$ URIs)
- ▷ inference (get out more than you put in!)

12.3 The Semantic Web

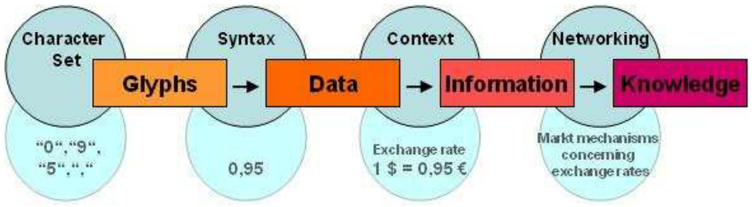
In this section we will introduce the semantic web. That tries to transform the World Wide Web from a human understandable web of multimedia documents into a "web of machine understandable data". In this context, "machine-understandable" means that machines can draw inferences from data they have access to, so that they can make use of the knowledge that is implicit – i.e. not explicitly stated, but can be derived from other information (by humans) – in the web.
We will now define the term semantic web and discuss the pertinent ideas involved. There are two central ones, we will cover here:

- Information and data come in different levels of explicitness; this is usually visualized by a "ladder" of information.
- if information is sufficiently machine-understandable, then we can automate drawing conclusions.

The Semantic Web

▷ **Definition 12.3.1.** The semantic web is the result including of semantic content in web pages with the aim of converting the WWW into a machine-understandable "web of data", where inference based services can add value to the ecosystem.

▷ **Idea:** Move web content up the ladder, use inference to make connections.



▷ **Example 12.3.2.** Information not explicitly represented (in one place)

Query: *Who was US president when Barak Obama was born?*

Google: *. . . BIRTH DATE: August 04, 1961. . .*

Query: *Who was US president in 1961?*

Google: *President: Dwight D. Eisenhower [. . .] John F. Kennedy (starting Jan. 20.)*

Humans understand the text and combine the information to get the answer. Machines need more than just text $\rightsquigarrow$ semantic web technology.

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 351 2025-02-06

The term "semantic web" was coined by Tim Berners Lee in analogy to semantic networks, only applied to the world wide web. And as for semantic networks, where we have inference processes that allow us the recover information that is not explicitly represented from the network (here the world-wide-web).

To see that problems have to be solved, to arrive at the semantic web, we will now look at a concrete example about the "semantics" in web pages. Here is one that looks typical enough.

What is the Information a User sees?

▷ **Example 12.3.3.** Take the following web-site with a conference announcement

WWW2002
The eleventh International World Wide Web Conference
Sheraton Waikiki Hotel
Honolulu, Hawaii, USA
7-11 May 2002

Registered participants coming from
Australia, Canada, Chile Denmark, France, Germany, Ghana, Hong Kong, India,
Ireland, Italy, Japan, Malta, New Zealand, The Netherlands, Norway,
Singapore, Switzerland, the United Kingdom, the United States, Vietnam, Zaire

On the 7th May Honolulu will provide the backdrop of the eleventh
International World Wide Web Conference.

Speakers confirmed
Tim Berners-Lee: Tim is the well known inventor of the Web,
Ian Foster: Ian is the pioneer of the Grid, the next generation internet.

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 352 2025-02-06

But as for semantic networks, what you as a human can see ("understand" really) is deceptive, so let us obfuscate the document to confuse your "semantic processor". This gives an impression of what the computer "sees".

What the machine sees

▷ **Example 12.3.4.** Here is what the machine "sees" from the conference announcement:

𝒲𝒲𝒲∈′′∈
𝒯⟨]]↕]⊑]\⊔⟨ℐ\⊔]∇\⊣⊔⟩≀\⊣↕𝒲≀∇↕[𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\⌋]
𝒮⟨]∇⊣⊔≀\𝒲⊣⟩∥⟩∥⟩ℋ≀⊔]↕
ℋ≀\≀↕⊓↕⊓⇔ℋ⊣⊒⊣⟩⟩⇔𝒰𝒮𝒜
↖∞∞ℳ⊣†∈′′∈

ℛ]}⟩∫⊔]∇]⌈√⊣∇⊔⟩⌋⟩√⊣\⊔∫]≀⇕⟩\}{∇≀⇕
𝒜⊓∫⊔∇⊣↕⟩⊣⇔𝒞⊣\⊣⌈⊣⇔𝒞⟨⟩↕]𝒟]\⇕⊣∇∥⇔ℱ∇⊣\⌋]⇔𝒢]∇⇕⊣\†⇔𝒢⟨⊣\⊣⇔ℋ≀\}𝒦≀\}⇔ℐ\⌈⟩⊣⇔

ℐ∇]↕⊣\⌈⇔ℐ⊔⊣↕†⇔𝒥⊣√⊣\⇔ℳ⊣↕⊔⊣⇔𝒩]⊒𝒵]⊣↕⊣\⌈⇔𝒯⟨]𝒩]⊔⟨]∇↕⊣\⌈∫⇔𝒩≀∇⊒⊣†⇔
𝒮⟩\}⊣√≀∇]⇔𝒮⊒⟩⊔‡]∇↕⊣\⌈⇔⊔⟨]𝒰\⟩⊔]⌈𝒦⟩\}⌈≀↕⇔⊔⟨]𝒰\⟩⊔]⌈𝒮⊔⊣⊔]∫⇔𝒱⟩]⊔\⊣↕⇔𝒵⊣⟩∇]

𝒪\⊔⟨]⊔⟨ℳ⊣†ℋ≀\≀↕⊓↕⊓⊒⟩↕↕√∇≀⊑⟩⌈]⊔⟨]⌊⊣⌋|⌈∇≀√≀{⊔⟨]]↕]⊑]\⊔⟨
ℐ\⊔]∇\⊣⊔⟩≀\⊣↕𝒲≀∇↕⌈𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\⌋]↙

𝒮√]⊣||]∇∫⌋≀\{⟩∇↕]⌈
𝒯⟩↕ℬ]∇\]∇∫↖ℒ]]¬𝒯⟩↕⟩∫⊔⟨]⊒]↕↕||\≀⊒\⟩\⊑]\⊔≀∇≀{⊔⟨]𝒲]⌊⇔
ℐ⊣\ℱ≀∫⊔]∇¬ℐ⊣\⟩∫⊔⟨]√⟩≀\]]∇≀{⊔⟨]𝒢∇⟩⌈⇔⊔⟨]\]§⊔}]\]∇⊣⊔⟩≀\⟩\⊔]∇\]⊔↙

Obviously, there is not much the computer understands, and as a consequence, there is not a lot the computer can support the reader with. So we have to "help" the computer by providing some meaning. Conventional wisdom is that we add some semantic/functional markup. Here we pick XML without loss of generality, and characterize some fragments of text e.g. as dates.

Solution: XML markup with "meaningful" Tags

▷ **Example 12.3.5.** Let's annotate (parts of) the meaning via XML markup

```
<title>𝒲𝒲𝒲∈′′∈
𝒯⟨]]↕]⊑]\⊔⟨ℐ\⊔]∇\⊣⊔⟩≀\⊣↕𝒲≀∇↕⌈𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\⌋]</title>
<place>𝒮⟨]∇⊣⊔≀\𝒲⊣⟩||⟩||⟩ℋ≀⊔]↕ℋ≀\≀↕⊓↕⊓⇔ℋ⊣⊒⊣⟩⟩⇔𝒰𝒮𝒜</place>
<date>↖∞∞ℳ⊣†∈′′∈</date>
<participants>ℛ]}⟩∫⊔]∇]⌈√⊣∇⊔⟩⌋⟩√⊣\⊔∫≀↕⟩\}∇≀↕
𝒜⊓∫⊔∇⊣↕⟩⊣⇔𝒞⊣\⊣⌈⊣⇔𝒞⟨⟩↕]𝒟]\↕⊣∇||⇔ℱ∇⊣\⌋]⇔𝒢]∇↕⊣\†⇔𝒢⟨⊣\⊣⇔ℋ≀\}𝒦≀\}⇔ℐ\⌈⟩⊣⇔
ℐ∇]↕⊣\⌈⇔ℐ⊔⊣↕†⇔𝒥⊣√⊣\⇔ℳ⊣↕⊔⊣⇔𝒩]⊒𝒵]⊣↕⊣\⌈⇔𝒯⟨]𝒩]⊔⟨]∇↕⊣\⌈∫⇔𝒩≀∇⊒⊣†⇔
𝒮⟩\}⊣√≀∇]⇔𝒮⊒⟩⊔‡]∇↕⊣\⌈⇔⊔⟨]𝒰\⟩⊔]⌈𝒦⟩\}⌈≀↕⇔⊔⟨]𝒰\⟩⊔]⌈𝒮⊔⊣⊔]∫⇔𝒱⟩]⊔\⊣↕⇔𝒵⊣⟩∇]
</participants>
<introduction>𝒪\⊔⟨]⊔⟨ℳ⊣†ℋ≀\≀↕⊓↕⊓⊒⟩↕↕√∇≀⊑⟩⌈]⌈]⊔⟨]⌊⊣⌋|⌈∇≀√≀{⊔⟨]]↕]⊑]\⊔⟨ℐ\⊔]∇↖
\⊣⊔⟩≀\⊣↕𝒲≀∇↕⌈𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\⌋]↙</introduction>
<program>𝒮√]⊣||]∇∫⌋≀\{⟩∇↕]⌈
<speaker>𝒯⟩↕ℬ]∇\]∇∫↖ℒ]]¬𝒯⟩↕⟩∫⊔⟨]⊒]↕↕||\≀⊒\⟩\⊑]\⊔≀∇≀{⊔⟨]𝒲]⌊</speaker>
<speaker>ℐ⊣\ℱ≀∫⊔]∇¬ℐ⊣\⟩∫⊔⟨]√⟩≀\]]∇≀{⊔⟨]𝒢∇⟩⌈⇔⊔⟨]\]§⊔}]\]∇⊣⊔⟩≀\⟩\⊔]∇\]⊔<speaker>
</program>
```

But does this really help? Is conventional wisdom correct?

What can we do with this?

▷ **Example 12.3.6.** Consider the following fragments:

ℜ⊔⟩⊔↕]⊤𝒲𝒲𝒲∈′′∈
𝒯⟨]]↕]⊑]\⊔⟨ℐ\⊔]∇\⊣⊔⟩≀\⊣↕𝒲≀∇↕⌈𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\⌋]ℜ∝⊔⟩⊔↕]⊤
ℜ√↕⊣⌋]⊤𝒮⟨]∇⊣⊔≀\𝒲⊣⟩||⟩||⟩ℋ≀⊔]↕ℋ≀\≀↕⊓↕⊓⇔ℋ⊣⊒⊣⟩⟩⇔𝒰𝒮𝒜ℜ∝√↕⊣⌋]⊤
ℜ⌈⊣⊔]⊤↖∞∞ℳ⊣†∈′′∈ℜ∝⌈⊣⊔]⊤

Given the markup above, a machine agent can

- ▷ parse ∞∞M⊣†∈′′∈ as the date May 7 11 2002 and add this to the user's calendar,
- ▷ parse S⟨]∇⊣⊔≀\W⊣⟩∥⟩∥⟩H≀⊔]↕H≀\≀↕⊓↕⊓⇔H⊣⊒⊣⟩⟩⇔USA as a destination and find flights.

▷ **But:** do not be deceived by your ability to understand English!

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 355 2025-02-06

To understand what a machine can understand we have to obfuscate the markup as well, since it does not carry any intrinsic meaning to the machine either.

What the machine sees of the XML

▷ **Example 12.3.7.** Here is what the machine sees of the XML

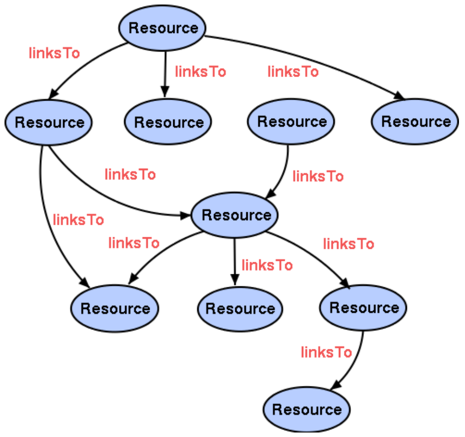
```
<title>WWW∈′′∈
T⟨]]↕]⊑]\⊔⟨I\⊔]∇\⊣⊔⟩≀\⊣↕W≀∇↕[W⟩⌈]W]⌊C≀\{]∇]\⌋]</⊔⟩⊔↕]>
<√↕⊣⌋]>S⟨]∇⊣⊔≀\W⊣⟩∥⟩∥⟩H≀⊔]↕H≀\≀↕⊓↕⊓⇔H⊣⊒⊣⟩⟩⇔USA</√↕⊣⌋]>
<⌈⊣⊔]>↖∞∞M⊣†∈′′∈</⌈⊣⊔]>
<√⊣∇⊔⟩⌋⟩√⊣\⊔∫>R]}⟩∫⊔]∇]⌈√⊣∇⊔⟩⌋⟩√⊣\⊔∫⌋≀⇕⟩\}{∇≀⇕
A⊓∫⊔∇⊣⇕⟩⊣⇔C⊣\⊣⌈⊣⇔C⟨⟩↕]D]\⇕⊣∇∥⇔F∇⊣\⌋]⇔G]∇⇕⊣\†⇔G⟨⊣\⊣⇔H≀\}K≀\}⇔I\⌈⟩⊣⇔
I∇]↕⊣\⌈⇔I⊔⊣⇕†⇔J⊣√⊣\⇔M⊣⇕⊔⊣⇔N]⊒Z]⊣↕⊣\⌈⇔T⟨]N]⊔⟨]∇↕⊣\⌈∫⇔N≀∇⊒⊣†⇔
S⟩\}⊣√≀∇]⇔S⊒⟩⊔‡]∇↕⊣\⌈⇔⊔⟨]U\⟩⊔]⌈K⟩\}⌈≀⇕⇔⊔⟨]U\⟩⊔]⌈S⊔⊣⊔]∫⇔V⟩]⊔\⊣⇕⇔Z⊣⟩∇]
</√⊣∇⊔⟩⌋⟩√⊣\⊔∫>
<⟩\⊔∇≀⌈⊓⌋⊔⟩≀\>O\⊔⟨]⊔⟨M⊣†H≀\≀↕⊓↕⊓⊒⟩↕↕√∇≀⊑⟩⌈]⊔⟨]⌊⊣⌋∥⌈∇≀√≀{⊔⟨]]↕]⊑]\⊔⟨I\⊔]∇\⊣↖
⊔⟩≀\⊣↕W≀∇↕[W⟩⌈]W]⌊C≀\{]∇]\⌋]↙</⟩\⊔∇≀⌈⊓⌋⊔⟩≀\>
<√∇≀}∇⊣⇕>S√]⊣∥]∇∫⌋≀\{⟩∇⇕]⌈
<∫√]⊣∥]∇>T⟩⇕B]∇\]∇∫↖L]]¬T⟩⇕⟩∫⊔⟨]⊒]↕↕∥\≀⊒\⟩\⊑]\⊔≀∇≀{⊔⟨]W]⌊</∫√]⊣∥]∇>
<∫√]⊣∥]∇>I⊣\F≀∫⊔]∇¬I⊣\⟩∫⊔⟨]√⟩≀\]]∇≀{⊔⟨]G∇⟩⌈⇔⊔⟨]\]§⊔}]\]∇⊣⊔⟩≀\⟩\⊔]∇\]⊔<∫√]⊣∥]∇>
</√∇≀}∇⊣⇕>
```

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 356 2025-02-06

So we have not really gained much either with the markup, we really have to give meaning to the markup as well, this is where techniques from semenatic web come into play.
To understand how we can make the web more semantic, let us first take stock of the current status of (markup on) the web. It is well-known that world-wide-web is a hypertext, where multimedia documents (text, images, videos, etc. and their fragments) are connected by hyperlinks. As we have seen, all of these are largely opaque (non-understandable), so we end up with the following situation (from the viewpoint of a machine).

The Current Web

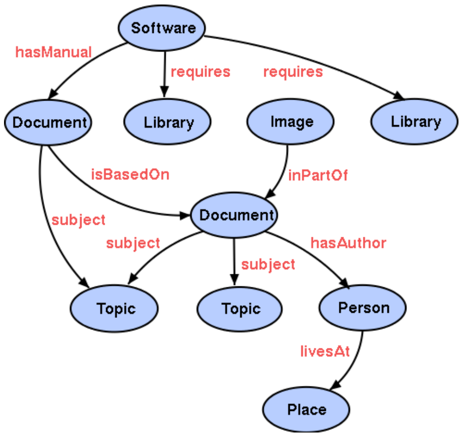
- ▷ **Resources:** identified by URIs, untyped
- ▷ **Links:** href, src, . . . limited, non-descriptive
- ▷ **User:** Exciting world - semantics of the resource, however, gleaned from content
- ▷ **Machine:** Very little information available - significance of the links only evident from the context around the anchor.



Let us now contrast this with the envisioned semantic web.

The Semantic Web

- ▷ **Resources:** Globally identified by URIs or Locally scoped (Blank), Extensible, Relational.
- ▷ **Links:** Identified by URIs, Extensible, Relational.
- ▷ **User:** Even more exciting world, richer user experience.
- ▷ **Machine:** More processable information is available (Data Web).
- ▷ **Computers and people:** Work, learn and exchange knowledge effectively.



Essentially, to make the web more machine-processable, we need to classify the resources by the concepts they represent and give the links a meaning in a way, that we can do inference with that. The ideas presented here gave rise to a set of technologies jointly called the “semantic web”, which we will now summarize before we return to our logical investigations of knowledge representation techniques.

Towards a “Machine-Actionable Web”

- ▷ **Recall:** We need external agreement on meaning of annotation tags.
- ▷ **Idea:** standardize them in a community process (e.g. DIN or ISO)
- ▷ **Problem:** Inflexible, Limited number of things can be expressed
- ▷ **Better:** Use ontologies to specify meaning of annotations

- ▹ Ontologies provide a vocabulary of terms
- ▹ New terms can be formed by combining existing ones
- ▹ Meaning (semantics) of such terms is formally specified
- ▹ Can also specify relationships between terms in multiple ontologies

▷ Inference with annotations and ontologies (get out more than you put in!)

- ▹ Standardize annotations in RDF [KC04] or RDFa [Her+13b] and ontologies on OWL [OWL09]
- ▹ Harvest RDF and RDFa in to a triplestore or OWL reasoner.
- ▹ Query that for implied knowledge (e.g. chaining multiple facts from Wikipedia)

  SPARQL: Who was US President when Barack Obama was Born?

  DBPedia: John F. Kennedy (was president in August 1961)

12.4 Semantic Networks and Ontologies

To get a feeling for ontologies and how they enable the "machine-actionable web" and how that helps us in DH, we take a look at "semantic networks", which are an early form of ontologies. They allow us to explain many of the basic functionalities of the "semantic web" without getting too much into details of the technologies involved. We will preview that at the end of this section and go into details **??**.

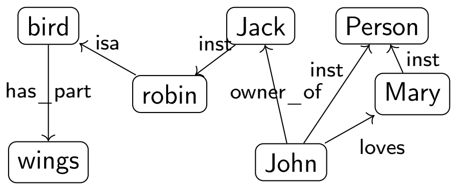
Semantic networks are a very simple way of arranging knowledge about objects and concepts and their relationships in a graph.

Semantic Networks [CQ69]

▷ **Definition 12.4.1.** A semantic network is a directed graph for representing knowledge:

- ▹ nodes represent objects and concepts (classes of objects) (e.g. John (object) and bird (concept))
- ▹ edges (called links) represent relations between these (isa, father_of, belongs_to)

▷ **Example 12.4.2.** A semantic network for birds and persons:



▷ **Problem:** How do we derive new information from such a network?

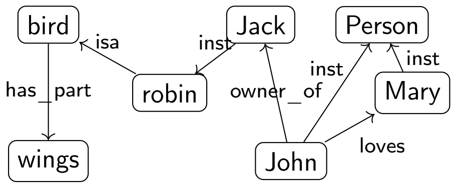
▷ **Idea:** Encode taxonomic information about objects and concepts in special links ("isa" and "inst") and specify property inheritance along them in the process model.

Even though the network in ?? is very intuitive (we immediately understand the concepts depicted), it is unclear how we (and more importantly a machine that does not associate meaning with the labels of the nodes and edges) can draw inferences from the “knowledge” represented.

Deriving Knowledge Implicit in Semantic Networks

▷ **Observation 12.4.3.** *There is more knowledge in a semantic network than is explicitly written down.*

▷ **Example 12.4.4.** In the network below, we “know” that *robins have wings* and in particular, *Jack has wings*.



▷ **Idea:** Links labeled with “isa” and “inst” are special: they propagate properties encoded by other links.

▷ **Definition 12.4.5.** We call links labeled by

- ▹ “isa” an inclusion or isa link (inclusion of concepts)
- ▹ “inst” instance or inst link (concept membership)

We now make the idea of “propagating properties” rigorous by defining the notion of derived relations, i.e. the relations that are left implicit in the network, but can be added without changing its meaning.

Deriving Knowledge Semantic Networks

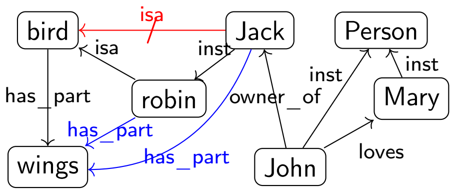
▷ **Definition 12.4.6 (Inference in Semantic Networks).** We call all link labels except “inst” and “isa” in a semantic network relations.

Let N be a semantic network and R a relation in N such that $A \xrightarrow{\text{isa}} B \xrightarrow{R} C$ or $A \xrightarrow{\text{inst}} B \xrightarrow{R} C$, then we can derive a relation $A \xrightarrow{R} C$ in N.

The process of deriving new concepts and relations from existing ones is called inference and concepts/relations that are only available via inference implicit (in a semantic network).

▷ **Intuition:** Derived relations represent knowledge that is implicit in the network; they could be added, but usually are not to avoid clutter.

▷ **Example 12.4.7.** Derived relations in ??



▷ **Slogan:** Get out more knowledge from a semantic networks than you put in.

Note that **??** does not quite allow to derive that *Jack is a bird* (did you spot that "isa" is not a relation that can be inferred?), even though we know it is true in the world. This shows us that inference in semantic networks has be to very carefully defined and may not be "complete", i.e. there are things that are true in the real world that our inference procedure does not capture.

Dually, if we are not careful, then the inference procedure might derive properties that are not true in the real world even if all the properties explicitly put into the network are. We call such an inference procedure unsound or incorrect.

These are two general phenomena we have to keep an eye on.

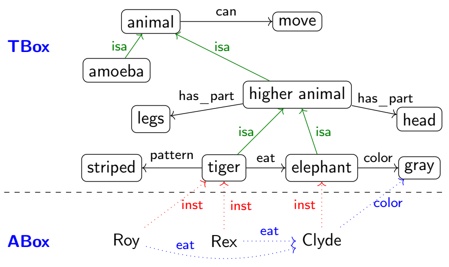
Another problem is that semantic networks (e.g. in **??**) confuse two kinds of concepts: individuals (represented by proper names like *John* and *Jack*) and concepts (nouns like *robin* and *bird*). Even though the isa and inst link already acknowledge this distinction, the "has_part" and "loves" relations are at different levels entirely, but not distinguished in the networks.

Terminologies and Assertions

▷ *Remark 12.4.8.* We should distinguish concepts from objects.

▷ **Definition 12.4.9.** We call the subgraph of a semantic network N spanned by the isa links and relations between concepts the terminology (or TBox, or the famous Isa Hierarchy) and the subgraph spanned by the inst links and relations between objects, the assertions (together the ABox) of N.

▷ **Example 12.4.10.** In this semantic network we keep objects concept apart notationally:



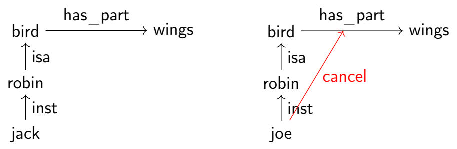
In particular we have objects "Rex", "Roy", and "Clyde", which have (derived) relations (e.g. *Clyde* is *gray*).

But there are severe shortcomings of semantic networks: the suggestive shape and node names give (humans) a false sense of meaning, and the inference rules are only given in the process model (the implementation of the semantic network processing system).

This makes it very difficult to assess the strength of the inference system and make assertions e.g. about completeness.

Limitations of Semantic Networks

- ▷ What is the meaning of a link?
  - ▷ link labels are very suggestive (misleading for humans)
  - ▷ meaning of link types defined in the process model (no denotational semantics)
- ▷ **Problem:** No distinction of optional and defining traits!
- ▷ **Example 12.4.11.** Consider a robin that has lost its wings in an accident:

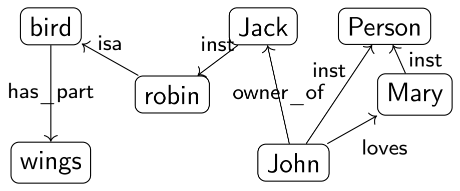


"Cancel-links" have been proposed, but their status and process model are debatable.

To alleviate the perceived drawbacks of semantic networks, we can contemplate another notation that is more linear and thus more easily implemented: function/argument notation.

Another Notation for Semantic Networks

- ▷ **Definition 12.4.12.** Function/argument notation for semantic networks
  - ▷ interprets nodes as arguments (reification to individuals)
  - ▷ interprets links as functions (predicates actually)
- ▷ **Example 12.4.13.**



isa(robin,bird)
haspart(bird,wings)
inst(Jack,robin)
owner_of(John, robin)
loves(John,Mary)

- ▷ **Evaluation:**
  - + linear notation (equivalent, but better to implement on a computer)
  - + easy to give process model by deduction (e.g. in Prolog)
  - − worse locality properties (networks are associative)

Indeed the function/argument notation is the immediate idea how one would naturally represent semantic networks for implementation.

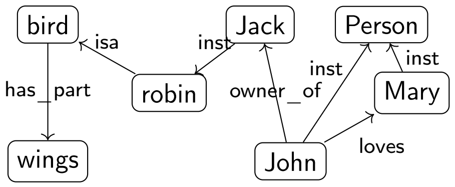
This notation has been also characterized as subject/predicate/object triples, alluding to simple (English) sentences. This will play a role in the "semantic web" later. The next slide is a bit outside of the scope of IWGS, but we want to go into this anyway.

We have been talking about the "procedural model" of a semantic network, which essentially specifies the inference algorithm that derives new knowledge in a network. There is an alternative to this: we can map the network language – function/argument notation for networks is an essential step for this – into a known language with an inference system. We call this kind of a mapping a "denotational semantics", here into a language called first-order logic.

Building on the function/argument notation from above, we can now give a formal semantics for semantic network: we translate them into first-order logic and use the semantics of that.

A Denotational Semantics for Semantic Networks

- **Observation:** If we handle isa and inst links specially in function/argument notation



robin $\subseteq$ bird
haspart(bird,wings)
Jack $\in$ robin
owner_of(John, Jack)
loves(John,Mary)

it looks like first-order logic, if we take

- $a \in S$ to mean $S(a)$ for an object a and a concept S.
- $A \subseteq B$ to mean $\forall X.A(X) \Rightarrow B(X)$ and concepts A and B
- $R(A, B)$ to mean $\forall X.A(X) \Rightarrow (\exists Y.B(Y) \wedge R(X, Y))$ for a relation R.

- **Idea:** Take first-order deduction as process model (gives inheritance for free)

Indeed, the semantics induced by the translation to first-order logic, gives the intuitive meaning to the semantic networks. Note that this only holds only for the features of semantic networks that are representable in this way, e.g. the "cancel links" shown above are not (and that is a feature, not a bug).

But even more importantly, the translation to first-order logic gives a first process model: we can use first-order inference to compute the set of inferences that can be drawn from a semantic network.

Based on the intuitions from semantic networks we can now come to general (semantic web) ontologies.

What is an Ontology

- **Definition 12.4.14.** An ontology is a formal model of (an aspect of) the world. It
  - introduces a vocabulary for the objects, concepts, and relations of a given domain,
  - specifies intended meaning of vocabulary in a description logic using
    - a set of axioms describing structure of the model
    - a set of facts describing some particular concrete situation

  The vocabulary together with the collection of axioms is often called a terminology (or TBox) and the collection of facts an ABox (assertions).

In addition to the represented axioms and facts, the description logic determines a number of derived ones.

▷ **Definition 12.4.15.** A vocabulary often includes names for classes and relationship (also called concepts, and properties).

▷ *Remark 12.4.16.* If the description logic has a reasoner, we can automatically

- ▷ detect inconsistent axiom systems
- ▷ compute class membership and taxonomies.

There is a whole collection of standardized languages and interoperable systems that facilitate dealing with (very large) ontologies in practice. We will only give a summary preview here, leaving the detailed discussion to ??.

Semantic Web Technology in a Nutshell

▷ Ontologies have become one of the standard devices for representing information about the Web and the world.

▷ **Definition 12.4.17.** This is facilitated and standardized by the semantic web technology stack:

- ▷ URIs for representing objects,
- ▷ RDF triples for representing facts,
- ▷ RDFa for annotating RDF triples in XML documents,
- ▷ OWL for representing TBoxes,
- ▷ triplestores for storing (lots of) RDF triples,
- ▷ SPARQL for querying ontologies,
- ▷ description logic reasoners for deciding ontology consistency and concept subsumption,
- ▷ Protégé for authoring and maintaining ontologies,

▷ Details ??.

Indeed, this list can be read as a technology roadmap for the WissKI system. We have already seen the most of the concepts in ??, we will discuss the technologies ??, but first we will have a look at the CIDOC CRM ontology that is used in WissKI.

12.5 CIDOC CRM: An Ontology for Cultural Heritage

We have seen that databases are not the only choice for representing data about cultural heritage. Indeed, the WissKI system chooses ontologies as a basis for representation and querying.

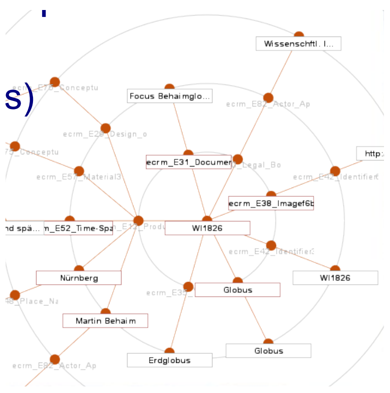
To ensure interoperability, WissKI is based on the ISO-standardized CIDOC CRM ontology, which we will now introduce and explore.

Now, we can instantiate what we have learned about ontology-based information systems to cultural heritage disciplines. We collect all the bits and pieces and hint at the technologies (details ??).

Ontologies for Cultural Artefacts

- ▷ **Idea:** Use ontologies for documenting cultural heritage.
  - ▷ flexible schemata (OWL)
  - ▷ easy data sharing
  - ▷ open standards, free tools
  - ▷ semantic querying via SPARQL
- ▷ **Idea:** We can use RDF like a Mindmap: RDF can
  - ▷ represent relations between objects
  - ▷ classify objects (web resources)

  RDFa for document annotation



- ▷ Reference ontologies for interoperability:
  - ▷ SUMO (Suggested Upper Model Ontology) [SUMO] for common knowledge,
  - ▷ FOAF (Friend-of-a-Friend) [FOAF14] for persons and relations,
  - ▷ CIDOC CRM for documentation of cultural heritage. (up next)

SOME RIGHTS RESERVED

So let us look at the CIDOC CRM ontology in more detail. It has been developed by the Documentation Committee of the ICOM (International Council of Museums) over more than 20 years and has been standardized by the ISO. Even more importantly for our purposes here, the CIDOC CRM has been implemented in the OWL format, which gives us the use of the semantic web technology stack.

CIDOC CRM (Conceptual Reference Model)

- ▷ **Definition 12.5.1.** CIDOC CRM provides an extensible ontology for concepts and information in cultural heritage and museum documentation. It is the international standard (ISO 21127:2014) for the controlled exchange of cultural heritage information. The central classes include
  - ▷ space time specified by title/identifier, place, era/period, time-span, and relationship to persistent items
  - ▷ events specified by title/identifier, beginning/ending of existence, participants (people, either individually or in groups), creation/modification of things (physical or conceptional), and relationship to persistent items
  - ▷ material things specified by title/identifier, place, the information object the material thing carries, part-of relationships, and relationship to persistent items
  - ▷ immaterial things specified by title/identifier, information objects (propositional or symbolic), conceptional things, and part-of relationships
- ▷ **Definition 12.5.2.** OWL implements CIDOC CRM in OWL

▷ Details about CIDOC CRM can be found at [CC] and about OWL at [ECRMb; ECRMa].

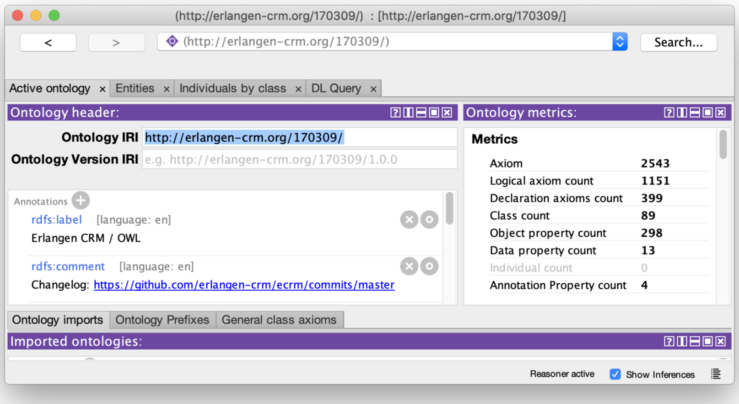
One of the advantages of having CIDOC CRM in OWL is that we can use semantic web technologies to deal with it. Here we use one of the practically most important tools: Protégé.

Protégé, an IDE for Ontology Development

▷ **Definition 12.5.3.** Protégé [Pro] is an integrated development environment for ontologies represented in the OWL family. It comprises

- ▷ a visual user interface for exploring and editing ontologies,
- ▷ a inference component to ensure ontology consistency and minimality,
- ▷ a facility for querying the loaded ontologies.

▷ **Example 12.5.4 (CIDOC CRM in Protégé).**


(http://erlangen-crm.org/170309/) : [http://erlangen-crm.org/170309/]

Ontology IRI http://erlangen-crm.org/170309/
Ontology Version IRI e.g. http://erlangen-crm.org/170309/1.0.0

Annotations
rdfs:label [language: en]
Erlangen CRM / OWL

rdfs:comment [language: en]
Changelog: https://github.com/erlangen-crm/ecrm/commits/master

Metrics

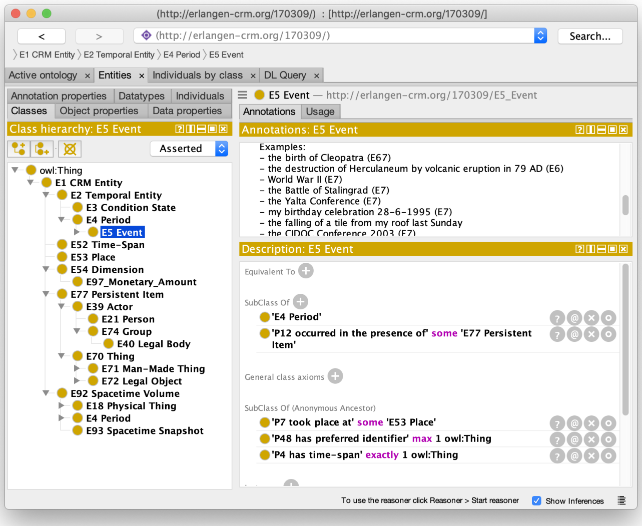
Axiom	2543
Logical axiom count	1151
Declaration axioms count	399
Class count	89
Object property count	298
Data property count	13
Individual count	0
Annotation Property count	4

The backbone of the CIDOC CRM ontology is formed by the concepts (called “classes” in OWL). They form an inheritance hierarchy – of which the top part is shown on the left of the Protégé window below. The ontology provides – usually relatively abstract classes for all objects related to cultural artefacts, their properties, and provenance.

CIDOC CRM Explored (Classes)

▷ **Idea:** Use semantic web technology to explore OWL.

▷ **CIDOC CRM Classes:** concept $\widehat{=}$ OWL “Class” (shown in Protege)



(http://erlangen-crm.org/170309/) : [http://erlangen-crm.org/170309/]

(http://erlangen-crm.org/170309/)

E1 CRM Entity 〉 E2 Temporal Entity 〉 E4 Period 〉 E5 Event

E5 Event — http://erlangen-crm.org/170309/E5_Event

Class hierarchy: E5 Event

```
owl:Thing
  E1 CRM Entity
    E2 Temporal Entity
      E3 Condition State
      E4 Period
        E5 Event
    E52 Time-Span
    E53 Place
    E54 Dimension
      E97_Monetary_Amount
    E77 Persistent Item
      E39 Actor
        E21 Person
        E74 Group
          E40 Legal Body
      E70 Thing
        E71 Man-Made Thing
        E72 Legal Object
    E92 Spacetime Volume
      E18 Physical Thing
      E4 Period
      E93 Spacetime Snapshot
```

Annotations: E5 Event

Examples:
- the birth of Cleopatra (E67)
- the destruction of Herculaneum by volcanic eruption in 79 AD (E6)
- World War II (E7)
- the Battle of Stalingrad (E7)
- the Yalta Conference (E7)
- my birthday celebration 28-6-1995 (E7)
- the falling of a tile from my roof last Sunday
- the CIDOC Conference 2003 (E7)

Description: E5 Event

Equivalent To

SubClass Of
- 'E4 Period'
- 'P12 occurred in the presence of' some 'E77 Persistent Item'

General class axioms

SubClass Of (Anonymous Ancestor)
- 'P7 took place at' some 'E53 Place'
- 'P48 has preferred identifier' max 1 owl:Thing
- 'P4 has time-span' exactly 1 owl:Thing

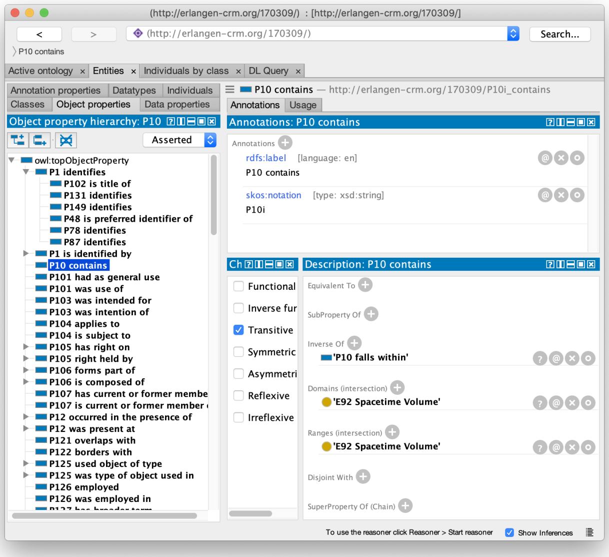
To use the reasoner click Reasoner > Start reasoner ✓ Show Inferences

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 372 2025-02-06

The concepts are complemented by the relations called “object properties” in OWL.

CIDOC CRM Explored (Relations)

▷ **CIDOC CRM Relations:** relation $\widehat{=}$ OWL “Object Property” (shown in Protege)



(http://erlangen-crm.org/170309/) : [http://erlangen-crm.org/170309/]

(http://erlangen-crm.org/170309/)

P10 contains

P10 contains — http://erlangen-crm.org/170309/P10i_contains

Object property hierarchy: P10

```
owl:topObjectProperty
  P1 identifies
    P102 is title of
    P131 identifies
    P149 identifies
    P48 is preferred identifier of
    P78 identifies
    P87 identifies
  P1 is identified by
  P10 contains
  P101 had as general use
  P101 was use of
  P103 was intended for
  P103 was intention of
  P104 applies to
  P104 is subject to
  P105 has right on
  P105 right held by
  P106 forms part of
  P106 is composed of
  P107 has current or former membe
  P107 is current or former member
  P12 occurred in the presence of
  P12 was present at
  P121 overlaps with
  P122 borders with
  P125 used object of type
  P125 was type of object used in
  P126 employed
  P126 was employed in
```

Annotations: P10 contains

Annotations
- rdfs:label [language: en]
  P10 contains
- skos:notation [type: xsd:string]
  P10i

Ch

- [ ] Functional
- [ ] Inverse fur
- [x] Transitive
- [ ] Symmetric
- [ ] Asymmetri
- [ ] Reflexive
- [ ] Irreflexive

Description: P10 contains

Equivalent To

SubProperty Of

Inverse Of
- 'P10 falls within'

Domains (intersection)
- 'E92 Spacetime Volume'

Ranges (intersection)
- 'E92 Spacetime Volume'

Disjoint With

SuperProperty Of (Chain)

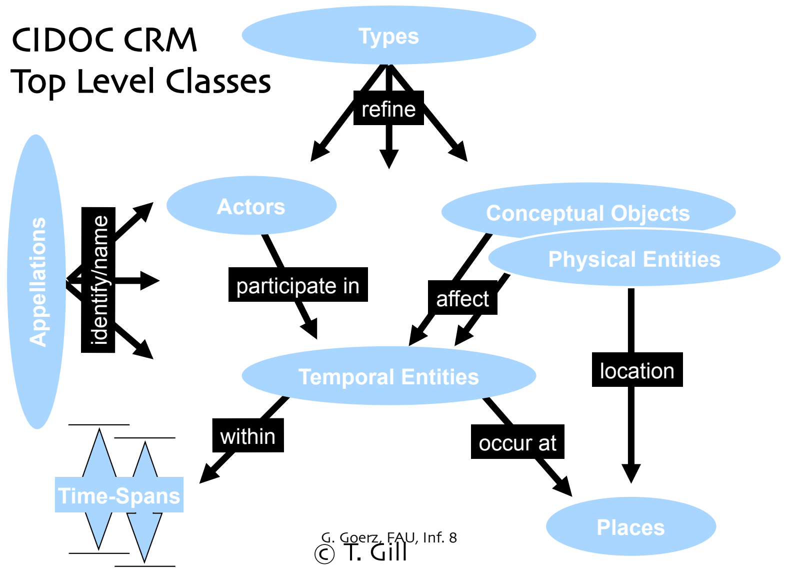
To use the reasoner click Reasoner > Start reasoner ✓ Show Inferences

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 373 2025-02-06

There are also a small number of “data properties”, i.e. properties whose values are concrete data like numbers, dates, or strings. They are less interesting structurally, but important in practice.

We can summarize the structure of the CIDOC CRM ontology in the following diagram.

CIDOC CRM Structure (Overview)



Now that we understand the CIDOC CRM ontology, we look into the process of modeling cultural artefacts.

CIDOC-CRM Modeling

▷ **This is all good and dandy** but how do I concretely model cultural artefacts?

▷ **Answer:** CIDOC CRM is only a TBox, we add an ABox of objects and facts.

▷ **Example 12.5.5.** *Albrecht Dürer painted Melencolia 1 in Nürnberg*
We have two units of information here: (model separately in CIDOC CRM)

1. *Albrecht Dürer painted Melencolia 1*
   CIDOC CRM modeling decisions:
   (a) A painting m is an "Information Carrier" (E84)
   (b) It was created in an "Production Event" q (E12)
   (c) m is related to q via the "was produced by" relation (P108i)
   (d) q was "carried out by" a "person" d (P14 E21)
   (e) d "is identified by" an "actor appellation" a (P131 E82)
   (f) a "has note" the string "Albrecht Dürer". (P3)
2. *this happened in the city of Nürnberg*
   CIDOC CRM modeling decisions:
   (a) A painting m is an "Information Carrier" (E84)
   (b) It was created in an "Production Event" q (E12)
   (c) m is related to q via the " produced by" relation (P108i)
   (d) q "took place at" a "place" p (P7 E53)
   (e) p "is identified by" a "place name" n (P48 E3)

(f) n “has note” the string "Nürnberg". (P3)

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 375 2025-02-06

If we look more closely at the objects and relations **??**, we see that

- a typical information unit results in a whole chain of objects connected by ontology relations
- parts of these chains are shared between information units

We address this now and introduce the concept of ontology groups and ontology paths for that.

CIDOC CRM Modelling (Ontology Paths)

▷ Modeling *Albrecht Dürer painted Melencolia 1 in Nürnberg* in CIDOC CRM

$$\begin{tikzcd}
 & & \boxed{d : E21} \arrow[r, "P131"] & \boxed{a : E82} \arrow[r, "P3"] & \texttt{"A. Dürer"} \\
\boxed{m : E84} \arrow[r, "P108i"] & \boxed{q : E12} \arrow[ur, "P14"] \arrow[dr, "P7"'] & & & \\
 & & \boxed{p : E53} \arrow[r, "P87"] & \boxed{n : E48} \arrow[r, "P3"] & \texttt{"Nürnberg"}
\end{tikzcd}$$

Note that we need to create the intermediary objects q, d, a, and n.

▷ **Problem:** That is a lot of work for something very simple.

▷ **Definition 12.5.6.** We call sequence of facts $s_i \xrightarrow{p_i} o_i$, where $s_i = o_{i-1}$ an ontology path and any subtree an ontology group.

▷ **Problem Reformulated:** A simple statement like *Albrecht Dürer painted Melencolia 1* becomes a whole ontology path in CIDOC CRM.

▷ **But:** we can reuse intermediary objects and facts, and need fine grained models for flexibility.

▷ **Idea:** Maybe systems can take some of the pain out of modeling. ($\rightsquigarrow$ WissKI)

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 376 2025-02-06

In **??**, we have already seen one of the peculiarities of modeling complex situations in ontologies: the use of events as intermediate objects. This is a general phenomenon when modeling with ontologies, which we have to get used. to

Event-Oriented Modeling in CIDOC CRM

▷ **Observation 12.5.7.** *Ontologies make it easy to model facts with transitive verbs, e.g. Albrecht Dürer <u>created</u> Melencolia 1* *(binary relation)*

▷ **Problem:** What about more complex situations with more arguments? E.g.

1. *Albrecht Dürer <u>created</u> Melencolia 1 with an etching needle* (ternary)
2. *Albrecht Dürer <u>created</u> Melencolia 1 with an etching needle in Nürnberg* (four arguments)
3. *Albrecht Dürer <u>created</u> Melencolia 1 with an etching needle in Nürnberg out of boredom* (five)

▷ **Standard Solution:** Introduce "events" tied to the verb and describe those

▷ **Example 12.5.8.** There was a creation event e with

1. *Albrecht Dürer* as the agent,
2. *Melencolia 1* as the product,
3. *an etching needle* as the means,
4. *boredom* as the reason,

▷ **Consequence:** More than 1/3 of CIDOC CRM classes are events of some kind.

This "event-oriented" thinking is unfamiliar at first and takes practice to become natural. As a rule of thumb one should proceed as in the Melencolia example above. We first identify the "participants" in the situation, if these are more than two, we need to introduce an appropriate event (select from the ones provided by CIDOC CRM) and then connect the event to the object currently under consideration, and all the "participants" to the event.

12.6 The Semantic Web Technology Stack

In this section we discuss how we can apply description logics in the real world, in particular, as a conceptual and algorithmic basis of the semantic web. That tries to transform the World Wide Web from a human-understandable web of multimedia documents into a "web of machine-understandable data". In this context, "machine-understandable" means that machines can draw inferences from data they have access to. Note that the discussion in this digression is not a full-blown introduction to RDF and OWL, we leave that to [SR14; Her+13a; Hit+12] and the respective W3C recommendations. Instead we introduce the ideas behind the mappings from a perspective of the description logics we have discussed above.
The most important component of the semantic web is a standardized language that can represent "data" about information on the Web in a machine-oriented way.

Resource Description Framework

▷ **Definition 12.6.1.** The Resource Description Framework (RDF) is a framework for describing resources on the web. It is an XML vocabulary developed by the W3C.

▷ **Note:** RDF is designed to be read and understood by computers, not to be displayed to people. (it shows)

▷ **Example 12.6.2.** RDF can be used for describing (all "objects on the WWW")

- ▷ properties for shopping items, such as price and availability
- ▷ time schedules for web events
- ▷ information about web pages (content, author, created and modified date)
- ▷ content and rating for web pictures
- ▷ content for search engines
- ▷ electronic libraries

Note that all these examples have in common that they are about "objects on the Web", which is

an aspect we will come to now.
"Objects on the Web" are traditionally called "resources", rather than defining them by their intrinsic properties – which would be ambitious and prone to change – we take an external property to define them: everything that has a URI is a web resource. This has repercussions on the design of RDF.

Resources and URIs

- ▷ RDF describes resources with properties and property values.
- ▷ RDF uses Web identifiers (URIs) to identify resources.
- ▷ **Definition 12.6.3.** A resource is anything that can have a URI, such as `http://www.fau.de`.
- ▷ **Definition 12.6.4.** A property is a resource that has a name, such as *author* or *homepage*, and a property value is the value of a property, such as *Michael Kohlhase* or `http://kwarc.info/kohlhase`. (a property value can be another resource)
- ▷ **Definition 12.6.5.** A RDF statement s (also known as a triple) consists of a resource (the subject of s), a property (the predicate of s), and a property value (the object of s). A set of RDF triples is called an RDF graph.
- ▷ **Example 12.6.6.** Statements: *[This slide]^{subj} has been [author]^{pred}ed by [Michael Kohlhase]^{obj}*

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 379 2025-02-06

The crucial observation here is that if we map "subjects" and "objects" to "individuals", and "predicates" to "relations", the RDF triples are just relational ABox statements of description logics. As a consequence, the techniques we developed apply.
Note: Actually, a RDF graph is technically a labeled multigraph, which allows multiple edges between any two nodes (the resources) and where nodes and edges are labeled by URIs.
We now come to the concrete syntax of RDF. This is a relatively conventional XML syntax that combines RDF statements with a common subject into a single "description" of that resource.

XML Syntax for RDF

- ▷ RDF is a concrete XML vocabulary for writing statements
- ▷ **Example 12.6.7.** The following RDF document could describe the slides as a resource

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22−rdf−syntax−ns#"
         xmlns:dc= "http://purl.org/dc/elements/1.1/">
  <rdf:Description about="https://.../CompLog/kr/en/rdf.tex">
    <dc:creator>Michael Kohlhase</dc:creator>
    <dc:source>http://www.w3schools.com/rdf</dc:source>
  </rdf:Description>
</rdf:RDF>
```

This RDF document makes two statements:

- ▷ The subject of both is given in the about attribute of the rdf:Description element

  - ▷ The predicates are given by the element names of its children
  - ▷ The objects are given in the elements as URIs or literal content.

▷ **Intuitively:** RDF is a web-scalable way to write down ABox information.

Note that XML namespaces play a crucial role in using element to encode the predicate URIs. Recall that an element name is a qualified name that consists of a namespace URI and a proper element name (without a colon character). Concatenating them gives a URI in our example the predicate URI induced by the dc:creator element is `http://purl.org/dc/elements/1.1/creator`. Note that as URIs go RDF URIs do not have to be URLs, but this one is and it references (is redirected to) the relevant part of the Dublin Core elements specification [DCM12].
RDF was deliberately designed as a standoff markup format, where URIs are used to annotate web resources by pointing to them, so that it can be used to give information about web resources without having to change them. But this also creates maintenance problems, since web resources may change or be deleted without warning.

RDFa gives authors a way to embed RDF triples into web resources and make keeping RDF statements about them more in sync.

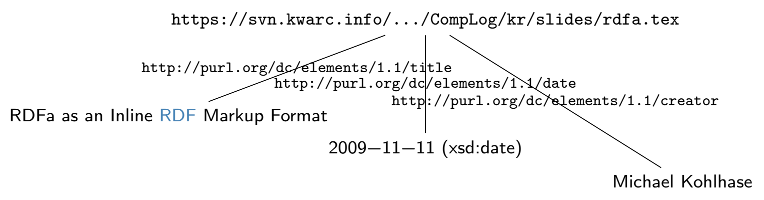
RDFa as an Inline RDF Markup Format

▷ **Problem:** RDF is a standoff markup format (annotate by URIs pointing into other files)

Definition 12.6.8. RDFa (RDF annotations) is a markup scheme for inline annotation (as XML attributes) of RDF triples.

▷ **Example 12.6.9.**

```
<div xmlns:dc="http://purl.org/dc/elements/1.1/" id="address">
  <h2 about="#address" property="dc:title">RDF as an Inline RDF Markup Format</h2>
  <h3 about="#address" property="dc:creator">Michael Kohlhase</h3>
  <em about="#address" property="dc:date" datatype="xsd:date"
      content="2009−11−11">November 11., 2009</em>
</div>
```


https://svn.kwarc.info/.../CompLog/kr/slides/rdfa.tex
http://purl.org/dc/elements/1.1/title
http://purl.org/dc/elements/1.1/date
http://purl.org/dc/elements/1.1/creator
RDFa as an Inline RDF Markup Format
2009−11−11 (xsd:date)
Michael Kohlhase

In the example above, the about and property attributes are reserved by RDFa and specify the subject and predicate of the RDF statement. The object consists of the body of the element, unless otherwise specified e.g. by the content and datatype attributes for literals content.
Let us now come back to the fact that RDF is just an XML syntax for ABox statements.

RDF as an ABox Language for the Semantic Web

▷ **Idea:** RDF triples are ABox entries h R s or $h{:}\varphi$.

▷ **Example 12.6.10.** h is the resource for Ian Horrocks, s is the resource for Ulrike Sattler, R is the relation "hasColleague", and φ is the class foaf:Person

```
<rdf:Description about="some.uri/person/ian_horrocks">
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
  <hasColleague resource="some.uri/person/uli_sattler"/>
</rdf:Description>
```

▷ **Idea:** Now, we need an similar language for TBoxes (based on $\mathcal{ALC}$)

In this situation, we want a standardized representation language for TBox information; OWL does just that: it standardizes a set of knowledge representation primitives and specifies a variety of concrete syntaxes for them. OWL is designed to be compatible with RDF, so that the two together can form an ontology language for the web.

OWL as an Ontology Language for the Semantic Web

▷ **Task:** Complement RDF (ABox) with a TBox language.

▷ **Idea:** Make use of resources that are values in rdf:type. (called Classes)

▷ **Definition 12.6.11.** OWL (the ontology web language) is a language for encoding TBox information about RDF classes.

▷ **Example 12.6.12 (A concept definition for "Mother").** Mother=Woman ⊓ Parent is represented as

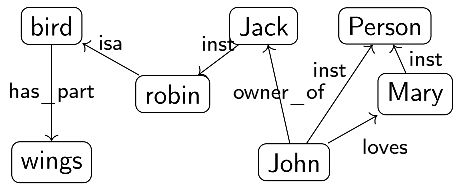
XML Syntax	Functional Syntax

`<EquivalentClasses>`
`  <Class IRI="Mother"/>`
`  <ObjectIntersectionOf>`
`    <Class IRI="Woman"/>`
`    <Class IRI="Parent"/>`
`  </ObjectIntersectionOf>`
`</EquivalentClasses>`
	`EquivalentClasses(`
	`  :Mother`
	`  ObjectIntersectionOf(`
	`    :Woman`
	`    :Parent`
	`  )`
	`)`

But there are also other syntaxes in regular use. We show the functional syntax which is inspired by the mathematical notation of relations.

Extended OWL Example in Functional Syntax

▷ **Example 12.6.13.** The semantic network from ?? can be expressed in OWL (in functional syntax)



```
ClassAssertion (:Jack :robin)
ClassAssertion(:John :person)
ClassAssertion (:Mary :person)
ObjectPropertyAssertion(:loves :John :Mary)
ObjectPropertyAssertion(:owner :John :Jack)
SubClassOf(:robin :bird)
SubClassOf (:bird ObjectSomeValuesFrom(:hasPart :wing))
```

- ▷ ClassAssertion formalizes the "inst" relation,
- ▷ ObjectPropertyAssertion formalizes relations,
- ▷ SubClassOf formalizes the "isa" relation,
- ▷ for the "has_part" relation, we have to specify that *all birds have a part that is a wing* or equivalently *the class of birds is a subclass of all objects that have some wing*.

We have introduced the ideas behind using description logics as the basis of a "machine-oriented web of data". While the first OWL specification (2004) had three sublanguages "OWL Lite", "OWL DL" and "OWL Full", of which only the middle was based on description logics, with the OWL2 Recommendation from 2009, the foundation in description logics was nearly universally accepted.

The semantic web hype is by now nearly over, the technology has reached the "plateau of productivity" with many applications being pursued in academia and industry. We will not go into these, but briefly instroduce one of the tools that make this work.

SPARQL an RDF Query language

- ▷ **Definition 12.6.14.** SPARQL, the "SPARQL Protocol and RDF Query Language" is an RDF query language, able to retrieve and manipulate data stored in RDF. The SPARQL language was standardized by the World Wide Web Consortium in 2008 [PS08].
- ▷ SPARQL is pronounced like the word "*sparkle*".
- ▷ **Definition 12.6.15.** A system is called a SPARQL endpoint, iff it answers SPARQL queries.
- ▷ **Example 12.6.16.** Query for person names and their e-mails from a triplestore with FOAF data.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
  ?person a foaf:Person.
  ?person foaf:name ?name.
  ?person foaf:mbox ?email.
}
```

SPARQL end-points can be used to build interesting applications, if fed with the appropriate data. An interesting – and by now paradigmatic – example is the DBPedia project, which builds a large ontology by analyzing Wikipedia fact boxes. These are in a standard HTML form which can be analyzed e.g. by regular expressions, and their entries are essentially already in triple form: The subject is the Wikipedia page they are on, the predicate is the key, and the object is either the URI on the object value (if it carries a link) or the value itself.

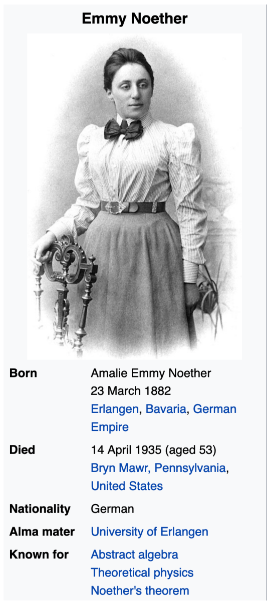
SPARQL Applications: DBPedia

▷ **Typical Application:** DBPedia screen-scrapes Wikipedia fact boxes for RDF triples and uses SPARQL for querying the induced triplestore.

▷ **Example 12.6.17 (DBPedia Query).** People who were born in Erlangen before 1900 (`http://dbpedia.org/snorql`)

```
SELECT ?name ?birth ?death ?person WHERE {
    ?person dbo:birthPlace :Erlangen .
    ?person dbo:birthDate ?birth .
    ?person foaf:name ?name .
    ?person dbo:deathDate ?death .
    FILTER (?birth < "1900—01—01"^^xsd:date) .
}
ORDER BY ?name
```

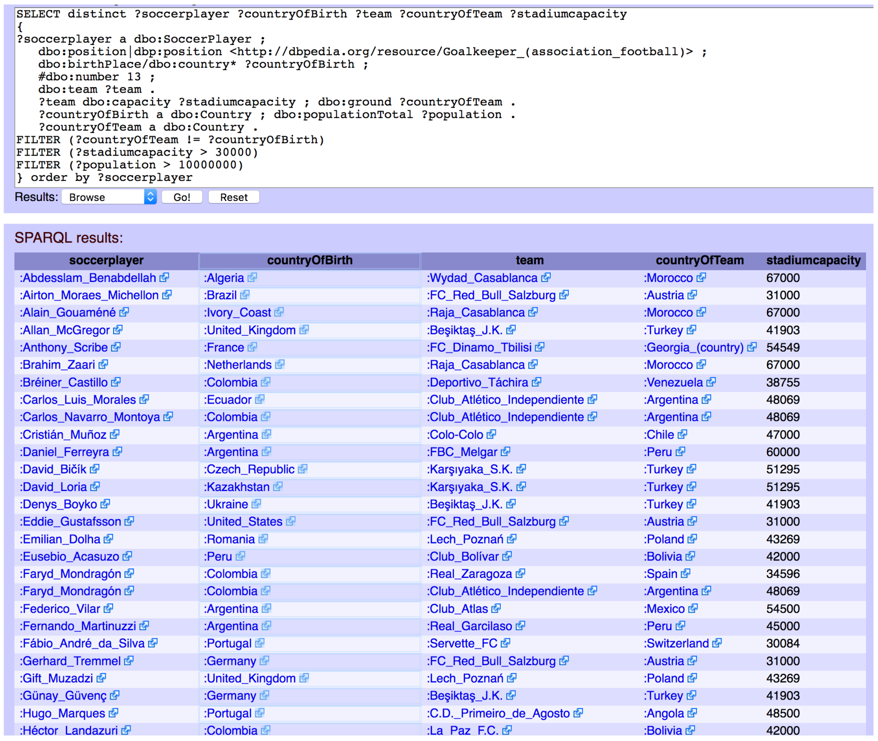
▷ The answers include Emmy Noether and Georg Simon Ohm.


Emmy Noether

Born	Amalie Emmy Noether 23 March 1882 Erlangen, Bavaria, German Empire
Died	14 April 1935 (aged 53) Bryn Mawr, Pennsylvania, United States
Nationality	German
Alma mater	University of Erlangen
Known for	Abstract algebra Theoretical physics Noether's theorem

A more complex DBPedia Query

▷ **Demo:** DBPedia `http://dbpedia.org/snorql/`
Query: Soccer players born in a country with more than 10 M inhabitants, who play as goalie in a club that has a stadium with more than 30.000 seats.
Answer: computed by DBPedia from a SPARQL query



```
SELECT distinct ?soccerplayer ?countryOfBirth ?team ?countryOfTeam ?stadiumcapacity
{
?soccerplayer a dbo:SoccerPlayer ;
   dbo:position|dbp:position <http://dbpedia.org/resource/Goalkeeper_(association_football)> ;
   dbo:birthPlace/dbo:country* ?countryOfBirth ;
   #dbo:number 13 ;
   dbo:team ?team .
   ?team dbo:capacity ?stadiumcapacity ; dbo:ground ?countryOfTeam .
   ?countryOfBirth a dbo:Country ; dbo:populationTotal ?population .
   ?countryOfTeam a dbo:Country .
FILTER (?countryOfTeam != ?countryOfBirth)
FILTER (?stadiumcapacity > 30000)
FILTER (?population > 10000000)
} order by ?soccerplayer
```

SPARQL results:

soccerplayer	countryOfBirth	team	countryOfTeam	stadiumcapacity
:Abdesslam_Benabdellah	:Algeria	:Wydad_Casablanca	:Morocco	67000
:Airton_Moraes_Michellon	:Brazil	:FC_Red_Bull_Salzburg	:Austria	31000
:Alain_Gouaméné	:Ivory_Coast	:Raja_Casablanca	:Morocco	67000
:Allan_McGregor	:United_Kingdom	:Beşiktaş_J.K.	:Turkey	41903
:Anthony_Scribe	:France	:FC_Dinamo_Tbilisi	:Georgia_(country)	54549
:Brahim_Zaari	:Netherlands	:Raja_Casablanca	:Morocco	67000
:Bréiner_Castillo	:Colombia	:Deportivo_Táchira	:Venezuela	38755
:Carlos_Luis_Morales	:Ecuador	:Club_Atlético_Independiente	:Argentina	48069
:Carlos_Navarro_Montoya	:Colombia	:Club_Atlético_Independiente	:Argentina	48069
:Cristián_Muñoz	:Argentina	:Colo-Colo	:Chile	47000
:Daniel_Ferreyra	:Argentina	:FBC_Melgar	:Peru	60000
:David_Bičík	:Czech_Republic	:Karşıyaka_S.K.	:Turkey	51295
:David_Loria	:Kazakhstan	:Karşıyaka_S.K.	:Turkey	51295
:Denys_Boyko	:Ukraine	:Beşiktaş_J.K.	:Turkey	41903
:Eddie_Gustafsson	:United_States	:FC_Red_Bull_Salzburg	:Austria	31000
:Emilian_Dolha	:Romania	:Lech_Poznań	:Poland	43269
:Eusebio_Acasuzo	:Peru	:Club_Bolívar	:Bolivia	42000
:Faryd_Mondragón	:Colombia	:Real_Zaragoza	:Spain	34596
:Faryd_Mondragón	:Colombia	:Club_Atlético_Independiente	:Argentina	48069
:Federico_Vilar	:Argentina	:Club_Atlas	:Mexico	54500
:Fernando_Martinuzzi	:Argentina	:Real_Garcilaso	:Peru	45000
:Fábio_André_da_Silva	:Portugal	:Servette_FC	:Switzerland	30084
:Gerhard_Tremmel	:Germany	:FC_Red_Bull_Salzburg	:Austria	31000
:Gift_Muzadzi	:United_Kingdom	:Lech_Poznań	:Poland	43269
:Günay_Güvenç	:Germany	:Beşiktaş_J.K.	:Turkey	41903
:Hugo_Marques	:Portugal	:C.D._Primeiro_de_Agosto	:Angola	48500
:Héctor_Landazuri	:Colombia	:La_Paz_F.C.	:Bolivia	42000

We conclude our survey of the semantic web technology stack with the notion of a triplestore, which refers to the database component, which stores vast collections of ABox triples.

Triple Stores: the Semantic Web Databases

▷ **Definition 12.6.18.** A triplestore or RDF store is a purpose-built database for the storage RDF graphs and retrieval of RDF triples usually through variants of SPARQL.

▷ Common triplestores include

- ▷ Virtuoso: `https://virtuoso.openlinksw.com/` (used in DBpedia)
- ▷ GraphDB: `http://graphdb.ontotext.com/` (often used in WissKI)
- ▷ blazegraph: `https://blazegraph.com/` (open source; used in WikiData)

▷ **Definition 12.6.19.** A description logic reasoner implements of reaonsing services based on a satisfiabiltiy test for description logics.

▷ Common description logic reasoners include

- ▷ FACT++: `http://owl.man.ac.uk/factplusplus/`
- ▷ HermiT: `http://www.hermit-reasoner.com/`

▷ **Intuition:** Triplestores concentrate on querying very large ABoxes with partial consideration of the TBox, while DL reasoners concentrate on the full set of ontology inference services, but fail on large ABoxes.

12.7 Ontologies vs. Databases

To understand ontologies better and contrast them to database systems to understand their respective possible role in documenting cultural artefacts. We start off with a definition of the concept and components of an ontology.

We will still keep our presentation of the material at a general level without committing to a particular ontology language or system.

We now consolidate our understanding of all these concepts with an example. We build an ontology by first contstructing a TBox and then a corresponding ABox.

Example: Hogwarts Ontology

▷ **Example 12.7.1.** Axioms describe the structure of the world,

```
Class HogwartsStudent = Student and attendsSchool Hogwarts
Class: HogwartsStudent ⊑ hasPet only (Owl or Cat or Toad)
ObjectProperty: hasPet Inverses: isPetOf
Class: Phoenix ⊑ isPetOf only Wizard
```

▷ **Example 12.7.2.** Facts describe some particular concrete situation,

```
Individual: Hedwig
   Types: Owl
Individual: HarryPotter
  Types: HogwartsStudent
  Facts: hasPet Hedwig
Individual: Fawkes
  Types: Phoenix
  Facts: isPetOf Dumbledore
```

It is very instructive to compare ontologies to databases. There are some similarities induced by the joint intention to represent structured data, but also some important differences, which will play a crucial role in our discussion later on.

Ontologies vs. Databases

▷ **Obvious Analogy:** In an ontology:

- ▷ axioms analogous to DB schema (structure and constraints on data)
- ▷ facts analogous to DB data
  - ▷ data instantiates schema, is consistent with schema constraints

▷ **But there are also important differences:**

Database:	Ontology:
▷ Closed world assumption (CWA)	▷ Open world assumption (OWA)
▷ Missing information treated as false	▷ Missing information treated as unknown
▷ Unique name assumption (UNA)	▷ No UNA
▷ Each individual has a single, unique name	▷ Individuals may have more than one name
▷ Schema behaves as constraints on structure of data	▷ Ontology axioms behave like implications (inference rules)
▷ Define legal database states.	▷ Entail implicit information

Let us elucidate these quite abstract concdepts and differences using a simple example, which we again take from the Hogwarts ontology (see ?? and ??).

DB vs. Ontology by Example (Querying)

▷ **Given the Ontology:**

```
Individual: HarryPotter
  Facts: hasFriend RonWeasley
         hasFriend HermioneGranger
         hasPet Hedwig
Individual: Draco Malfoy
```

▷ **Query:** Is Draco Malfoy a friend of HarryPotter?

- ▷ DB: No
- ▷ Ontology: Don't Know (OWA: didn't say Draco was not Harry's friend)

▷ **Counting Query:** How many friends does Harry Potter have?

- ▷ DB: 2
- ▷ Ontology: at least 1 (No UNA: Ron and Hermione may be 2 names for same person)

▷ **How about:** if we add

```
DifferentIndividuals: RonWeasley HermioneGranger
```

- ▷ DB: 2
- ▷ Ontology: at least 2 (OWA: Harry may have more friends we didn't mention yet)

▷ **And:** if we also add

```
Individual: HarryPotter
  Types: hasFriend only RonWeasley or HermioneGranger
```

- DB: 2
- Ontology: 2

We continue our example with the behavior if we insert new information to the Hogwarts ontology. Again, databases and ontology systems react differently.

DB vs. Ontology by Example (Insertion)

- **Given:** the ontology from ?? and ?? insert

```
Individual: Dumbledore
Individual: Fawkes
  Types: Phoenix
  Facts: isPetOf Dumbledore
```

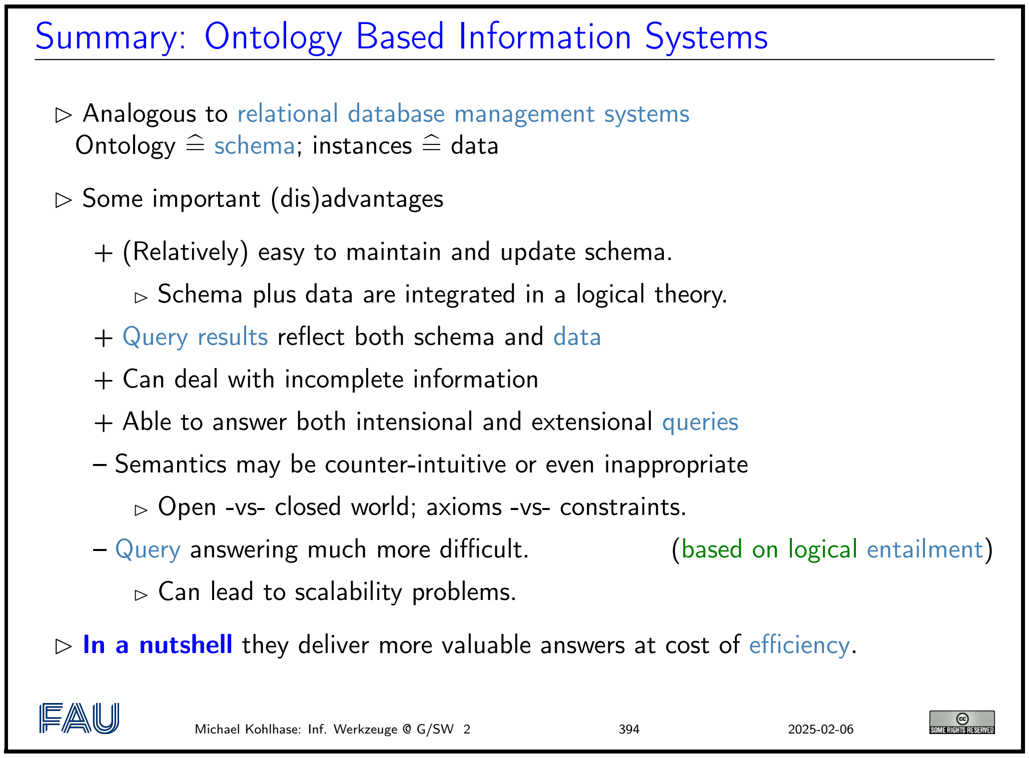
- **System Response:**
  - DB: Update rejected: constraint violation
    - Range of hasPet is Human; Dumbledore is not (CWA)
  - Ontology Reasoner:
    - Infer that Dumbledore is Human
    - Also infer that Dumbledore is a Wizard (only a Wizard can have a phoenix as a pet)

Finally, we come to one of the central disciplines in which to compare databases and ontology based information systems: query answering. Here we see a crucial difference: ontology queries are semantic, i.e. they take both axioms and facts into account.

DB vs. Ontology by Example: Query Answering

- DB schema plays no role in query answering (efficiently implementable)
- Ontology axioms play a powerful and crucial role in QA
  - Answer may include implicitly derived facts
  - Can answer conceptual as well as extensional queries
    E.g., *Can a* Muggle *have a* Phoenix *for a* pet?
  - May have very high worst case complexity ($\widehat{=}$ terrible running time)
    Implementations may still behave well in typical cases.
- **Definition 12.7.3.** We call a query language semantic, iff query answering involves derived axioms and facts.
- **Observation 12.7.4.** *Ontology queries are semantic, while database queries are not.*

We will now summarize what we have learned about ontology-based information systems.



Summary: Ontology Based Information Systems

▷ Analogous to relational database management systems
Ontology $\widehat{=}$ schema; instances $\widehat{=}$ data

▷ Some important (dis)advantages

- + (Relatively) easy to maintain and update schema.
  - ▷ Schema plus data are integrated in a logical theory.
- + Query results reflect both schema and data
- + Can deal with incomplete information
- + Able to answer both intensional and extensional queries
- – Semantics may be counter-intuitive or even inappropriate
  - ▷ Open -vs- closed world; axioms -vs- constraints.
- – Query answering much more difficult. (based on logical entailment)
  - ▷ Can lead to scalability problems.

▷ **In a nutshell** they deliver more valuable answers at cost of efficiency.

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 394 2025-02-06

12.8 Exercises

Problem 8.1 (Function/Argument Form of a Semantic Network)
Write the semantic network from **??** in function/argument notation.

Problem 8.2 (Evaluation of Semantic Networks)
Using the example from **??**, discuss the pros and cons – give two of each - of semantic networks.

Problem 8.3 (Semantic Web Technology)
Semantic web technology comes in two parts, RDF and OWL. Briefly describe their roles in the semantic web. How do they relate to $\mathcal{ALC}$?

Problem 8.4

1. Install the Protege System from `http://protege.stanford.edu/` on your computer and
2. use it to represent the following knowledge into an ABox:
   (a) *Vincent is the brother of Cecilia who is George's daughter.*
   (b) *Ruth is George's niece and Paul her brother.*
   (c) *Frida is George's mother.*
3. Define a TBox of family relationships (compliant to the common understanding) that is sufficiently rich so that the following relationships can be inferred (discuss the inferences).
   (a) *Paul is Cecilia's cousin.*
   (b) *Frida is Ruth's and Vincent's grandmother.*
   (c) *George has a brother or sister.*

Chapter 13

The WissKI System: A Virtual Research Environment for Cultural Heritage

We will now come to the WissKI system itself, which positions itself as a virtual research environment for cultural heritage. Indeed it is a comprehensive, ontology-based information system for documenting, studying, and presenting our cultural heritage.

Before we go into the technicalities of the WissKI system itself, let us recall the requirements and motivations.

WissKI: a Virtual Research Env. for Cultural Heritage

- ▷ **Definition 13.0.1.** WissKI is a virtual research environment (VRE) for managing scholarly data and documenting cultural heritage.
- ▷ **Requirements:** For a virtual research environment for cultural heritage, we need
  - ▷ scientific communication about and documentation of the cultural heritage
  - ▷ networking knowledge from different disciplines (transdisciplinarity)
  - ▷ high-quality data acquisition and analysis
  - ▷ safeguarding authorship, authenticity, persistence
  - ▷ support of scientific publication
- ▷ WissKI was developed by the research group of Prof. Günther Görtz at FAU Erlangen-Nürnberg and is now used in hundreds of DH projects across Germany.
- ▷ FAU supports cultural heritage research by providing hosted WissKI instances.
  - ▷ See `https://wisski.data.fau.de` for details
  - ▷ We will use an instance for the Kirmes paintings in the homework assignments.

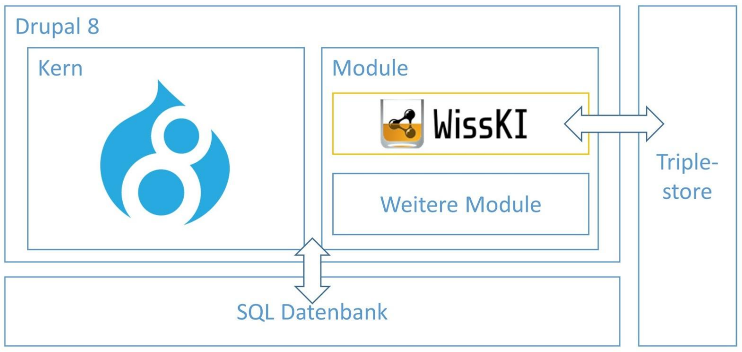
SOME RIGHTS RESERVED

13.1 WissKI extends Drupal

The first thing about the WissKI system is that it is realized as an extension of the drupal web content management system, which already provides many of the features (e.g. user management, web authoring, collaboration, . . .) a VRE needs to implement.

WissKI System Architecture

- ▷ Software basis: drupal CMS (content management system)
  - ▷ large, active community, extensible by drupal modules
  - ▷ provides much of the functionality of a VRE out of the box.

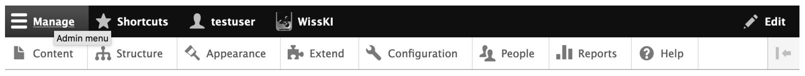


We now give a general overview of the drupal system, and introduce the concepts we need for understanding WissKI system. Naturally, this does now do the drupal WCMS justice. For an introduction we refer readers to [Gla17; Tom17] and the drupal web site [Dru].

Drupal: A Web Content Managemt Framework

- ▷ **Definition 13.1.1.** Drupal is an open source web content management application. It combines CMS functionality with knowledge management via RDF.
- ▷ **Definition 13.1.2.** Drupal allows to configure web pages modularly from content blocks, which can be
  - ▷ static content, i.e. supplied by a module,
  - ▷ user supplied content, or
  - ▷ views, i.e. listings of content fragments from other blocks.

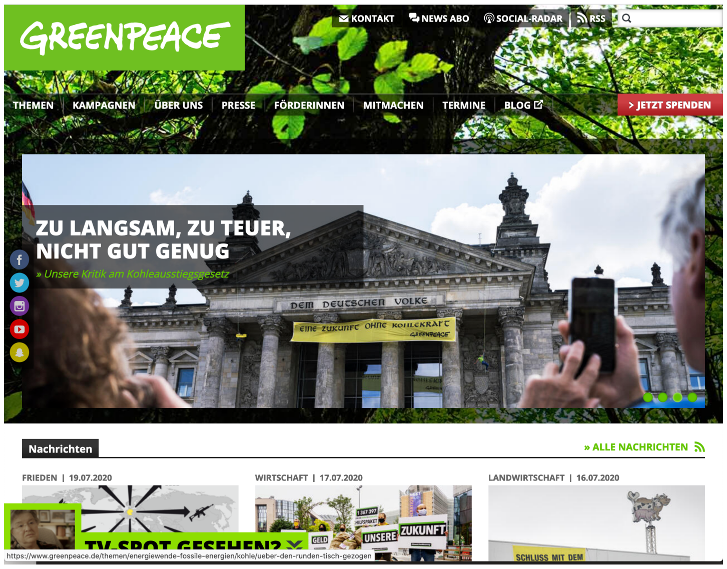
  These can be assembled into web pages via a visual interface: the config bar.



To fortify our intuition about the concepts introduced above, let us try to find them in an existing web page.

Assembling a Web Site via Drupal Blocks (Example)

▷ **Example 13.1.3 (Greenpeace via Drupal).** Can you find the blocks?



We now come to one of the most important features used in WissKI: drupal is modular and extensible; this allows us to build the features for an ontology-based information system as drupal modules.

Drupal Modules and Themes

▷ **Idea:** Drupal is designed to be modular and extensible (so it can adapt to the ever-changing web)

▷ **Definition 13.1.4 (Modular Design).** Drupal functionality is structured into

- ▷ drupal core – the basic CMS functionality
- ▷ modules which contribute e.g. new block types (∼ 45.000)
- ▷ themes which contribute new UI layouts (∼ 2800)

Drupal core is the vanilla system as downloaded, modules and themes must be installed and configured separately via the config bar.

▷ The drupal core functionalities include

- ▷ user/account management
- ▷ menu management,
- ▷ RSS feeds,
- ▷ taxonomy,
- ▷ page layout customization (via blocks and views),
- ▷ system administration

This brings us to the central data acquisition subsystem in drupal, which we will use to build our system. Much of the actual data in the drupal system is internally stored in terms of dictionaries: systems of key/value pairs.

Bundles and Fields in Drupal (Data Entry)

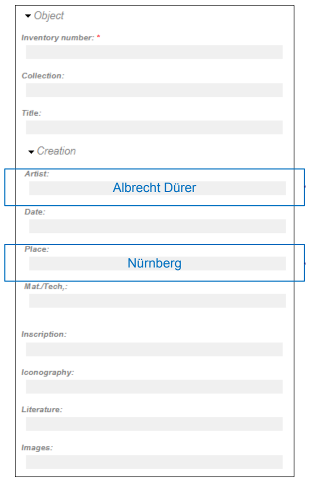
▷ **Definition 13.1.5.** Drupal has a special data type called a bundle, which is essentially a dictionary: it contains key/value pairs called fields.

- ▷ bundles can be nested ⇝ sub bundles.
- ▷ fields also have data type information, etc. to support editing.

▷ drupal presents bundles as

- ▷ HTML lists for reading
- ▷ HTML forms for data entry/editing

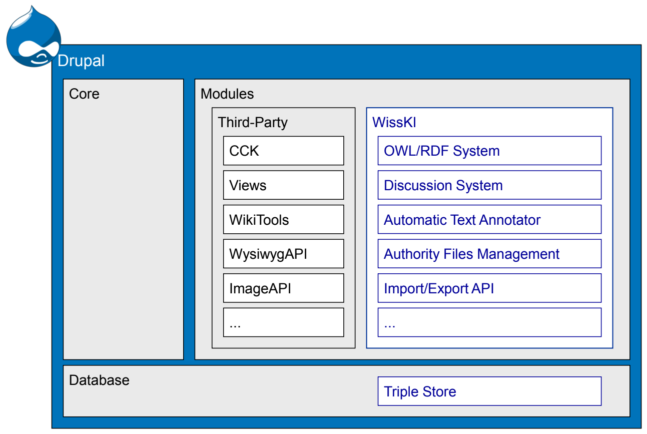
▷ Drupal bundles induce blocks that can be used for data entry and presentation.



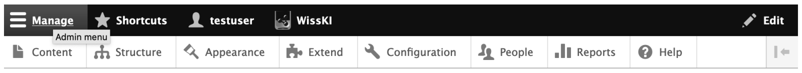
Now we can summarize the WissKI architecture in a simple equation. While this glosses over many of finer points in the system, it is important to keep this in mind for working with the system in practice.

WissKI System Architecture (Recap)

▷ WissKI = drupal + CIDOC CRM + triplestore + WissKI modules



▷ **Note:** Much of WissKI functionality is configurable via the drupal config bar.



13.2 Dealing with Ontology Paths: The WissKI Pathbuilder

We now come to what is probably the defining feature of WissKI: the WissKI path builder. It solves the problem that with ontologies, even for simple facts we have to generate entire ontology paths.

The WissKI Path Builder (Idea)

▷ **Recall:** *Albrecht Dürer painted Melencolia 1 in Nürnberg*

$$\begin{tikzcd}
 & & \boxed{d:E21} \arrow[r, "P131"] & \boxed{a:E82} \arrow[r, "P3"] & \texttt{"A. Dürer"} \\
\boxed{m:E84} \arrow[r, "P108i"] & \boxed{q:E12} \arrow[ur, "P14"] \arrow[dr, "P7"] & & & \\
 & & \boxed{p:E53} \arrow[r, "P87"] & \boxed{n:E48} \arrow[r, "P3"] & \texttt{"Nürnberg"}
\end{tikzcd}$$

▷ **Idea:** Hide the complexity induced by the ontology from the user

- ▷ Form-based interaction with categories and fields (as in a RDBMS UI)

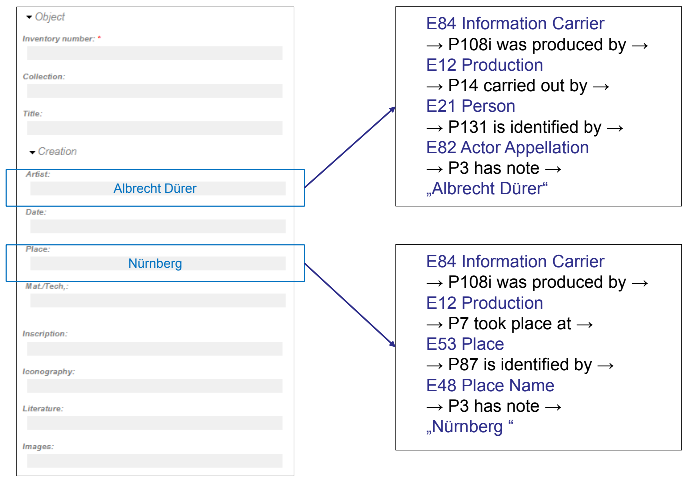
▷ **Definition 13.2.1.** The WissKI path builder maps ontology groups and ontology paths to drupal bundles and fields.

- ▷ ontology groups become data entry forms (bundles) for the root entities,
- ▷ their fields are mapped to ontology paths.
- ▷ subtrees in the ontology become sub-bundles. (shared objects)

Even though we have introduced all the necessary concepts above, the best way of understanding this is to look at our running example again: the path builder induces a data entry form that allows us to enter a whole set of ontology paths, introducing and sharing intermediary objects along the way.

The WissKI Path Builder (Example)

▷ **Example 13.2.2 (A WissKI Group).**



Albrecht Dürer

Nürnberg

E84 Information Carrier
→ P108i was produced by →
E12 Production
→ P14 carried out by →
E21 Person
→ P131 is identified by →
E82 Actor Appellation
→ P3 has note →
„Albrecht Dürer"

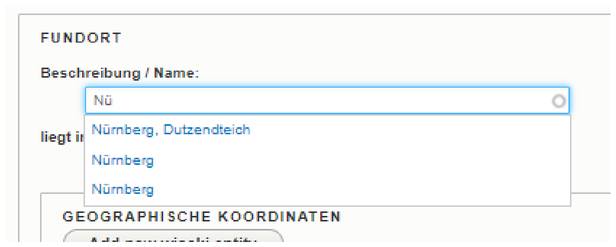
E84 Information Carrier
→ P108i was produced by →
E12 Production
→ P7 took place at →
E53 Place
→ P87 is identified by →
E48 Place Name
→ P3 has note →
„Nürnberg "

If we look at the data entry form on the left of ??, then we see that we only enter strings, not the objects we mean. So there is the problem of disambiguating which objects that are then linked to some object via CIDOC CRM relations we actually mean with the string.

Sharing and Disambiguation in Path Builders

▷ **Observation 13.2.3.** *Sometimes we want to refer to existing entities in WissKI.*

▷ **Example 13.2.4 (Referring to Nürnberg).** (We love tab completion)



FUNDORT

Beschreibung / Name:

Nü

Nürnberg, Dutzendteich

Nürnberg

Nürnberg

▷ **Example 13.2.5 (To What).** Albrecht Dürer created all his etchings in Nürnberg.

▷ **Problem:** (In paths) we are creating lots of objects, which ones to offer?

▷ **Idea:** Mark the entities we might want to reuse on paths while specifying them.

▷ **Definition 13.2.6.** A disambiguation point in a path marks an entity that can be re used in data acquisition.

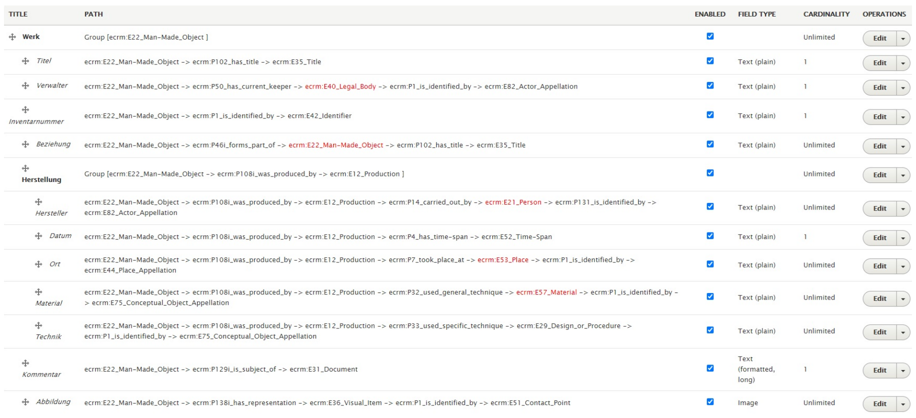
▷ **Example 13.2.7.** Disambiguation points are highlighted in red on paths.

$$\begin{tikzcd}
 & & \boxed{d:E21} \arrow[r, "\text{P131}"] & \boxed{a:E82} \arrow[r, "\text{P3}"] & \texttt{"A. Dürer"} \\
\boxed{m:E84} \arrow[r, "\text{P108i}"] & \boxed{q:E12} \arrow[ur, "\text{P14}"] \arrow[dr, "\text{P7}"] & & & \\
 & & \boxed{p:E53} \arrow[r, "\text{P87}"] & \boxed{n:E48} \arrow[r, "\text{P3}"] & \texttt{"Nürnberg"}
\end{tikzcd}$$

Now we can have a look at how drupal sees (and shows) path builders

Specifying/Maintaining WissKI Path Builders

- **Recall:** A WissKI path builder maps ontology groups and ontology paths to drupal bundles and fields.
- **Example 13.2.8 (Specifying a WissKI Path Builder).**

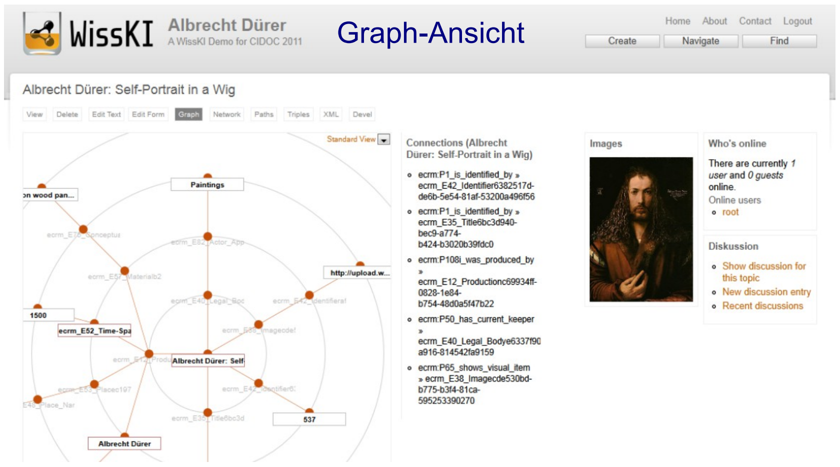


TITLE	PATH	ENABLED	FIELD TYPE	CARDINALITY	OPERATIONS
Werk	Group [ecrm:E22_Man-Made_Object]	✓		Unlimited	Edit
Titel	ecrm:E22_Man-Made_Object -> ecrm:P102_has_title -> ecrm:E35_Title	✓	Text (plain)	1	Edit
Verwalter	ecrm:E22_Man-Made_Object -> ecrm:P50_has_current_keeper -> ecrm:E40_Legal_Body -> ecrm:P1_is_identified_by -> ecrm:E82_Actor_Appellation	✓	Text (plain)	1	Edit
Inventarnummer	ecrm:E22_Man-Made_Object -> ecrm:P1_is_identified_by -> ecrm:E42_Identifier	✓	Text (plain)	1	Edit
Beziehung	ecrm:E22_Man-Made_Object -> ecrm:P46i_forms_part_of -> ecrm:E22_Man-Made_Object -> ecrm:P102_has_title -> ecrm:E35_Title	✓	Text (plain)	Unlimited	Edit
Herstellung	Group [ecrm:E22_Man-Made_Object -> ecrm:P108i_was_produced_by -> ecrm:E12_Production]	✓		Unlimited	Edit
Hersteller	ecrm:E22_Man-Made_Object -> ecrm:P108i_was_produced_by -> ecrm:E12_Production -> ecrm:P14_carried_out_by -> ecrm:E21_Person -> ecrm:P131_is_identified_by -> ecrm:E82_Actor_Appellation	✓	Text (plain)	Unlimited	Edit
Datum	ecrm:E22_Man-Made_Object -> ecrm:P108i_was_produced_by -> ecrm:E12_Production -> ecrm:P4_has_time-span -> ecrm:E52_Time-Span	✓	Text (plain)	1	Edit
Ort	ecrm:E22_Man-Made_Object -> ecrm:P108i_was_produced_by -> ecrm:E12_Production -> ecrm:P7_took_place_at -> ecrm:E53_Place -> ecrm:P1_is_identified_by -> ecrm:E44_Place_Appellation	✓	Text (plain)	Unlimited	Edit
Material	ecrm:E22_Man-Made_Object -> ecrm:P108i_was_produced_by -> ecrm:E12_Production -> ecrm:P32_used_general_technique -> ecrm:E57_Material -> ecrm:P1_is_identified_by -> ecrm:E75_Conceptual_Object_Appellation	✓	Text (plain)	Unlimited	Edit
Technik	ecrm:E22_Man-Made_Object -> ecrm:P108i_was_produced_by -> ecrm:E12_Production -> ecrm:P33_used_specific_technique -> ecrm:E29_Design_or_Procedure -> ecrm:P1_is_identified_by -> ecrm:E75_Conceptual_Object_Appellation	✓	Text (plain)	Unlimited	Edit
Kommentar	ecrm:E22_Man-Made_Object -> ecrm:P129i_is_subject_of -> ecrm:E31_Document	✓	Text (formatted, long)	1	Edit
Abbildung	ecrm:E22_Man-Made_Object -> ecrm:P138i_has_representation -> ecrm:E36_Visual_Item -> ecrm:P1_is_identified_by -> ecrm:E51_Contact_Point	✓	Image	Unlimited	Edit

Of course all paths of an ontology group can be visualized as a graph. WissKI supports this as well.

WissKI Path Builders as Graphs

- **Example 13.2.9 (A WissKI Path Builder as a Graph).**



Albrecht Dürer: Self-Portrait in a Wig

Connections (Albrecht Dürer: Self-Portrait in a Wig)

- ecrm:P1_is_identified_by » ecrm_E42_Identifier6382517d-de6b-5e54-81af-53200a496f56
- ecrm:P1_is_identified_by » ecrm_E35_Title6bc3d940-bec9-a774-b424-b3020b39fdc0
- ecrm:P108i_was_produced_by » ecrm_E12_Productionc69934ff-0828-1e84-b754-48d0a5f47b22
- ecrm:P50_has_current_keeper » ecrm_E40_Legal_Bodye6337f90 a916-814542fa9159
- ecrm:P65_shows_visual_item » ecrm_E38_Imagecde530bd-b775-b3f4-81ca-595253390270

There are currently *1 user* and *0 guests* online.
Online users
- root

- Show discussion for this topic
- New discussion entry
- Recent discussions

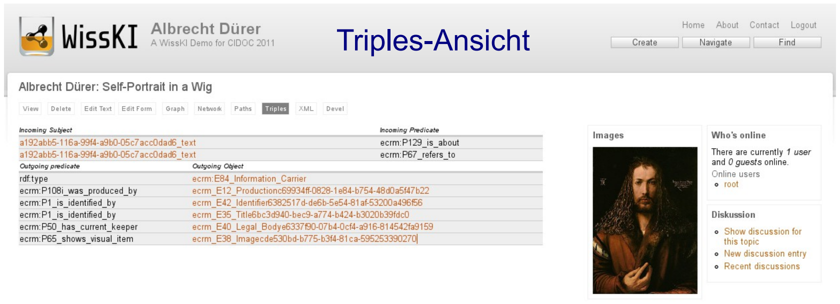
- Very nice and helpful, but does not work currently!

And finally, a path builder can be seen as a set of triples indeed this is the default export format for path builders.
Of course all paths of an ontology group can be visualized as a graph. WissKI supports this as well.

WissKI Path Builders as Triples

▷ **Of course** we can view path builders as sets of triples.

▷ **Example 13.2.10 (A WissKI Path Builder as Triples).**


WissKI
Albrecht Dürer
A WissKI Demo for CIDOC 2011
Triples-Ansicht
Home About Contact Logout
Create Navigate Find
View Delete Edit Text Edit Form Graph Network Paths Triples XML Devel
Images
Who's online
There are currently 1 user and 0 guests online.
Online users
root
Diskussion
Show discussion for this topic
New discussion entry
Recent discussions

Albrecht Dürer: Self-Portrait in a Wig

Incoming Subject	*Incoming Predicate*
a192abb5-116a-99f4-a9b0-05c7acc0dad6_text	ecrm:P129_is_about
a192abb5-116a-99f4-a9b0-05c7acc0dad6_text	ecrm:P67_refers_to

Outgoing predicate	*Outgoing Object*
rdf:type	ecrm:E84_Information_Carrier
ecrm:P108i_was_produced_by	ecrm_E12_Productionc69934ff-0828-1e84-b754-48d0a5f47b22
ecrm:P1_is_identified_by	ecrm_E42_Identifier6382517d-de6b-5e54-81af-53200a496f56
ecrm:P1_is_identified_by	ecrm_E35_Title6bc3d940-bec9-a774-b424-b3020b39fdc0
ecrm:P50_has_current_keeper	ecrm_E40_Legal_Bodye6337f90-07b4-0cf4-a916-814542fa9159
ecrm:P65_shows_visual_item	ecrm_E38_Imagecde530bd-b775-b3f4-81ca-595253390270

▷ Such an export also allows standardized communication.

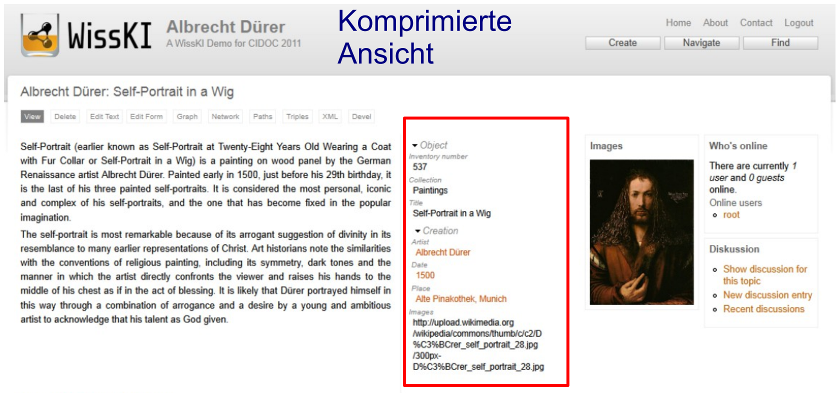
But of course, path builders can not only be used as data acquisition devices. They also define drupal blocks which can be used for data visualization (akin to fact boxes in Wikipedia).

Data Presentation using Path Builders in WissKI

▷ Path builders can be used as drupal blocks for data presentation.

▷ For every object o, aggregate the values of the paths starting in o.

▷ **Example 13.2.11 (Compressed View).**


WissKI
Albrecht Dürer
A WissKI Demo for CIDOC 2011
Komprimierte Ansicht
Home About Contact Logout
Create Navigate Find
View Delete Edit Text Edit Form Graph Network Paths Triples XML Devel
Images
Who's online
There are currently 1 user and 0 guests online.
Online users
root
Diskussion
Show discussion for this topic
New discussion entry
Recent discussions

Albrecht Dürer: Self-Portrait in a Wig

Self-Portrait (earlier known as Self-Portrait at Twenty-Eight Years Old Wearing a Coat with Fur Collar or Self-Portrait in a Wig) is a painting on wood panel by the German Renaissance artist Albrecht Dürer. Painted early in 1500, just before his 29th birthday, it is the last of his three painted self-portraits. It is considered the most personal, iconic and complex of his self-portraits, and the one that has become fixed in the popular imagination.

The self-portrait is most remarkable because of its arrogant suggestion of divinity in its resemblance to many earlier representations of Christ. Art historians note the similarities with the conventions of religious painting, including its symmetry, dark tones and the manner in which the artist directly confronts the viewer and raises his hands to the middle of his chest as if in the act of blessing. It is likely that Dürer portrayed himself in this way through a combination of arrogance and a desire by a young and ambitious artist to acknowledge that his talent as God given.

▾ Object
Inventory number
537
Collection
Paintings
Title
Self-Portrait in a Wig
▾ Creation
Artist
Albrecht Dürer
Date
1500
Place
Alte Pinakothek, Munich
Images
http://upload.wikimedia.org/wikipedia/commons/thumb/c/c2/D%C3%BCrer_self_portrait_28.jpg/300px-D%C3%BCrer_self_portrait_28.jpg

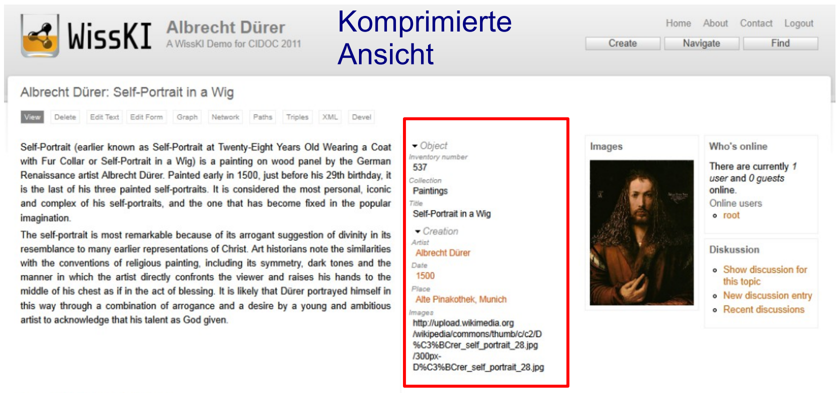
SOME RIGHTS RESERVED

13.3 The WissKI Link Block

The WissKI Link Block (Idea)

▷ **Observation 13.3.1.** *For an entity in a RDF graph, both the outgoing and the incoming edges are important for understanding.*

▷ **Example 13.3.2.** This view only shows the outgoing edges!



Albrecht Dürer: Self-Portrait in a Wig

Self-Portrait (earlier known as Self-Portrait at Twenty-Eight Years Old Wearing a Coat with Fur Collar or Self-Portrait in a Wig) is a painting on wood panel by the German Renaissance artist Albrecht Dürer. Painted early in 1500, just before his 29th birthday, it is the last of his three painted self-portraits. It is considered the most personal, iconic and complex of his self-portraits, and the one that has become fixed in the popular imagination.

The self-portrait is most remarkable because of its arrogant suggestion of divinity in its resemblance to many earlier representations of Christ. Art historians note the similarities with the conventions of religious painting, including its symmetry, dark tones and the manner in which the artist directly confronts the viewer and raises his hands to the middle of his chest as if in the act of blessing. It is likely that Dürer portrayed himself in this way through a combination of arrogance and a desire by a young and ambitious artist to acknowledge that his talent as God given.

Object
Inventory number
537
Collection
Paintings
Title
Self-Portrait in a Wig
Creation
Artist
Albrecht Dürer
Date
1500
Place
Alte Pinakothek, Munich
Images
http://upload.wikimedia.org/wikipedia/commons/thumb/c/c2/D%C3%BCrer_self_portrait_28.jpg/300px-D%C3%BCrer_self_portrait_28.jpg

There are currently *1 user* and *0 guests* online.

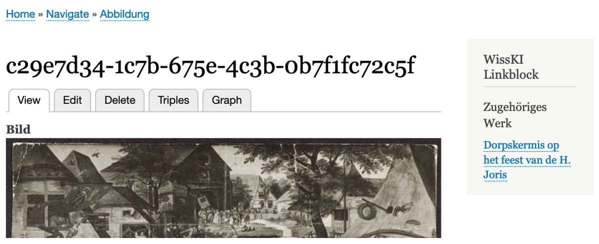
▷ **Idea:** Add a block with “incoming links” to the page, use the path builder.

SOME RIGHTS RESERVED

Link Blocks (Definition)

▷ **Definition 13.3.3.** Let p be a drupal page for an ontology group g, then a WissKI link block is a special drupal block with associated path builder, whose ontology paths all end in g.

▷ **Example 13.3.4 (A link block for Images).**


Home » Navigate » Abbildung
c29e7d34-1c7b-675e-4c3b-0b7f1fc72c5f
View
Edit
Delete
Triples
Graph
Bild
WissKI Linkblock
Zugehöriges Werk
Dorpskermis op het feest van de H. Joris

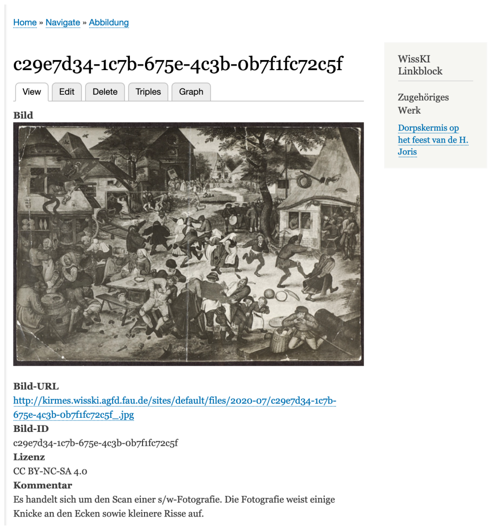
Note the difference between

- ▷ a “work” – the original painting Pieter Brueghel created in 1628
- ▷ and an “image of the work” – a b/w photograph of the “work”.

This particular link block mediates between these two.

A Link Block in the Wild (the full Picture)

▷ **Example 13.3.5 (A link block for Images).**



Home » Navigate » Abbildung

c29e7d34-1c7b-675e-4c3b-0b7f1fc72c5f

View Edit Delete Triples Graph

Bild

Bild-URL
http://kirmes.wisski.agfd.fau.de/sites/default/files/2020-07/c29e7d34-1c7b-675e-4c3b-0b7f1fc72c5f_.jpg
Bild-ID
c29e7d34-1c7b-675e-4c3b-0b7f1fc72c5f
Lizenz
CC BY-NC-SA 4.0
Kommentar
Es handelt sich um den Scan einer s/w-Fotografie. Die Fotografie weist einige Knicke an den Ecken sowie kleinere Risse auf.

WissKI Linkblock

Zugehöriges Werk

Dorpskermis op het feest van de H. Joris

▷ **outgoing relations** below the image,

▷ **incoming ones** in the link block

Making Link Blocks via the Path Builder

▷ How to make a link block in page p for group g? (Details at [WH])

1. create a block via the config bar and place it on p.
2. associate it with a link block path builder
3. model paths into g in the path builder (various source groups)

▷ **Idea:** You essentially know link block paths already: If you have already modeled a path $g, r_1, \ldots, r_n, s$ for a group s, then you have a path $s, r_n^{-1}, \ldots, r_1^{-1}, g$, where r_i^{-1} are the inverse roles of r_i (exist in CIDOC CRM)

$$\begin{tikzcd}
 & & \boxed{d : E21} \arrow[r, "\text{P131}"] \arrow[dl, bend left, "\text{P14i}"] & \boxed{a : E82} \arrow[r, "\text{P3}"] & \texttt{"A. Dürer"} \\
\boxed{m : E84} \arrow[r, shift left, "\text{P108i}"] & \boxed{q : E12} \arrow[l, shift left, "\text{P108}"] \arrow[ur, bend left, "\text{P14}"] \arrow[dr, bend left, "\text{P7}"] & & & \\
 & & \boxed{p : E53} \arrow[ul, bend left, "\text{P7i}"] \arrow[r, "\text{P87}"] & \boxed{n : E48} \arrow[r, "\text{P3}"] & \texttt{"Nürnberg"}
\end{tikzcd}$$

▷ **Note:** With this setup, you never have to fill out the link block paths!

13.4 Cultural Heritage Research: Querying WissKI Resources

So far, we have concentrated on the WissKI system, and how that can be used for data acquisition and documentation of cultural artefacts. While we did this we lost view of the most important aspect: what are we doing data acquisition for? Arguably this is cultural heritage research – and we mean this in an inclusive manner – this could be academic research or researching for a school project or article in a newspaper.

This research and how the WissKI system can support is what we will go into now.

Research in WissKI

▷ **So far** we have seen how to acquire complex knowledge about cultural artefacts using CIDOC CRM ABoxes.

▷ **Question:** But how do we do research using WissKI?

▷ **Answer:** Finding patterns, inherent connections, . . . in the data.

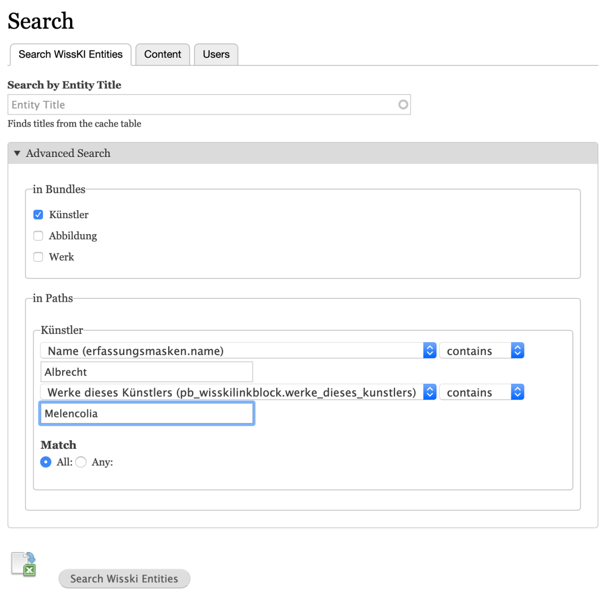
▷ **But how?:** That depends on the kind of research you want to do. Here are some WissKI research tools

1. we can use drupal search on the data.
2. We can formulate our own queries in SPARQL
3. We can pre-configure various queries in drupal views.

The simplest form of "research" is just being able to search over the objects that have been created. This is one of the basic facilities WissKI offers out of the box. Already that can be quite useful.

Drupal Search in WissKI

▷ **Example 13.4.1.**


Search

Search WissKI Entities | Content | Users

Search by Entity Title

Entity Title

Finds titles from the cache table

▼ Advanced Search

in Bundles

☑ Künstler
☐ Abbildung
☐ Werk

in Paths

Künstler

Name (erfassungsmasken.name) contains

Albrecht

Werke dieses Künstlers (pb_wisskilinkblock.werke_dieses_kunstlers) contains

Melencolia

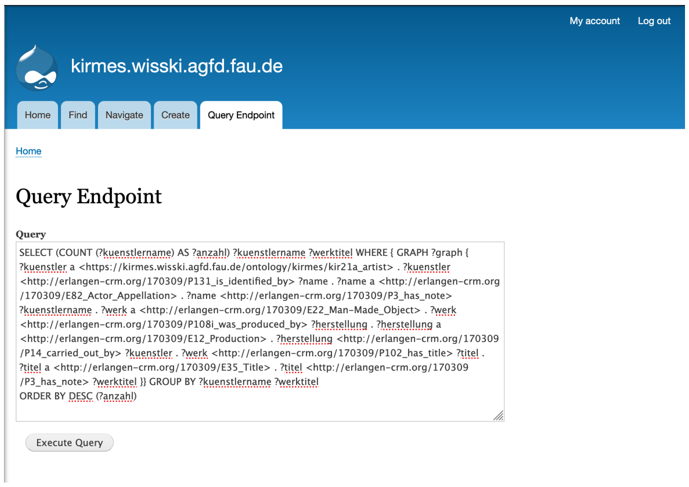
Match

[✓ All:] Any:

Search Wisski Entities

SPARQL Endpoint in WissKI

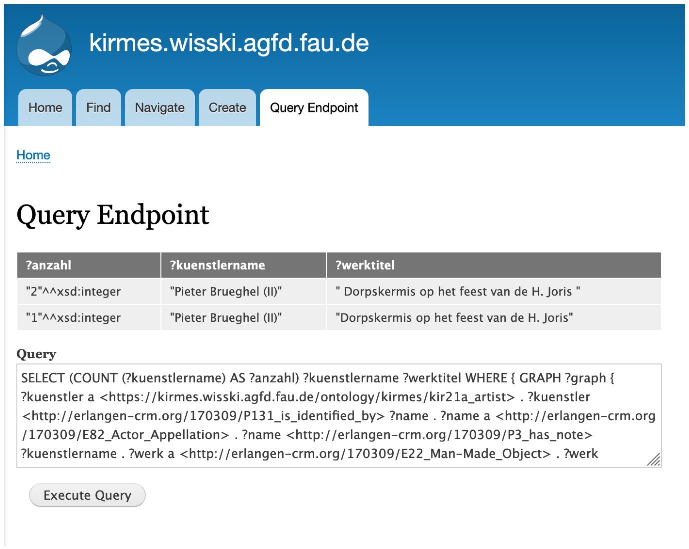
▷ **Example 13.4.2.** Find kirmes paintings and their painters and count them



Query Endpoint

Query

```
SELECT (COUNT (?kuenstlername) AS ?anzahl) ?kuenstlername ?werktitel WHERE { GRAPH ?graph {
?kuenstler a <https://kirmes.wisski.agfd.fau.de/ontology/kirmes/kir21a_artist> . ?kuenstler
<http://erlangen-crm.org/170309/P131_is_identified_by> ?name . ?name a <http://erlangen-crm.org
/170309/E82_Actor_Appellation> . ?name <http://erlangen-crm.org/170309/P3_has_note>
?kuenstlername . ?werk a <http://erlangen-crm.org/170309/E22_Man-Made_Object> . ?werk
<http://erlangen-crm.org/170309/P108i_was_produced_by> ?herstellung . ?herstellung a
<http://erlangen-crm.org/170309/E12_Production> . ?herstellung <http://erlangen-crm.org/170309
/P14_carried_out_by> ?kuenstler . ?werk <http://erlangen-crm.org/170309/P102_has_title> ?titel .
?titel a <http://erlangen-crm.org/170309/E35_Title> . ?titel <http://erlangen-crm.org/170309
/P3_has_note> ?werktitel }} GROUP BY ?kuenstlername ?werktitel
ORDER BY DESC (?anzahl)
```



Query Endpoint

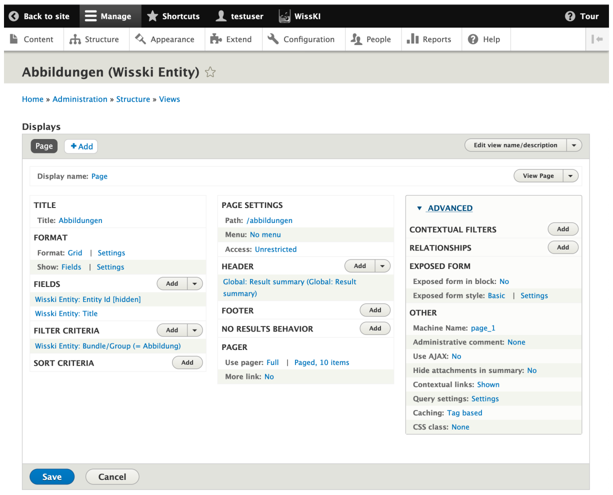
?anzahl	?kuenstlername	?werktitel
"2"^^xsd:integer	"Pieter Brueghel (II)"	" Dorpskermis op het feest van de H. Joris "
"1"^^xsd:integer	"Pieter Brueghel (II)"	"Dorpskermis op het feest van de H. Joris"

Query

```
SELECT (COUNT (?kuenstlername) AS ?anzahl) ?kuenstlername ?werktitel WHERE { GRAPH ?graph {
?kuenstler a <https://kirmes.wisski.agfd.fau.de/ontology/kirmes/kir21a_artist> . ?kuenstler
<http://erlangen-crm.org/170309/P131_is_identified_by> ?name . ?name a <http://erlangen-crm.org
/170309/E82_Actor_Appellation> . ?name <http://erlangen-crm.org/170309/P3_has_note>
?kuenstlername . ?werk a <http://erlangen-crm.org/170309/E22_Man-Made_Object> . ?werk
```

Data Presentation via Views in WissKI

▷ **Example 13.4.3 (Configuring a View).** This makes a drupal block.



Abbildungen (Wisski Entity)

Home » Administration » Structure » Views

Displays

Page | + Add | Edit view name/description

Display name: Page | View Page

TITLE
Title: Abbildungen

FORMAT
Format: Grid | Settings
Show: Fields | Settings

FIELDS (Add)
Wisski Entity: Entity Id [hidden]
Wisski Entity: Title

FILTER CRITERIA (Add)
Wisski Entity: Bundle/Group (= Abbildung)

SORT CRITERIA (Add)

PAGE SETTINGS
Path: /abbildungen
Menu: No menu
Access: Unrestricted

HEADER (Add)
Global: Result summary (Global: Result summary)

FOOTER (Add)

NO RESULTS BEHAVIOR (Add)

PAGER
Use pager: Full | Paged, 10 items
More link: No

▼ ADVANCED

CONTEXTUAL FILTERS (Add)

RELATIONSHIPS (Add)

EXPOSED FORM
Exposed form in block: No
Exposed form style: Basic | Settings

OTHER
Machine Name: page_1
Administrative comment: None
Use AJAX: No
Hide attachments in summary: No
Contextual links: Shown
Query settings: Settings
Caching: Tag based
CSS class: None

Save | Cancel

Drupal generates a SPARQL query, aggregates results into a block.

This Research is WissKI-instance-local

▷ **Observation 13.4.4.** *All these research queries only work in the current WissKI instance.*

▷ **Observation 13.4.5.** *There is probably much more about the entities you are interested in outside your particular WissKI instance.*

▷ **Problem:** How to make use of this?

▷ **Solution:** We need to do two things

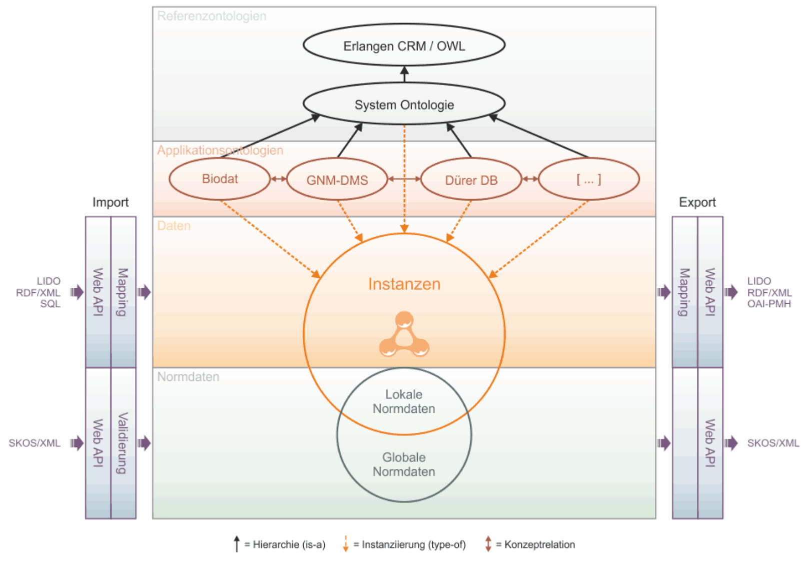
1. Make use of other people's ABoxes
2. Provide your ABox to other people.

This practice is called linked open data. (up next)

13.5 Application Ontologies in WissKI

WissKI Information Architecture (Ontologies)

▷ Ontologies, instances, and export formats

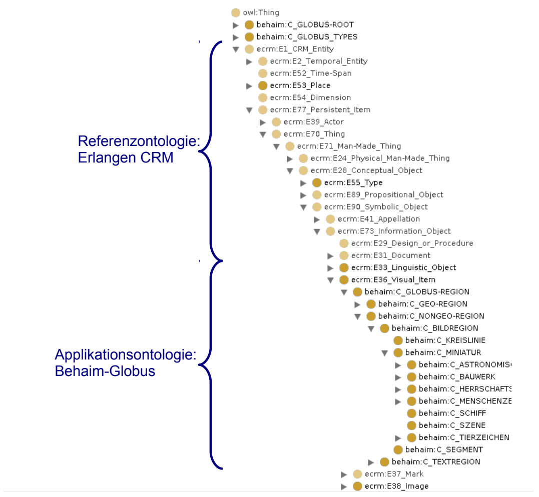


Application Ontologies extend CIDOC CRM

▷ **Observation 13.5.1.** *Sometimes we need more than CIDOC CRM.*

▷ **Definition 13.5.2.** A WissKI application ontology is one that extends CIDOC CRM, without changing it.

▷ **Example 13.5.3 (Behaim Application Ontology).**



```
owl:Thing
▶ behaim:C_GLOBUS-ROOT
▶ behaim:C_GLOBUS_TYPES
▼ ecrm:E1_CRM_Entity
  ▶ ecrm:E2_Temporal_Entity
    ecrm:E52_Time-Span
  ▶ ecrm:E53_Place
    ecrm:E54_Dimension
  ▼ ecrm:E77_Persistent_Item
    ▶ ecrm:E39_Actor
    ▼ ecrm:E70_Thing
      ▼ ecrm:E71_Man-Made_Thing
        ▶ ecrm:E24_Physical_Man-Made_Thing
        ▼ ecrm:E28_Conceptual_Object
          ▶ ecrm:E55_Type
          ▶ ecrm:E89_Propositional_Object
          ▼ ecrm:E90_Symbolic_Object
            ▶ ecrm:E41_Appellation
            ▼ ecrm:E73_Information_Object
                ecrm:E29_Design_or_Procedure
              ▶ ecrm:E31_Document
              ▶ ecrm:E33_Linguistic_Object
              ▼ ecrm:E36_Visual_Item
                ▼ behaim:C_GLOBUS-REGION
                  ▶ behaim:C_GEO-REGION
                  ▼ behaim:C_NONGEO-REGION
                    ▼ behaim:C_BILDREGION
                        behaim:C_KREISLINIE
                      ▼ behaim:C_MINIATUR
                        ▶ behaim:C_ASTRONOMIS
                        ▶ behaim:C_BAUWERK
                        ▶ behaim:C_HERRSCHAFTS
                        ▶ behaim:C_MENSCHENZE
                          behaim:C_SCHIFF
                          behaim:C_SZENE
                        ▶ behaim:C_TIERZEICHEN
                        behaim:C_SEGMENT
                    ▶ behaim:C_TEXTREGION
                ▶ ecrm:E37_Mark
                ▶ ecrm:E38_Image
```

Making an Application Ontology

- ▷ The "current ontology" of a WissKI instance can be configured via the config bar via the "WissKI ontology" module.
- ▷ The application ontology should import CIDOC CRM.
- ▷ **Idea:** Use Protégé for that.

13.6 The Linked Open Data Cloud

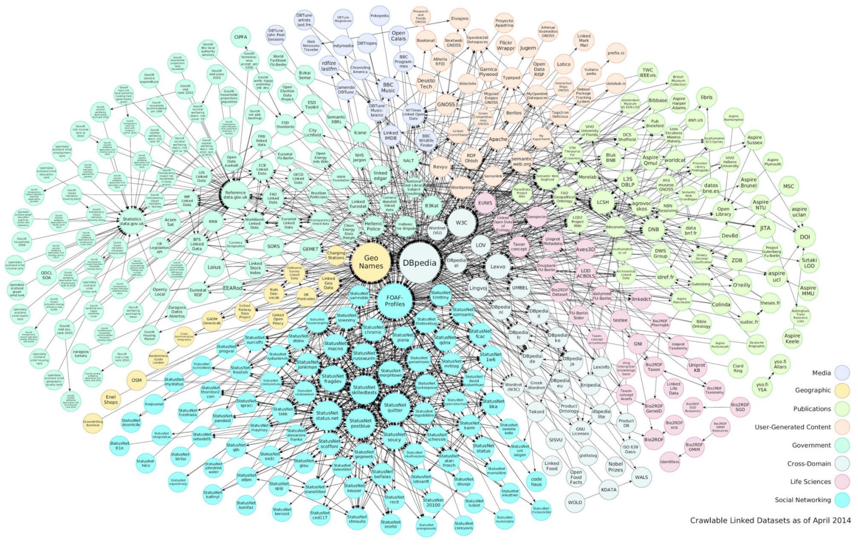
Linked Open Data

- ▷ **Definition 13.6.1.** Linked data is structured data in which classified objects are interlinked via relations with other objects so that the data becomes more useful through semantic queries and access methods.
- ▷ **Definition 13.6.2.** Linked open data (LOD) is linked data which is released under an open license, which does not impede its reuse by the community.
- ▷ **Definition 13.6.3.** Given the semantic web technology stack, we can create interoperable ontologies and interlinked data sets, we call their totality the linked open data cloud.
- ▷ **Recall the LOD Incentives:**
  - ▹ incentivize other authors to extend/improve the LOD $\rightsquigarrow$ more/better data can be generated at a lower cost.
  - ▹ generate *attention* to the LOD and recognition for authors $\rightsquigarrow$ this gives alternative revenue models for authors.

By **??** the linked open data cloud is the totality of linked open data that has been published. [LOD] tracks (the larger parts of) it. This gives us a sense of the extent of this giant network of knowledge expressed as triples.

The Linked Open Data Cloud

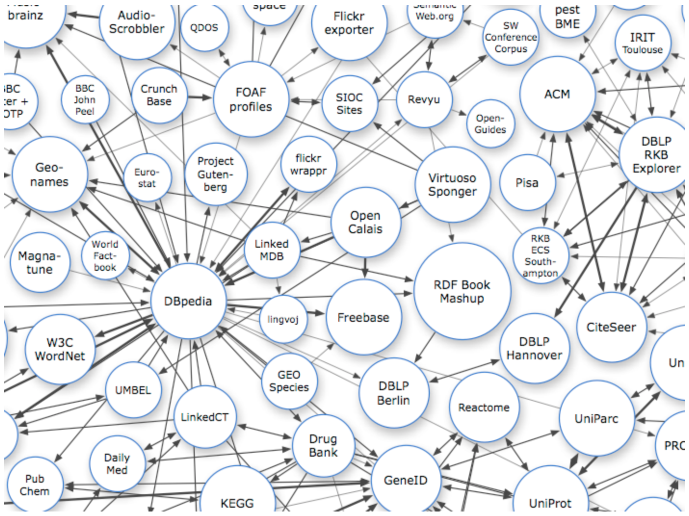
- ▷ The linked open data cloud in 2014 (today much bigger, but unreadable)



We now “zoom in” on this picture to get a better sense”. Each of the circles in the picture is a data set of at least 1000 triples. The DBPedia in the center of this fragment has 3 billion triples alone (in 2014).

The Linked Open Data Cloud

▷ Zooming in (data sets and their – interlinked – ontologies)



The ideas of the linked open data cloud directly apply knowledge about cultural artefacts as we formalize them in the WissKI system: we can directly reference objects from the cloud in WissKI.

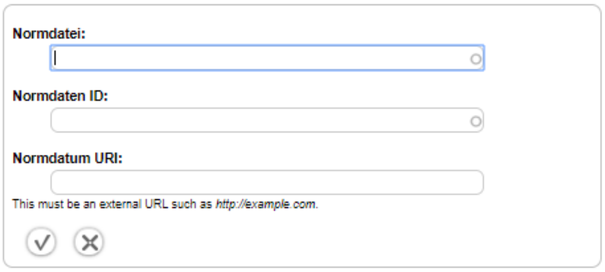
Using the LOD-Cloud in WissKI

- **Idea:** Do not re-model entities that already exist (in the LOD Cloud)
- **Problem:** Most of the LOD Cloud is about things we do not want.
- But there are some sources that are useful
  - the GND (Gemeinsame Normdatei [GND]), an authority file for personal/corporate names and keywords from literary catalogs,
  - geonames[GN], a geographical database with more than 25M names and locations
  - Wikipedia
- **Observation 13.6.4.** *All of them provide URIs for real world entities, which is just what we need for objects in RDF triples.*
- **Definition 13.6.5.** WissKI provides special modules called adapters for GND and geonames.

Using linked open data in WissKI actually makes for higher-quality digitizations, as they are more interoperable. Unfortunately, WissKI only supports the two adapters we mention above. There are many many more that would be useful.
Let us now see how to concretely use an adapter, here for the geonames service.

Using Geonames in WissKI (Example)

1. **Example 13.6.6.** We want to use the "Meilwald" (Erlangen) in WissKI.
2. make a sub-ontology groups "norm data" in the WissKI path builder
3. The induced sub-bundle looks like this:

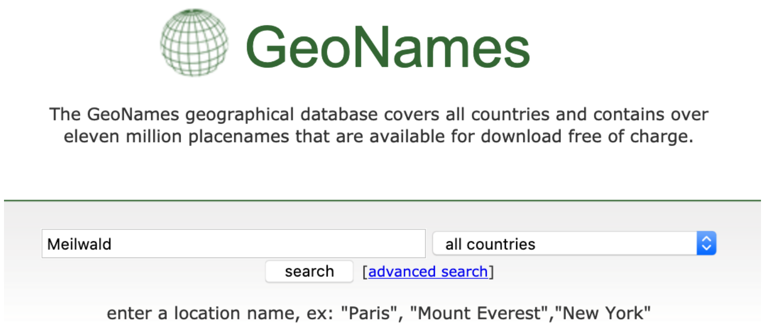

Normdatei:

Normdaten ID:

Normdatum URI:

This must be an external URL such as *http://example.com*.

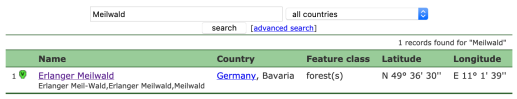
4. We enter `https://geodata.org` for "Normdatei" and go there to find out the URI for "Meilwald" which goes into "Normdatum URI".



The GeoNames geographical database covers all countries and contains over eleven million placenames that are available for download free of charge.

enter a location name, ex: "Paris", "Mount Everest","New York"

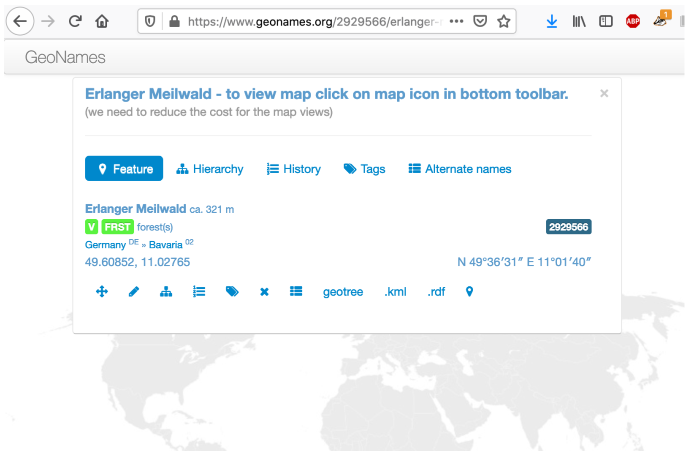
5. there may be multiple results (here only one)



1 records found for "Meilwald"

	Name	Country	Feature class	Latitude	Longitude
1	Erlanger Meilwald Erlanger Meil-Wald,Erlanger Meilwald,Meilwald	Germany, Bavaria	forest(s)	N 49° 36' 30''	E 11° 1' 39''

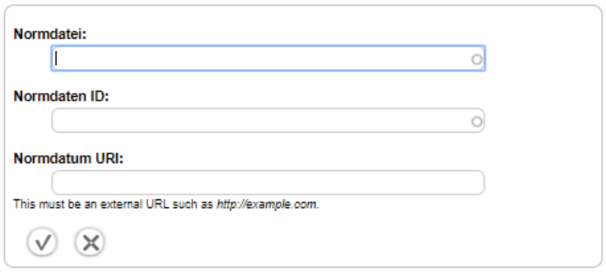
6. Select/click the intended one, check the details



Erlanger Meilwald - to view map click on map icon in bottom toolbar.
(we need to reduce the cost for the map views)

Erlanger Meilwald ca. 321 m
V FRST forest(s) 2929566
Germany DE » Bavaria 02
49.60852, 11.02765 N 49°36′31″ E 11°01′40″

7. Enter the URL from the URL bar into "Normdatum URI".



This must be an external URL such as http://example.com.

If we – as we did here – tell the story of using authority files in WissKI from a linked open data perspective, a curious asymmetry becomes apparent: WissKI is using LOD resources, but is – by and large – not contributing LOD resources back to the "public domain" of linked open cultural heritage data.

Towards a WissKI Commons in the LOD Cloud

- ▷ **Recap:** We can directly refer to (URIs of) external objects in WissKI.
- ▷ **Observation 13.6.7.** *The most interesting source for references to cultural artefacts are other WissKI instances.*
- ▷ **Problem:** A WissKI is an island, unless it exports its data! (few do)
- ▷ **Idea:** We need a LOD cloud of cultural heritage research data under to foster object centric research in the humanities.
- ▷ **Definition 13.6.8.** We call the part of this resource that can be created by aggregating WissKI exports the WissKI commons.
- ▷ **Observation 13.6.9.** *WissKI exports meet the FAIR principles quite nicely already.*
- ▷ We will be working on a FAU WissKI commons in the next years. (help wanted)

This asymmetry is a very serious problem, since cultural heritage research is not profiting as much from digitizations as it could. Keeping data in WissKI silos – this is what we do when we are not exporting WissKI data and referencing objects from other WissKI instances – leads to fragmentation of the research community and to duplication of work.

Chapter 14

Legal Foundations of Information Technology

In this chapter, we cover a topic that is a very important secondary aspect of our work as knowledge workers that – at best – create immaterial things: the legal foundations that regulate how the fruits of our labor are appreciated (and – importantly – recompensated), and what we have to do to respect people's personal data. ⚠ The content of this chapter are about legal matters, but are written by a computer scientist, i.e. not a legal expert. They should considered as an introduction of the fundamental concepts involved, and definitely not as legal advice. For that, contact an intellectual property lawyer.
That being said, we expect that understanding the concepts covered in this chapter will help you with getting most out of this conversation.

14.1 Intellectual Property

The first complex of questions centers around the assessment of the products of work of knowledge/information workers, which are largely intangible, and about questions of recompensation for such work.

Intellectual Property: Concept

- ▷ **Question:** Intellectual labour creates (intangible) objects, can they be owned?
- ▷ **Answer:** Yes: in certain circumstances they are property like tangible objects.
- ▷ **Definition 14.1.1.** The concept of intellectual property motivates a set of laws that regulate property rights rights on intangible objects, in particular
  - ▷ Patents grant exploitation rights on original ideas.
  - ▷ Copyrights grant personal and exploitation rights on expressions of ideas.
  - ▷ Industrial design rights protect the visual design of objects beyond their function.
  - ▷ Trademarks protect the signs that identify a legal entity or its products to establish brand recognition.
- ▷ **Intent:** Property like treatment of intangibles will foster innovation by giving individuals and organizations material incentives.

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 427 2025-02-06

To understand intellectual property better, let us recap the concepts of property and ownership in general.

Background: Property and Ownership in General

- ▷ **Definition 14.1.2.** Ownership is the state or fact of exclusive rights and control over property, which may be a physical object, land/real estate or intangible object.
- ▷ **Definition 14.1.3.** Ownership involves multiple rights (the property rights), which may be separated and held by different parties.
- ▷ **Definition 14.1.4.** There are various legal entities (e.g. persons, states, companies, associations, ...) that can have ownership over a property p. We call them the owners of p.
- ▷ *Remark 14.1.5.* Depending on the nature of the property, an owner of property has the right to consume, alter, share, redefine, rent, mortgage, pawn, sell, exchange, transfer, give away or destroy it, or to exclude others from doing these things, as well as to perhaps abandon it.
- ▷ *Remark 14.1.6.* The process and mechanics of ownership are fairly complex: one can gain, transfer, and lose ownership of property in a number of ways.

These concepts are the basis for many other concepts such as money, trade, debt, bankruptcy, and the criminality of theft. Ownership is the key building block in the development of the capitalist socio-economic system, must influentially developed in Adam Smith's book *An Inquiry into the Nature and Causes of the Wealth of Nations* [Smi76] from 1776.
Naturally, many of the concepts are hotly debated. Especially due to the fact that intuitions and legal systems about property have evolved around the more tangible forms of properties that cannot be simply duplicated and indeed multiplied by copying them. In particular, other intangibles like physical laws or mathematical theorems cannot be property.

Intellectual Property: Problems

- ▷ **Delineation Problems:** How can we distinguish the product of human work, from "discoveries", of e.g. algorithms, facts, genome, algorithms. (not property)
- ▷ **Philosophical Problems:** The implied analogy with physical property (like land or an automobile) fails because physical property is generally rivalrous while intellectual works are non-rivalrous (the enjoyment of the copy does not prevent enjoyment of the original).
- ▷ **Practical Problems:** There is widespread criticism of the concept of intellectual property in general and the respective laws in particular.
  - ▷ (Software) patents are often used to stifle innovation in practice. (patent trolls)
  - ▷ Copyright is seen to help big corporations and to hurt the innovating individuals.

We will not go into the philosophical debates around intellectual property here, but concentrate on the legal foundations that are in force now and regulate IP issues. We will see that groups holding alternative views of intellectual properties have learned to use current IP laws to their advantage and have built systems and even whole sections of the software economy on this basis. Many of the concepts we will discuss here are regulated by laws, which are (ultimately) subject to national legislative and juridicative systems. Therefore, none of them can be discussed without an understanding of the different jurisdictions. Of course, we cannot go into particulars here, therefore we will make use of the classification of jurisdictions into two large legal traditions to get an overview. For any concrete decisions, the details of the particular jurisdiction have to be checked.

Legal Traditions

- ▷ The various legal systems of the world can be grouped into "traditions".
- ▷ **Definition 14.1.7.** Legal systems in the common law tradition are usually based on case law, they are often derived from the British system.
- ▷ **Definition 14.1.8.** Legal systems in the civil law tradition are usually based on explicitly codified laws (civil codes).
- ▷ As a rule of thumb all English-speaking countries have systems in the common law tradition, whereas the rest of the world follows a civil law tradition.

Another prerequisite for understanding intellectual property concepts is the historical development of the legal frameworks and the practice how intellectual property law is synchronized internationally.

Historic/International Aspects of Intellectual Property Law

- ▷ **Early History:** In late antiquity and the middle ages IP matters were regulated by royal privileges
- ▷ **History of Patent Laws:** First in Venice 1474, Statutes of Monopolies in England 1624, US/France 1790/1...
- ▷ **History of Copyright Laws:** Statue of Anne 1762, France: 1793, ...
- ▷ **Problem:** In an increasingly globalized world, national IP laws are not enough.
- ▷ **Definition 14.1.9.** The Berne convention process is a series of international treaties that try to harmonize international IP laws. It started with the original Berne convention 1886 and went through revision in 1896, 1908, 1914, 1928, 1948, 1967, 1971, and 1979.
- ▷ The World Intellectual Property Organization Copyright Treaty was adopted in 1996 to address the issues raised by information technology and the internet, which were not addressed by the Berne Convention.
- ▷ **Definition 14.1.10.** The Anti Counterfeiting Trade Agreement (ACTA) is a multinational treaty on international standards for intellectual property rights enforcement.
- ▷ With its focus on enforcement ACTA is seen my many to break fundamental human information rights, criminalize FLOSS.

14.2 Copyright

In this section, we go into more detail about a central concept of intellectual property law: copyright is the component most of IP law applicable to the individual computer scientist. Therefore a basic understanding should be part of any CS education. We start with a definition of what works can be copyrighted, and then progress to the rights this affords to the copyright holder.

Copyrightable Works

▷ **Definition 14.2.1.** A copyrightable work is any artefact of human labor that fits into one of the following eight categories:

- ▷ Literary works: Any work expressed in letters, numbers, or symbols, regardless of medium. (computer source code is also considered to be a literary work.)
- ▷ Musical works: Original musical compositions.
- ▷ Sound recordings of musical works. (different licensing)
- ▷ Dramatic works: literary works that direct a performance through written instructions.
- ▷ Choreographic works must be “fixed,” either through notation or video recording.
- ▷ Pictorial, graphic and sculptural work (PGS works): Any two dimensional or three dimensional art work
- ▷ Audiovisual works: work that combines audio and visual components. (e.g. films, television programs)
- ▷ Architectural works. (copyright only extends to aesthetics)

▷ The categories are interpreted quite liberally (e.g. for computer code).

▷ There are various requirements to make a work copyrightable: it has to

- ▷ exhibit a certain originality. (“Schöpfungshöhe”)
- ▷ require a certain amount of labor and diligence. (“sweat of the brow” doctrine)

In short almost all products of intellectual work are copyrightable, but this does not mean copyright applies to all those works. Indeed there is a large body of works that are “out of copyright”, and can be used by everyone. Indeed it is one of the intentions of intellectual property laws to increase the body of intellectual resources a society a draw upon to create wealth. Therefore copyright is limited by regulations that limit the duration of copyright and exempts some classes of works from copyright (e.g. because they have already been paid for by society).

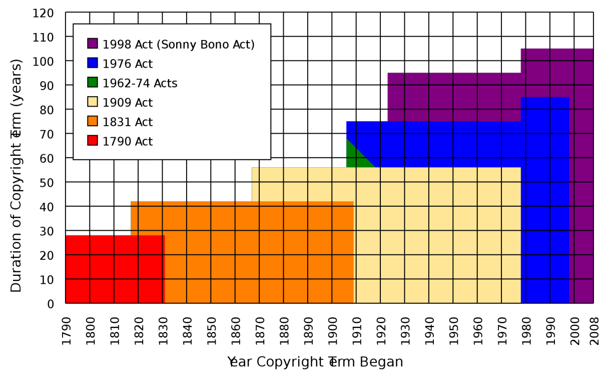
Limitations of Copyrightabilitiy: The Public Domain

▷ **Definition 14.2.2.** A work is said to be in the public domain, if no copyright applies, otherwise it is called copyrighted.

▷ **Example 14.2.3.** Works made by US government employees (in their work time) are in the public domain directly. (Rationale: taxpayer already paid for them)

▷ **Copyright expires:** usually 70 years after the death of the creator.

▷ **Example 14.2.4 (US Copyright Terms).** Some people claim that US copyright terms are extended, whenever Disney's Mickey Mouse would become public domain.



Now that we have established, which works are copyrighted — i.e. to which works are intellectual property, we now turn to the rights owning such a property entails.

Rights under Copyright Law

▷ **Definition 14.2.5.** The copyright is a collection of rights on a copyrighted work;

▷ Personal rights: the owner of the copyright may

▷ determine whether and how the work is published (right to publish)

▷ determine whether and how her authorship is acknowledged. (right of attribution)

▷ to object to any distortion, mutilation or other modification of the work, which would be prejudicial to his honor or reputation. (droit de respect)

▷ Exploitation rights: the owner of a copyright has the exclusive right to do, or authorize to do any of the following:

▷ to reproduce the copyrighted work in copies (or phonorecords);

▷ to prepare derivative works based upon the copyrighted work;

▷ to distribute copies of the work to the public by sale, rental, lease, or lending;

▷ to perform the copyrighted work publicly;

▷ to display the copyrighted work publicly; and

▷ to perform the copyrighted work publicly by means of a digital-audio transmission.

▷ *Remark 14.2.6.* Formally, it is not the copyrightable work that can be owned itself, but the copyright.

▷ **Definition 14.2.7.** The use of a copyrighted material, by anyone other than the owner of the copyright, amounts to copyright infringement only when the use is such that it conflicts with any one or more of the exclusive rights conferred to the owner of the copyright.

Initially, and by default the copyright of an intellectual work is owned by the creator. But – as with any property – copyrights can be transferred. We will now go into the details.

Copyright Holder

- **Definition 14.2.8.** The copyright holder is the legal entity that owns the copyright to a copyrighted work.
- By default, the original creator of a copyrightable work holds the copyright.
- In most jurisdictions, no registration or declaration is necessary. (but copyright ownership may be difficult to prove in court)
- Copyright is considered intellectual property, and can be transferred to others. (e.g. sold to a publisher or bequeathed)
- **Definition 14.2.9 (Work for Hire).** A work made for hire (WFH) is a work created by an employee as part of his or her job, or under the explicit guidance or under the terms of a contract.
- **Observation 14.2.10.** *In jurisdictions from the common law tradition, the copyright holder of a WFH is the employer, in jurisdictions from the civil law tradition, the author, unless the respective contract regulates it otherwise.*

Again, the rights of the copyright holder are mediated by usage rights of society; recall that intellectual property laws are originally designed to increase the intellectual resources available to society.

Limitations of Copyright (Citation/Fair Use)

- There are limitations to the exclusivity of rights of the copyright holder. (some things cannot be forbidden)
- **Citation Rights:** Civil law jurisdictions allow citations of (extracts of) copyrighted works for scientific or artistic discussions. (note that the right of attribution still applies)
- In the civil law tradition, there are similar rights:
- **Definition 14.2.11 (Fair Use/Fair Dealing Doctrines).** Case law in common law traditions has established a fair use doctrine, which allows e.g.
  - making safety copies of software and audiovisual data,
  - lending of books in public libraries,
  - citing for scientific and educational purposes, or
  - excerpts in search engine.

Fair use is established in court on a case-by-case taking into account the purpose (commercial/educational), the nature of the work the amount of the excerpt, the effect on the marketability of the work.

14.3 Licensing

Given that intellectual property law grants a set of exclusive rights to the owner, we will now look at ways and mechanisms how usage rights can be bestowed on others. This process is called licensing, and it has enormous effects on the way software is produced, marketed, and consumed. Again, we will focus on copyright issues and how innovative license agreements have created the open source movement and economy.

Licensing: the Transfer of Rights

▷ **Remember:** The copyright holder has exclusive rights to a copyrighted work.

▷ **In particular:** All others have only fair use rights. (but we can transfer rights)

▷ **Definition 14.3.1.** A license is an authorization (by the licensor) to use the licensed material (by the licensee).

▷ **Note:** a license is a regular contract (about intellectual property) that is handled just like any other contract. (it can stipulate anything the licensor and licensees agree on) in particular a license may

- ▷ involve term, territory, or renewal provisions,
- ▷ require paying a fee and/or proving a capability, or
- ▷ require to keep the licensor informed on a type of activity, and to give them the opportunity to set conditions and limitations.

▷ **Mass Licensing of Computer Software:** Software vendors usually license software under extensive end user license agreement (EULA) entered into upon the installation of that software on a computer. The license authorizes the user to install the software on a limited number of computers.

Copyright law was originally designed to give authors of literary works — e.g. novelists and playwrights — revenue streams and regulate how publishers and theatre companies can distribute and display them so that society can enjoy more of their work.

With the inclusion of software as “literary works” under copyright law the basic parameters of the system changed considerably:

- modern software development is much more a collaborative and diversified effort than literary writing,
- re-use of software components is a decisive factor in software,
- software can be distributed in compiled form to be executable which limits inspection and re-use, and
- distribution costs for digital media are negligible compared to printing.

As a consequence, much software development has been industrialized by large enterprises, who become copyright holder as the software was created as work for hire. This has led to software quasi-monopolies, which are prone to stifling innovation and thus counteract the intentions of intellectual property laws.
The Free/Open Source Software movement attempts to use the intellectual property laws themselves to counteract their negative side effects on innovation and collaboration and the (perceived) freedom of the programmer.

Free/Libre/Open-Source Licenses

- ▷ **Recall:** Software is treated as literary works wrt. copyright law.
- ▷ **But:** Software is different from literary works wrt. distribution channels. (and that is what copyright law regulates)
- ▷ **In particular:** When literary works are distributed, you get all there is, software is usually distributed in binary format, you cannot understand/cite/modify/fix it.
- ▷ **So:** Compilation can be seen as a technical means to enforce copyright. (seen as an impediment to freedom of fair use)
- ▷ **Recall:** IP laws (in particular patent law) was introduced explicitly for two things:
  - ▷ incentivize innovation, (by granting exclusive exploitation rights)
  - ▷ spread innovation. (by publishing ideas and processes)

  Compilation breaks the second tenet! (and may thus stifle innovation)
- ▷ **Idea:** We should create a public domain of source code.
- ▷ **Definition 14.3.2.** Free/Libre/Open Source Software (FLOSS or just open source) is software that is and licensed via licenses that ensure that its source code is available.
- ▷ Almost all of the internet infrastructure is (now) FLOSS; so are the Linux and Android operating systems and applications like OpenOffice and The GIMP.

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 438 2025-02-06

The relatively complex name Free/Libre/Open Source comes from the fact that the English[1] word "free" has two meanings: free as in "freedom" and free as in "free beer". The initial name "free software" confused issues and thus led to problems in public perception of the movement. Indeed Richard Stallman's initial motivation was to ensure the freedom of the programmer to create software, and only used cost-free software to expand the software public domain. To disambiguate some people started using the French "libre" which only had the "freedom" reading of "free". The term "open source" was eventually adopted in 1998 to have a politically less loaded label.

The main tool in brining about a public domain of open source software was the use of licenses that are cleverly crafted to guarantee usage rights to the public and inspire programmers to license their works as open source systems. The most influential license here is the GNU public license which we cover as a paradigmatic example.

GPL/Copyleft: Creating a FLOSS Public Domain?

[1]the movement originated in the USA

▷ **Problem:** How do we get people to contribute source code to the FLOSS public domain?

▷ **Idea:** Use special licenses to:

- ▷ allow others to use/fix/modify our source code and (derivative works)
- ▷ require them to release modifications to the FLOSS public domain if they do.

▷ **Definition 14.3.3.** A copyleft license is a license which requires that allows derivative works, but requires that they be licensed with the same license.

▷ **Definition 14.3.4.** The General Public License (GPL) is a copyleft license for FLOSS software originally written by Richard Stallman in 1989. It requires that the source code of GPL-licensed software be made available.

▷ The GPL was the first copyleft license to see extensive use, and continues to dominate the licensing of FLOSS software.

▷ FLOSS based development can reduce development and testing costs. (but community involvement must be managed)

▷ Various software companies have developed successful business models based on FLOSS licensing models. (e.g. Red Hat, Mozilla, IBM, ...)

Note:
that the GPL does not make any restrictions on possible uses of the software. In particular, it does not restrict commercial use of the copyrighted software. Indeed it tries to allow commercial use without restricting the freedom of programmers. If the unencumbered distribution of source code makes some business models (which are considered as "extortion" by the open-source proponents) intractable, this needs to be compensated by new, innovative business models. Indeed, such business models have been developed, and have led to an "open-source economy" which now constitutes a non-trivial part of the software industry.
With the great success of open source software, the central ideas have been adapted to other classes of copyrightable works; again to create and enlarge a public domain of resources that allow re-use, derived works, and distribution.

Open Content/Data via Open Licenses

▷ **Recall:** FLOSS licenses have created a vibrant public domain for software.

▷ **How about:** (not so different from software)

- ▷ other copyrightable works: musics, videos, literatures, technical documents.
- ▷ data (including research data).

▷ **Idea:** Adapt the FLOSS license ideas to the particular domain $X \rightsquigarrow$ open X.

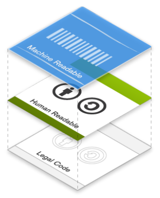
- ▷ Open content: pictures, music, video, documents, ... $\rightsquigarrow$ Creative Commons
- ▷ Open data: data from science, government, and organizations, ...
  $\rightsquigarrow$ Open Data Commons [ODC].
- ▷ Open licenses for many other domains X.

- ▷ **Why open communities grow:** Open X licenses give strong incentives to join: they
  - ▷ incentivize other authors to extend/improve the X
    $\rightsquigarrow$ more/better X can be generate at a lower cost.
  - ▷ generate attention to the X andrecognition for authors
    $\rightsquigarrow$ this gives alternative revenue models for authors.
- ▷ **Open X Slogan:** Publish X early, publish X often!

We will now discuss the probably most prominent example of a system of "open X licenses": the Creative Commons licenses. This system of licenses has been adapted from the software-oriented licenses by some of the most prominent IP lawyers of their time.

Creative Commons a System of Open Content Licenses

▷ **Definition 14.3.5.** The Creative Commons license are
  - ▷ a common legal vocabulary for sharing content
  - ▷ to create a kind of "public domain" using licensing
  - ▷ presented in three layers (human/lawyer/machine)-readable



▷ **Definition 14.3.6.** The CC licenses stipulate that (cf. `http://www.creativecommons.org`)
  - ▷ Creators retain the copyright on their works.
  - ▷ Creators license their works to the world with under the CC provisions:

+/- attribution	(must reference the author)
+/- commercial use	(can be restricted)
+/- derivative works	(can allow modification)
+/- share alike (copyleft)	(modifications must be donated back)

The Creative Commons licenses are continually gaining traction, as they give copyright holders strong secondary incentives (and the moral high ground). Correspondingly, the Creative Commons of freely usable works is continually growing, which is exactly what the CC licenses were created for.

14.4 Information Privacy

The last big topic in this chapter is information privacy. This affects us in IWGS in a different way than the previous ones. As providers of information systems we are subject to regulations that require us to keep user's personal data private to the extent possible and keep inform users informed of what happens to it.

Information/Data Privacy

- **Definition 14.4.1.** The principle of information privacy (also called data privacy) comprises the idea that humans have the right to control who can access their personal data.
- Information privacy concerns exist wherever personal data is collected and stored – in digital form or otherwise. In particular in the following contexts:
  - healthcare records,
  - criminal justice investigations and proceedings,
  - financial institutions and transactions,
  - biological traits, such as ethnicity or genetic material, and
  - residence and geographic records.
- Information privacy is becoming a growing concern with the advent of the internet and web search engines that make access to information easy and efficient.
- The "reasonable expectation of privacy" is regulated by special laws.
- These laws differ considerably by jurisdiction; The EU has particularly stringent regulations. (and you are subject to these.)
- **Intuition:** Acquisition and storage of personal data is only legal for the purposes of the respective transaction, must be minimized, and distribution of personal data is generally forbidden with few exceptions. Users have to be informed about collection of personal data.

The legal basis for information privacy at least for the EU – is the GDPR, the most current information privacy legislation. We will go into the details in the next couple of slides.

The General Data Protection Regulation (GDPR)

- **Definition 14.4.2.** The General Data Protection Regulation (GDPR) is a EU regulation created in 2016 to harmonize information privacy regulations within Europe.

  The GDPR applies to data controllers, i.e organizations that process personal data of EU citizens (the data subjects).
- **Remark:** The GDPR sanctions violations to its mandates with substantial punishments up to 20€ or 4% of annual worldwide turnover.
- *Remark 14.4.3.* As an EU regulation, the GDPR is directly effective in all EU member countries. (enforced since 2018)
- **Axiom 14.4.4.** *The GDPR applies to data controllers outside the EU, iff they*

  1. *offer goods or services to EU citizens, or*
  2. *monitor their behavior.*

Organizational Measures for Information Privacy (GDPR)

- ▷ **Definition 14.4.5.** Physical access control: Unauthorized persons may not be granted physical access to data processing equipment that process personal data. (⇝ locks, access control systems)
- ▷ **Definition 14.4.6.** System access control: Unauthorized users may not use systems that process personal data. (⇝ passwords, firewalls, . . .)
- ▷ **Definition 14.4.7.** Information access control: Users may only access those data they are authorized to access. (⇝ access control lists, safe boxes for storage media, encryption)
- ▷ **Definition 14.4.8.** Data transfer control: Personal data may not be copied during transmission between systems. (⇝ encryption)
- ▷ **Definition 14.4.9.** Input control: It must be possible to review retroactively who entered, changed, or deleted personal data. (⇝ authentication, journaling)
- ▷ **Definition 14.4.10.** Availability control: Personal data have to be protected against loss and accidental destruction. (⇝ physical/building safety, backups)
- ▷ **Definition 14.4.11.** Obligation of separation: Personal data that was acquired for separate purposes has to be processed separately.

SOME RIGHTS RESERVED

Personally Data (GDPR)

- ▷ **Definition 14.4.12.** A person is called identifiable if it can be identified by a direct identifier (e.g., passport information) that can identify a person uniquely, or a combination of one or more quasi-identifiers, i.e. factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that allow to recognize that person; we call such a combination identifying.
- ▷ **Definition 14.4.13.** We collectively call direct identifiers and identifying collections of quasi-identifiers personally identifying information (PII).
- ▷ **Example 14.4.14.** Quasi-identifiers include name, date of birth, race, location, . . .
- ▷ **Definition 14.4.15.** Personal data (also called personal information) is any information relating to an identified or identifiable person.
- ▷ **Example 14.4.16.** The color name "red" by itself is not personal data, but stored as part of a data subject's record as their "favorite color" is personal data; it is the connection to the person that makes it personal data, not the value itself.
- ▷ **Axiom 14.4.17.** *Under the GDPR, any personal data a site collects must be either anonymized, i.e. PII deleted, or pseudonymized (with the data subject's PII consistently replaced with aliases).*
- ▷ **Intuition:** With pseudonymization data controllers can still do data analysis that would be impossible with anonymization.

Customer-Service Requirements (GDPR)

▷ Visitors must be notified of data the site collects from them and explicitly consent to that information-gathering. (This site uses cookies $\rightsquigarrow$ Agree)

▷ Data controllers must notify data subjects in a timely way (72h) if any of their personal data held by the site is breached.

▷ The data controller needs to specify a data-protection officer (DPO).

▷ Data subjects have the right to have their presence on the site erased.

▷ Data subjects can request the disclosure all data the data controller collected on them. (if the request is in writing, the answer must be on paper)

14.5 Exercises

Problem 5.1 (Problems with Intellectual Property)
State two problems of treating intangibles as (intellectual) property.

Problem 5.2 (CopyLeft)
Briefly state the copyleft clause in the GNU Public License or in the Creative Commons licenses, and explain how it works.

Problem 5.3 (Public Domain)

1. When do we speak of a work as being "in the pubic domain"?
2. state a use of a work that would not be allowed if it was licensed under the GNU General Public License (GPL) instead of being in he public domain.

Chapter 15

Collaboration and Project Management

To facilitate group work – both for the IWGS-II project and future projects down the line, we will start off the semester by looking at state-of-the art project and content management systems and directly use that in the project.

We will concentrate on two parts of such a system:

- collaborative, versioned document/program development via GIT (see **??**)
- issue tracking and management via GitHub/GitLab (**??**).

Systems like GitLab or GitHub also offer additional features like developer communication, continuous integration, automated deployment, monitoring and security management (collectively called DevOps) which are way beyond the scope of IWGS.

15.1 Revision Control Systems

We address a very important topic for project management: supporting the life-cycle of project documents, data, and software in a collaborative process. In this section we discuss how we can use a set of tools that have been developed for supporting collaborative development of large program collections can be used for general project artefact management.

We will first introduce the problems and attempts at solutions and the introduce two classes of revision control systems and discuss their paradigmatic systems.

15.1.1 Dealing with Large/Distributed Projects and Document Collections

In this subsection we will look at problems in managing the artefacts of large projects. Such projects range from technical documentation for complex systems over knowledge collections like the Wikipedia, to software collections like the Linux kernel. They have in common that a *large group of authors/developers* manage a *large artefact collection* over a *long period of time*.

Web Development Scenario

▷ **Example 15.1.1.**

1. Your boss told you to develop an interactive website.
2. You already have an early prototype.

3. You have a great idea for a new feature and you want to surprise your boss with an even better prototype, so you have worked on it for two days.

▷ **Problem 1:** When you present it to your boss, she only wants the basics done. What do you do? **Idea 1:** You make a copy of your file, store it away and delete the feature from your current document. **Problem 2:** What if you worked on the html, css and the .js files for the new feature? **Idea 2:** You make a copy of your folder, store it away and delete the feature from all your current documents. **Problem 3:** What if you finished the basics and now your boss wants the cool feature? **Idea 3:** You go to the stored-away folder, search for the code fragments of the feature and you copy them over to the newest version of your files. **Problem 4:** What if your boss notices that you need help programming and employs someone? **Idea 4:** Your colleague will get a copy of your latest folder and both of you work on the project. At some point you will join the most current files and the most current code fragments. **Problem 5:** Let's say that you use dropbox for collaboration.

- ▷ What if your colleague introduced a bug?
- ▷ What if your colleague deleted a file by accident?

Intuition: Sharing is fine, (bug) tracking not, backup is also not possible on a broad scale.

How do we collaborate?

▷ Direct collaboration (the human-to-human aspect)

- ▷ meetings for brainstorming/conflict management
- ▷ calls for current hot problem solving

▷ Indirect, artefact-based collaboration (the system aspect)

- ▷ mails, messages, reports, links, . . . , code fragments

▷ **Idea:** Support by artefact-based collaboration by a computer system:

- ▷ Communication management
- ▷ Project management via issue tracking
- ▷ Local and distributed change management

▷ Such systems are called revision control systems a.k.a. RCS.

Collaboration Support by RCS

▷ **Revisions:** A revision control system (RCS) copies snapshots of all project changes in files/subfolders for you.

▷ **Control:** A RCS helps you control all collaborators's revisions over time.

- ▷ Complexity is hidden
- ▷ Tools for browsing your project history
- ▷ Tools for collaborating in a project

▷ **System:**

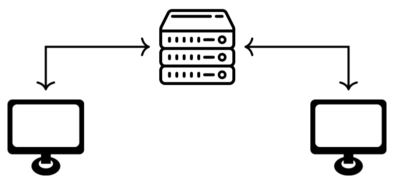
- ▷ Repository $\widehat{=}$ collection of all revisions + special information (order, what, who) for a project.
- ▷ You decide on which changes count toward a version e.g. code fragments in index.html and style.css for one feature, but not your list of passwords.
- ▷ Committing $\widehat{=}$ the act of telling the RCS that you are finished (for now).

Architecture of Revision Control Systems

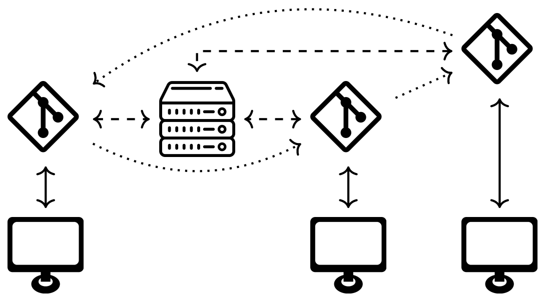
▷ **Observation:** We distinguish three large classes of RCS.

▷ In local RCS, a working copy uses a repository on the same machine.

▷ In a centralized RCS, the repository is on a central repository server.



▷ In a distributed RCS, working copy, use local repositories, which can communicate change to the web server or other local repositories.



▷ We will go through these in explaining the respective features as we go along.

GIT as a Revision Control System for IWGS

- ▷ GIT is a powerful distributed revision control system.
- ▷ GIT is the current dominant RCS, exceeding 90% adoption in open source projects and high utilization in industry.
- ▷ GIT features a well-designed set of primitive revision control actions, from which complex behaviours can be composed.
- ▷ **In particular,**
  the GIT revision control actions can implement local, centralized, and distributed revision control.
- ▷ We use GIT as the model for revision control systems in IWGS.

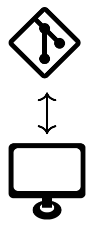
15.1.2 Local Revision Control: Versioning

We start out with the basics of local revision control systems. This architecture essentially provides access to past versions with minimal hassle.

We first introduce the concepts and then make them concrete using the git system, which we will use throughout the IWGS course.

Revision Control Systems

- ▷ **Definition 15.1.2.** A revision control system (RCS) a software system that tracks the change process of a document collection via a federation of repositories. Each step in the development history is called a revision.
- ▷ **Definition 15.1.3.** In a RCS, users do not directly work on the repository, but on a working copy that is synchronized with the repository.
- ▷ **Definition 15.1.4.** A local RCS supports the following revision control actions:
  1. initialize: creates a new repository with empty head revision (a.k.a. head).
  2. checkout: given a revision identifier – by default the head creates a new working copy from the repository.
  3. add: places a file in the working copy under control of the RCS.
  4. commit: transmits the differences between the head and the working copy to the repository, which patches the head.



- ▷ **Observation 15.1.5.** *The user's commits determine the revisions in a RCS.*
- ▷ **Remark:** Revision control systems usually store the head revision explicitly and can compute development histories via reverse diffs.

?? and ?? are very general, so that they can cover a wide variety of architectures.

Don't drink and write code!: RCS even allow to checkout to a specific revision that is not the head, e.g. if an author wants to base her work on that – or wants to revert some changes. In fact, most RCS support branching: committing different development lines to the repository, but we will not go into this here and leave the discussion for later when we discuss distributed revision control systems where branching is the main mechanism of operation. Before we become more concrete, let us have a look at the basic ingredient of revision control systems: computing differences, applying them to documents, and reconciling differences.

Computing and Managing Differences with `diff` & `patch`

▷ **Definition 15.1.6.** `diff` is a file comparison utility that computes differencess between two strings or text files: the source f_1 and the target f_2. Differencess are output linewise in a differences $\delta(f_1, f_2)$.

▷ **Definition 15.1.7.** `patch` is a sister utility that applies a differences $\delta := \delta(f_1, f_2)$ to f_1 – resulting in f_2; we say it patches f_1 with δ.

▷ **Example 15.1.8.** We compare two simple text files:

The quick brown	The quack brown	1c1,2
fox jumps over		< The quick brown
the lazy dog	fox jumps over	---
	the loozy dog	> The quack brown
		>
		3c4
		< the lazy dog

		> the loozy dog

▷ **Definition 15.1.9.** A differences consists of a sequence of hunks that in turn consist of a locator which indicates the source line number followed by the lines deleted in the source and added in the target.

FAU Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 453 2025-02-06

In practice, – unlike in our didactic example – differencess are usually (much) smaller than either the source or the target. This makes the design decision of passing around differencess instead of files in revision control systems efficient.

15.1.3 GIT as a local Revision Control System

Now that we understand the concepts, let us see how we can use them in practice. For this we assume that students have installed GIT on their computers, so that they can use it; [CS14, section 1.5] gives an excellent introduction. For this subsection, we explain GIT workflows for local revision control, e.g. for a single user who wants to keep track (and revive) past versions of their code or document collections.

We we explain GIT functionality "from scratch", and do not presuppose a repository management system.

Working with GIT

▷ **Observation:** GIT can be used in many situations.

▷ **On your Laptop:** for software development

- ▷ Download GIT from `https://git-scm.com/downloads`, install (you want to use it on your local machine)
- ▷ We will use GIT from the shell on your system (`macOS` or `linux`) or `GitBash`, a shell that comes with your GIT download (`MSWindows`). (graphical front ends exist but often hinder understanding)
- ▷ Test whether your installation works: git version

▷ **In jupyterLab:** For the IWGS homeworks.

- ▷ You can use the jupyterLab terminal (the resident shell)
- ▷ There is a visual GIT integration into jupyterLab, see the GIT logo ◆ on the left.

SOME RIGHTS RESERVED

In all of our concrete examples, we will use `UNIX` shell commands; for `MSWindows` users should use the GIT shell, a GIT enhanced version of the `UNIX` shell that comes with the GIT distribution, and *not* the `MSWindows` command prompt. There are graphical front-ends for the GIT client, but our experience shows that using shell commands helps understand the concepts and workflows much better.

Working with GIT (Initializing a Local Repository)

▷ Download GIT from `https://qgit-scm.com/downloads`, install (you want to use it on your local machine)

▷ We will use git from the shell on your system (`macOS` or `linux`) or `GitBash` that comes with your GIT download (`MSWindows`).(graphical front ends exist but hinder understanding)

▷ Test whether your installation works: git version (should be ≥ 2.30)

▷ **Definition 15.1.10.** git init initializes a local repository:

- ▷ git init turns the current directory into a GIT working copy by adding a local repository as a hidden .git folder.
- ▷ git init ⟨⟨name⟩⟩ makes working copy + local repository in the ⟨⟨name⟩⟩ subdirectory.

SOME RIGHTS RESERVED

We will now come to a GIT peculiarity that is important to understand for working with GIT: Often we only want to commit only a subset of the changed files – e.g. because the changes already constitute a achievement of their own or we want to split the development into multiple commits. There are essentially two ways of achieving this.

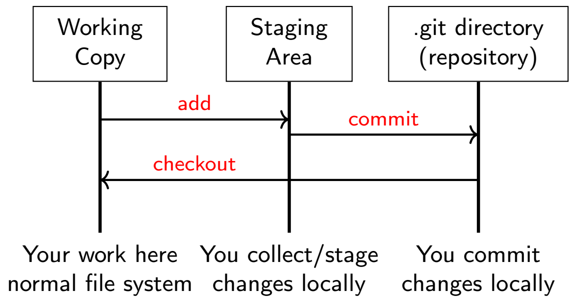
1. giving the commit action a list of files to be committed, or

2. marking files for a future commit this is called staging.

The second method is more flexible, since we do not have to remember which files participate in a commit and we can stage files as we go along. Therefore GIT uses this method, even though it adds conceptual complexity – actually, the first method can be recovered by syntactic sugar.

Working with GIT (Staging and Committing)

▷ **Overview:** GIT local workflow: staging files for commit using git add


Working Copy
Staging Area
.git directory (repository)
add
commit
checkout
Your work here normal file system
You collect/stage changes locally
You commit changes locally

commits acts only on staged files ⇝ git add foo.tex (GIT must know about them)

Working with GIT (Staging and Committing)

▷ Basic GIT commands: (many variants and options ⇝ study them)

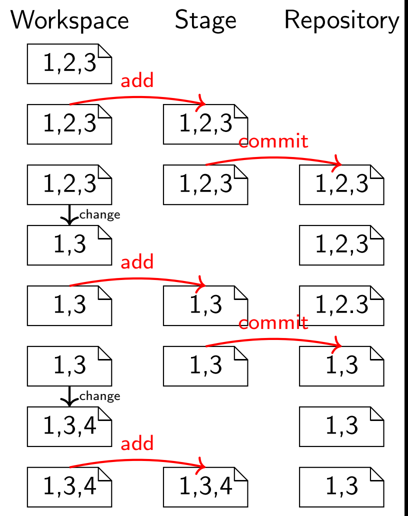
git add ⟨⟨file/dir⟩⟩	stages a file or directory ⟨⟨file/dir⟩⟩
git add ——all	stages all changes in the current folder
git reset HEAD ⟨⟨file/dir⟩⟩	unstages ⟨⟨file/dir⟩⟩
git commit —m'⟨⟨msg⟩⟩'	commits staged files with commit message ⟨⟨msg⟩⟩
git status	gives information about the working copy.

We have only shown the most basic commands here. There are many other commands an options that make your life much easier. For instance, the —a option is very useful for git commit: it automatically stages all the changed files. git commit —am'foo' commits all your change in the current directory (which is often what you want).
Let us now fortify our intuition on working with GIT by exhibiting a typical (but elementary) workflow.

An Example Git Workflow

▷ **Example 15.1.11.** A typical, elementary workflow in GIT in a shell.

```
> git init
Initialized empty Git repository in /tmp
> echo "1,2,3" > test.txt
> git add test.txt
> git commit —m'initializing'

> echo "1,3" > test.txt
> git status
On branch master
Changes not staged for commit:
  (use "git add <file>..." to update ...
  (use "git checkout —— <file>..." to...
        modified: test.txt
no changes added to commit
(use "git add" and/or "git commit —a")

> git add test.txt
> git commit —m'bla' test.txt
> echo "1,3,4" > test.txt
> git add test.txt
```



Note that the shell command echo ⟨⟨string⟩⟩ > ⟨⟨file⟩⟩ updates the contents of the file ⟨⟨file⟩⟩ to ⟨⟨string⟩⟩ or creates ⟨⟨file⟩⟩ with this content in the first place. We use this command to make the file changes visible in the shell on the left side.

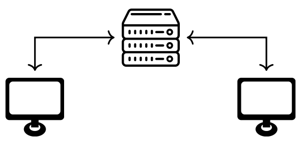
15.1.4 Centralized Revision Control: Collaboration

With this, we can now understand the revision control worfklows in our concrete system. In its simplest form, a revision control system, can be understood using the Subversion system that is heavily used in open source projects that have a relatively hierarchical development model.

Collaboration via Centralized RCS

▷ **Definition 15.1.12.** A centralized revision control system features

- ▷ a single, central repository server (for current revision and reverse diffs)
- ▷ local working copies (asynchronous checkouts, updates, commits)



They are kept synchronized by passing around diffs and patching the repository and working copies. Conflicts are resolved by (three-way) merge.

The revision control actions are those of a local RCS plus

- ▷ clone: fetch the current revision from repository server and checkout a new working copy.
- ▷ pull: fetch the pending differences between the revision of the working copy and the revision of the repository server and merges them into the working copy.

- ▷ push: if the working copy and the repository are based on the same revision, then transmit the differences to the repository server and update the revision there.

fetch and push are dual operations. Just as fetch is integrated into the pull, push is usually integrated into commit for centralized RCS.

For revision control systems we need more than just `diff` and `patch`. When we are sending around diffs along non-linear development histories, then we also have to reconcile diffs that come via different paths.

Merging Differences

▷ There are basically two ways of merging the differences of files into one.

▷ **Definition 15.1.13.** In two way merge, an automated procedure tries to combine two different files by copying over differences by guessing or asking the user.

▷ **Definition 15.1.14.** In a three way merge the files are f_1 and f_2 are assumed to be created by changing a joint original (the parent) p by editing.

If there are hunks h_1 in $\delta(f_1,p)$ and h_1 in $\delta(f_2,p)$ that affect the same line in p, then we call the pair (h_1,h_2) a conflict.

The result of a three way merge are two diffs $\mu^3{}_i(f_1,f_2,p)$, which contain the non-conflicting differences of $\delta(f_i,p)$ and (representations called conflict markers of) the conflicts.

▷ **Note:** In revision control systems conflicts must be resolved by choosing one of the alternatives or creating a manually merged revision before changes can be commited.

Merging Differences with `merge3`

▷ **Definition 15.1.15.** The `merge3` tool computes a three way merge.

▷ **Example 15.1.16.** We compare two simple text files with a parent:

mine.txt	your.txt	parent.txt	conflict marker
This is the file. Hello	This is the file. hello	This is the file. hi	This is the file. <<<<<<< mine.txt Hello ||||||| parent.txt hi ======= hello >>>>>>> your.txt

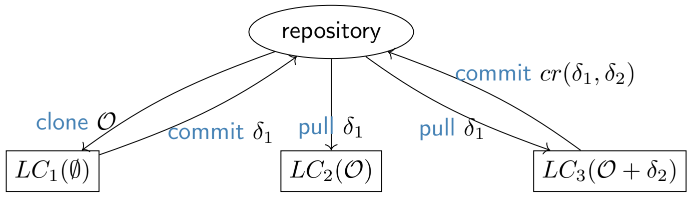
▷ **Remark:** The conflict markers in actual RCSs are similar, but may vary.

▷ **Note:** There are good visual merge3 tools that help you cope with merges. Some text editors also have support for resolving conflict markers.

▷ **Remark:** There are analoga to `diff` and `patch` for other file formats, but in practice, revision control is mostly restricted to text files.

Collaboration via Centralized RCS (Example)

▷ **Example 15.1.17 (A Workflow with three Working Copies).**



In the workflow of **??** is a typical one:

1. A first user checks out a new working copy LC_1, from the empty repository, adds a couple of files – we denote the new document collection at this point with $\mathcal{O}$, and commits the difference δ_1 between the working copy and $\mathcal{O}$ to the repository which δ_1 logs it as "revision 1".

2. There is another repository LC_2, which has been checked out earlier (i.e. based on "revision 0"), and which is now no longer in sync with the repository. So we can pull (i.e. patch) it to "revision 1" by transferring δ_1 to LC_2, which thus has same content as LC_1, namely $\mathcal{O}$.

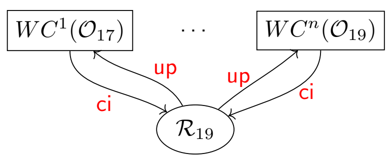
3. For a third repository LC_3 which has been checked out at "revision 0" we assume that it has been changed by adding different files, the difference being δ_2. Note that as these changes are relative to "revision 0", they cannot simply be committed to the repository. Therefore we need to pull it. As LC_3 already contains changes, this amounts to a merge of δ_1 and δ_2 to get a new local copy that is essentially $\mathcal{O}+\delta_2$, which is now relative to "revision 1". This can now be committed to the repository to form "revision 2".

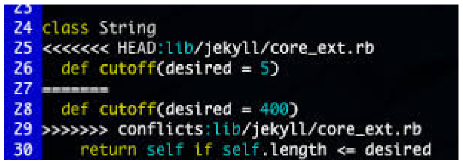
Note: that in all of this it does not matter who the authors of the respective changes and the owners of the respective working copies are. They might be different persons, or a single author might have multiple working copies, e.g. one one the work computer, one on a laptop, and one on the home desktop. They are all held in sync by pulls, commits.
With this basic mechanism, we can already model quite complex and collaborative workflows. The basic idea is simple: we just use the pull/commit cycle to synchronize a set of working copies.

Collaboration via Revision Control

▷ **Idea:** We can use revision control for collaboration with multiple working copies.

▷ **Diff-Based Collaboration:** Centralized RCS takes care of the synchronization:





```
class String
<<<<<<< HEAD:lib/jekyll/core_ext.rb
  def cutoff(desired = 5)
=======
  def cutoff(desired = 400)
>>>>>>> conflicts:lib/jekyll/core_ext.rb
    return self if self.length <= desired
```

- ▷ you can only commit, if your revision is the head (otherwise update)
- ▷ update merges the changes into your working copy.
- ▷ If there are changes on the same line, you have a conflict, which must be resolved.

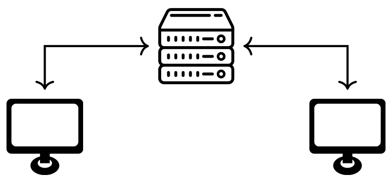
Note: that these collaborative workflows can be asynchronous. In particular working copies can lag behind the repository as long as they want – they only have to synchronize for commits. This gives a lot of freedom in the development process.
Also note: that unless the repository and the working copies are on the same computer which is somewhat unlikely. Commits and updates are only possible while online, this sometimes prevents authors/developers from grouping changes logically as they have to collect them until they are online again.

15.1.5 GIT as a centralized RCS

In this subsection, we cover GIT-based collaborative workflows for centralized revision control, as they occur in small collaborative projects, where a simple centralized structure suffices. Again, we we explain GIT functionality "from scratch" without using a repository management system.

Recap: Centralized RCS

▷ **Idea:** In a centralized RCS, the repository resides on a repository server.



▷ **Problem:** We need some generalizations over local RCS:

- ▷ Identifying the repository server.
- ▷ Pushing and fetching over the network.

Working with GIT (Remote Repositories)

▷ **Recap:** A repository can be connected to one or several remote repositories.

▷ GIT commands for working with remote repositories:

git remote add ⟪name⟫ ⟪URI⟫	gives the repos at ⟪URI⟫ the name ⟪name⟫
git remote	shows names of all remote repositories

▷ git remote –v shows the remote repositories e.g.

```
MiKo:collaboration kohlhase$ git remote –v
origin https://gitlab.cs.fau.de/iwgs–ss19/collaboration.git (fetch)
origin https://gitlab.cs.fau.de/iwgs–ss19/collaboration.git (push)
```

▷ git remote add ⟪name⟫ ⟪URI⟫ adds remote repositories e.g.

```
kohlhase$ git remote add upstream git@gl.kwarc.info:test/collab.git
kohlhase$ git remote –v
origin https://gitlab.cs.fau.de/iwgs–ss19/collaboration.git (fetch)
origin https://gitlab.cs.fau.de:iwgs–ss19/collaboration.git (push)
upstream https://gl.kwarc.info:test/collab.git (fetch)
upstream https://gl.kwarc.info:test/collab.git (push)
```

▷ We can now pull/push to the new remote repository, e.g. git push upstream master

▷ **Note:** git push is just syntactic sugar for git push origin master

Before you start, you should configure some global options for GIT to make your life easier and the documentation of your interactions on the repository server more systematic.

Configuring GIT on your Computer

▷ **Background:** Configuration sets sensible defaults. (Saves you typing)

▷ **Definition 15.1.18.** git config allows to view and set configuration options

git config ––list	shows configuration
git config ⟪key⟫ ⟪value⟫	sets config option ⟪key⟫ to value ⟪value⟫

▷ **Example 15.1.19 (Name and E-Mail).**

```
git config ––global user.name "John Doe"
git config ––global user.email johndoe@example.com
```

▷ **Example 15.1.20 (Default Repositories and Branches).** Always pull the branch called main from the repository called origin.

```
git config branch.master.remote origin
git config branch.master.merge refs/heads/master
```

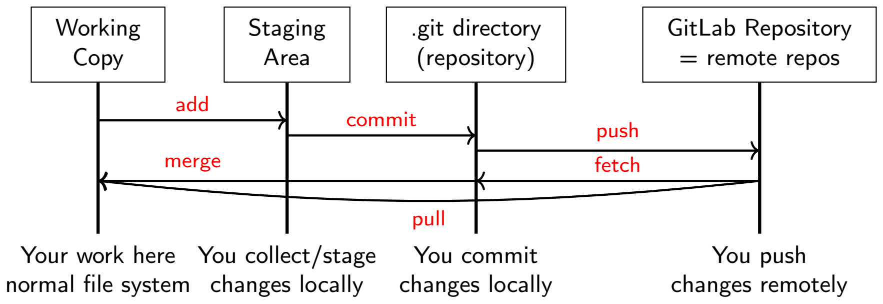
Replace git push origin master with a simple git push.

Working with Remote Repositories: Pushing and Pulling

▷ GIT commands for working with remote repositories

git clone ⟪URI⟫	clones the repos at ⟪URI⟫
git push ⟪repos⟫ ⟪branch⟫	pushes all commits to branch ⟪branch⟫ on ⟪repos⟫
git pull ⟪repos⟫ ⟪branch⟫	fetches and merges branch ⟪branch⟫ from ⟪repos⟫

▷ **Overview:** GIT centralized workflow: pushing and pulling to a remote repository


Working Copy
Staging Area
.git directory (repository)
GitLab Repository = remote repos
add
commit
push
fetch
merge
pull
Your work here normal file system
You collect/stage changes locally
You commit changes locally
You push changes remotely

Working with GIT (Cloning a Remote Repository)

▷ **Alternative:** Clone a remote repository, i.e. git init + git pull

```
git clone https://gitlab.cs.fau.de/iwgs−ss19/collaboration.git
Cloning into 'collaboration'...
Username for 'https://gitlab.cs.fau.de': yp70uzyj
Password for 'https://yp70uzyj@gitlab.cs.fau.de':
...
```

15.1.6 Distributed Revision Control

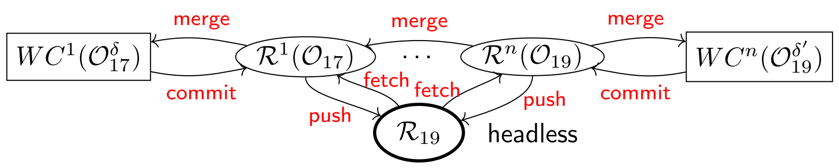
We now introduce distributed revision control systems using the GIT system as an example.

Distributed Version Control

▷ **Problems with Centralized Revision Control:**

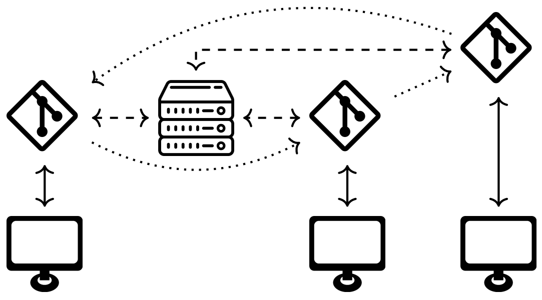
1. We can only commit when online! (but we work on the train)
2. All collaboration goes via one, central repository. (prescribes workflow)

▷ **Idea:** Distribute the repositories and move patches between them.


merge
merge
merge
$WC^1(\mathcal{O}^{\delta}_{17})$
$\mathcal{R}^1(\mathcal{O}_{17})$
...
$\mathcal{R}^n(\mathcal{O}_{19})$
$WC^n(\mathcal{O}^{\delta'}_{19})$
commit
fetch
fetch
push
commit
push
$\mathcal{R}_{19}$
headless

1. local commits to local repositories

2. all repositories created equal (flexible organization)



▷ **Definition 15.1.21.** We call a revision control system distributed, iff it allows multiple repositories that can exchanged patches.

▷ **Definition 15.1.22.** We call a repository headless (or bare), if used without a working copy.

▷ **Observation:** Putting a headless repository onto a web server, yields a repository server.

The concept of distributed revision control systems is motivated by the two shortcomings at the top of the slide, which can be remedies by a single – if relatively radical idea: allowing lots of repositories that can communicate with each other by exchanging patches. Local repositories allow commits while offline and distributed repositories allow for flexible architectures.

We now come to the most prominent of the distributed revision control system: GIT. It implements the concepts motivated above. Somewhat paradoxically, the distributed nature of the workflows makes it simpler and more efficient to implement.

Distributed Version Control with GIT

▷ **Definition 15.1.23.** GIT is a distributed revision control system that features

- ▹ local repositories for each working copy.
- ▹ multiple remote repositories connected to a local repository
  - ▹ clone a remote repository $\leadsto$ make local repository+working copy
  - ▹ local repository changes can be fetched from and pushed to a remote repository (the upstream/downstream repositories).
- ▹ branches and forks (remote upstream repository)

▷ **Software Support:** Facilitates working with GIT:

- ▹ GitHub, a repository hosting service at `http://GitHub.com` (free public/private repositories)
- ▹ GitLab, an open source repository management system and repository hosting service at `http://GitLab.com` (free public/private repositories)

15.1.7 Working with GIT in large Projects

In this subsection, we will (further) discuss the concepts for using GIT in large, long-lived projects. This is less important for IWGS, since projects are rather small. But we want to at least make students aware of GIT branching and the GIT flow paradigm, and we want to clear up the mystery of which GIT often speaks of master.

We can now come back to the topic, where GIT really shines: branching. The main reason for this is that merging is so well supported in GIT. Together with the distributed "local-repository" architecture, this allows for very flexible organization of workflows. We will discuss the basics of branch-based and fork-based workflows here.

GIT Branches and Forks

▷ GIT special commands for making, switching, and merging branches.

git branch ⟨⟨branch⟩⟩	makes a branch with name ⟨⟨name⟩⟩
git checkout ⟨⟨branch⟩⟩	switches a working copy to branch ⟨⟨branch⟩⟩
git branch −v	shows all branches
git branch −d ⟨⟨branch⟩⟩	deletes branch ⟨⟨branch⟩⟩

▷ **Intuition:** In GIT branches are very similar to repositories, but more lightweight.
Repositories can have different permissions; branches inhert these.

▷ **Fork-based Collaboration:** If you want to contribute to a repository $\mathcal{R}$ you have no push-rights on,

1. clone $\mathcal{R}$ to a new repository $\mathcal{R}'$ you own (i.e. fork it; $\mathcal{R}'$ is a fork of $\mathcal{R}$)
2. develop your contribution on $\mathcal{R}'$.
3. ask $\mathcal{R}$s owners to pull from $\mathcal{R}'$ (pull request)

GIT repository management systems like GitHub and GitLab support this.

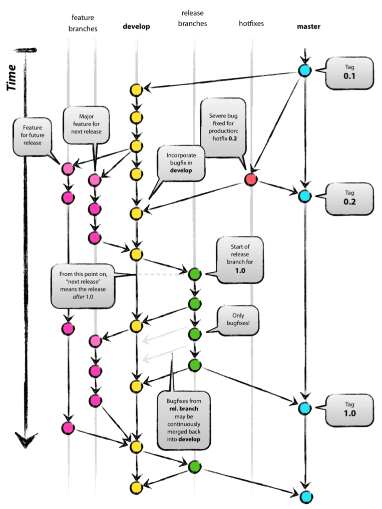
What we have seen above, let us briefly discuss an elaborate workflow suitable for large development teams, which has become known under the name "GitFlow".

GitFlow: An Elaborate Development Model based on GIT

▷ **Definition 15.1.24 (Development Model).** [Dri10] suggests GIT flow, which includes:

- ▷ A main branch called main that all other branches merge into.
- ▷ New functionality is developed "feature-by-feature" on feature branches.
- ▷ A development branch (usually called devel) that integrates all feature branches and is merged into master once the integrated functionality is stable.
- ▷ (possibly) release branches for every release; they collect bugfixes, but no new features.

▷ Most large software development projects adopt aspects of GIT flow.


feature branches
develop
release branches
hotfixes
master
Time
Tag 0.1
Feature for future release
Major feature for next release
Severe bug fixed for production: hotfix 0.2
Incorporate bugfix in develop
Tag 0.2
Start of release branch for 1.0
From this point on, "next release" means the release after 1.0
Only bugfixes!
Bugfixes from rel. branch may be continuously merged back into develop
Tag 1.0

15.2 Working with GIT and GitLab/GitHub

In principle we know all we need for running GIT in practice. But if we want to make use of remote repositories – and without that, we lack most of the advantages of revision control systems – we have to deploy a web server which takes on the upstream repositories.

Even though this is relatively simple to set up, there are now dedicated web applications that supply repositories and additional project managment infrastructure.

Working with GitLab/GitHub

▷ GIT it sufficient to set up a remote repository. (but tedious [CS14, chapter 4])

▷ **Idea:** Use a GIT repository manager like GitLab/GitHub (we use GitLab)

▷ **Definition 15.2.1.** A repository management system is an web application that supports the administration of a repository server and manages authentication and authorization.

▷ **Example 15.2.2.** GitLab is an open source repository management system and repository hosting service at `http://GitLab.com`.(free public/private repositories)

▷ **Definition 15.2.3.** A repository hosting service is a web based repository management system that also offers storage space for repositories.

▷ **Example 15.2.4.** GitHub is a repository hosting service at `http://GitHub.com`. (free public repositories)

GitHub is now the default hosting service for open source software development, it hosts more than 190 Million repositories (March 2020).

We could be using GitHub for IWGS – and we would probably do so for an open-source software proejct – but we will use the FAU offering: a GitLab instance that offers repository hosting to all

FAU members and log in via FAU IDM. The instructors of IWGS have installed a special group for repository hosting.

Working with GitLab/GitHub (continued)

▷ **Definition 15.2.5.** Often, repository management systems organize repositories (called projects in GitLab) hierarchically into groups (also called namespace) and provide a personal group to all users.

▷ **Concretely:** we use the FAU GitLab: `https://gitlab.cs.fau.de`

1. sign in with the FAU Single Sign On (aka. FAU IDM account)
2. this makes an account there and gives you a personal group `https://gitlab.cs.fau.de/`⟪SSID⟫
3. IWGS has a course group `https://gitlab.cs.fau.de/iwgs-ss19`(the course project goes there)
4. ⚠ Note that the SSO credentials are *only* for log in! You will have to set a password (or upload an SSH Key, see below) seperately to push. *Using the SSO credentials for authentication during push will not work!* ⚠

Now we are ready to play with GitLab, and please do, there is nothing you can do wrong. And – that is the beauty of revision control systems – few things you cannot undo.

Making Repositories on GitLab

▷ Make a new project with ⊞⌄, play with it (you can always delete it)

▷ **Definition 15.2.6.** Group/project visibility can be one of three states:

- ▷ Private: Project access must be granted explicitly to each user.
- ▷ Internal: The project can be accessed by any authenticated user.
- ▷ Public: The project can be accessed without any authentication.

Private and public make most sense in our setting.

▷ **Exercise:** Make a repository, clone it locally, add a file to it, commit that, let your friends clone/change/commit it, merge their changes, ... (see the homework)

Using GitLab for the IWGS Project

▷ Make a in a member

To understand what these visibility levels mean, we have to talk about authorization in GitLab, i.e. how we can manage what interactions a particular (class of) user is allowed to do.

Authorization in GitLab: Managing Access Permissions

▷ **Definition 15.2.7.** Authorization refers to a set of rules that determine who is allowed to do what.

▷ **Definition 15.2.8.** Authorization is often operationalized by assigning permission levels and binding the authorization to execute particular interactions to permission levels.

▷ **Definition 15.2.9.** GitLab has five permission levels for repositories:

1. guests can clone and see/report issues . . .
2. reporters can also assign issues . . .
3. developers can also push, create branches . . .
4. maintainers can also assign permission levels . . .
5. owners can also delete repository . . .

▷ **Intuition:** In a public repository, everyone is guest, in a internal one, logged in users are.

15.3 Excursion: Authentication with SSH

We now come to a topic that is of practical relevance, whenever we work with web applications that work with restricted resources – in this case the content of your private repositories: authentication. Generally, there are two authentication methods: the one via passwords built into HTTPS and ssh-authentication, which we will briefly discuss here, since it is the more convenient method for interacting with GitLab (and GitHub).

Before we come to ssh-authentication, let us clarify the concept of authentication in general.

Authentication

▷ **Definition 15.3.1.** Authentication is the process of ascertaining that somebody really is who they claim to be.

▷ **Definition 15.3.2.** Authentication can be performed by assertaining an authentication factor, i.e. testing for something the user

- ▹ *knows*, e.g. a password or answer to a security question – kwowledge factor
- ▹ *has*, e.g. an ID card, key, implanted device, software token, – ownership factorx
- ▹ *is* or *does*, e.g. a fingerprint, retinal pattern, DNA sequence, or voice – inheritance factor.

▷ **Note:** Password-based authentication is known to be problematic. (and you have to remember/type it)

▷ **One Problem:** Server and user must both know the password to authenticate passwords are symmetric keys: the server can leak them.

SOME RIGHTS RESERVED

We now come to an authentication method that leaves the user out of the loop completely. It workd via cryptographic keys, which are exchanged between the GIT client and server. In this particular setup, we make use of public key cryptography, which only transfers public keys and keeps the private keys local; minimizing the user of passwords and leakage.

The details of this are quite involved, so we only give a very brief introduction of the moving parts.

Authentication by Cryptographic Public Keys

▷ **Definition 15.3.3.** Cryptography is the practice of transmitting a plain text t by encoding it into a cipher text t', to hide its content from anyone but the legitimate reciever who can decode t' to t.

▷ **Definition 15.3.4.** Public key cryptography split the key into an encode key e and a decode key d

- ▷ key e can encode a text t to t', but only d can decode t' to t.

▷ **Definition 15.3.5 (Public Key Authentication).** built into the SSH communication protocol.

1. user generates key pair (e,d), deposits d on server as certificate, keeps e secret.
2. user encodes a text t with e to t' send $t+t'$ to server
3. server decodes t' to t'' with d and verifies $t=t'' \rightsquigarrow$ OK, iff $t=t''$.

▷ **Advantage:** Passwords canot be leaked, need not be transmitted, retyped.

SOME RIGHTS RESERVED

In practice, working with SSH-based authentication is quite easy to work with: we have to generate a public/private key pair – there are standard utilities for that, deposit the public key in GitLab, and then use clone using the SSH URI supplied by GitLab.

Working with GIT (Cloning a Remote Repository with SSH)

▷ **Alternative:** Clone a remote repository via SSH URL

```
kohlhase$ git clone git@gitlab.cs.fau.de:iwgs—ss19/collaboration.git
Cloning into 'collaboration'...
remote: Enumerating objects: 12, done.
remote: Counting objects: 100% (12/12), done.
remote: Compressing objects: 100% (5/5), done.
remote: Total 12 (delta 1), reused 0 (delta 0)
Receiving objects: 100% (12/12), done.
Resolving deltas: 100% (1/1), done.
```

▷ **But we need a key pair** for this to work.
Go to `https://gitlab.cs.fau.de/profile/keys` and follow the instructions there

- ▷ **essentially**: generate a key pair, copy one into GitLab.

We will now complement revision control systems, as discussed above, with issue tracking systems. The former support dealing with changes in the collaborative development of document collections, the latter support the collaboratie management of issues the reasons for changes.

15.4 Bug/Issue Tracking Systems

In this section we will discuss issue tracking systems, which support the collaborative management of reports on a particular problem, feature request or general task, as well as its its status and other relevant data. These systems originated from tracking systems for help desks and in software engineering, but have evolved into general project planning systems.

issue tracking systems

We will mainly look at systems that originate from software engineering applications here.

Bug/Issue Tracking Systems

▷ **Definition 15.4.1.** An issue tracker (also called issue tracking system simply bugtracker) is a software application that keeps track of reported issues i.e. software bugs, tasks, and feature requests – in software development projects.

▷ **Example 15.4.2.** There are many open-source and commercial bugtrackers

- ▹ bugzilla: `http://bugzilla.org` (Mozilla's bugtracker)
- ▹ TRAC: `http://trac.edgewall.org` (mostly for Subversion)
- ▹ GitHub: `http://github.com` (probably the most used)
- ▹ GitLab: `http://gitlab.com` (open source version of GitHub)
- ▹ JIRA: `https://www.atlassian.com/software/jira` (proprietary)

▷ Most bugtrackers are web applications and also integrate a wiki and integrate a revision control system via extended markdown.

It is no coincidence that issue trackers often come bundled with revision control systems; they form the perfect complement: while the latter track large digital artefacts over extended development cycles, the issue trackers track the tasks induced by the development over the same time frame. It is natural that the two should be well-synchronized for a successful development project.

Issue trackers manage issues and track their status over its whole lifetime – from the initial report to its resolution. This results in a particular set of components that are present in all systems.

The Anatomy of an Issue

▷ **Definition 15.4.3.** An issue (or bug report) specifies

- ▹ title: a short and descriptive overview (one line)
- ▹ description: a precise description of the expected and actual behavior, giving exact reference to the component, version, and environment in which the bug occurs. (bugs must be reproducible and localizable)
- ▹ issue metadata: who, when, what, why, state, … (see below)

- ▷ conversation: a forum like facility for disussing an issue.
- ▷ attachment: e.g. a screen shot, set of inputs, etc.

▷ **Definition 15.4.4.** A feature request is an issue that only specifies the expected behavior and proposes ways of implementing that.

The conversation of an issue is a lightweight text category, which should be efficient to write, but has some structure to make reading and understanding the concepts and details involved. In particular, it is important to be able to refer to the program code, other issues, other developers, commits, etc.

Most bugtrackers use the markdown format, which strikes a good balance between structure and brevity of markup codes and extend it with bugtrackers specific markup.

We use the opportunity to introduce markdown in general before we come to the extensions.

Markdown a simple Markup Format Generating HTML

.

▷ **Idea:** We can translate between markup formats.

▷ **Definition 15.4.5.** Markdown is a family of markup formats whose control words are unobtrusive and easy to write in a text editor. It is intended to be converted to HTML and other formats for display.

▷ **Example 15.4.6.** Markdown is used in applications that want to make user input easy and efficient, e.g. wikis and issue tracking systems.

▷ **Workflow:** Users write markdown, which is formatted to HTML and then served for display.

▷ A good cheet-sheet for markdown control words can be found at `https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet`.

Instead of introducing the markdown syntax systematically, let us look at an example that shows the most prominent control words in action and see how things look in a markdown-based application (and behind the scenes as HTML).

Markdown a simple Markup Language Generating HTML

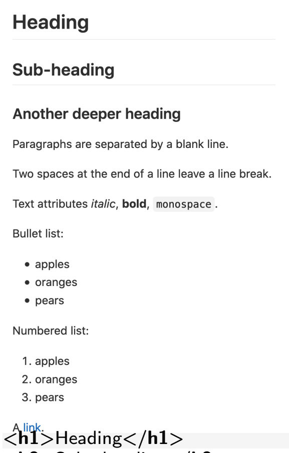
▷ **Example 15.4.7.** We show the most important Markdown commands.

Markdown syntax	Generated HTML

Markdown syntax:

```
# Heading
## Sub—heading
### Another deeper heading

Paragraphs are separated
by a blank line.

Two spaces at the end of a
line leave a line break.

Text attributes _italic_,
**bold**, ‘monospace‘.

Bullet list:
  * apples
  * oranges
  * pears

Numbered list:
  1. apples
  2. oranges
  3. pears

A [link](http://example.com).
```

Generated HTML:



Heading

Sub-heading

Another deeper heading

Paragraphs are separated by a blank line.

Two spaces at the end of a line leave a line break.

Text attributes *italic*, **bold**, `monospace`.

Bullet list:

- apples
- oranges
- pears

Numbered list:

1. apples
2. oranges
3. pears

A link.

```
<h1>Heading</h1>
<h2>Sub—heading</h2>
<h3>Another deeper heading</h3>
<p>Paragraphs are separated by a blank line.</p>
<p>Two spaces at the end of a
  line leave a<br/> line break.</p>
<p>Text attributes <em>italic</em>,
<strong>bold</strong>,
<code>monospace</code>.</p>
<p>Bullet list:</p>
<ul>
<li>apples</li>
<li>oranges</li>
<li>pears</li>
</ul>
<p>Numbered list:</p>
<ol>
<li>apples</li>
<li>oranges</li>
<li>pears</li>
</ol>
<p>A <a href="http://example.com">link</a>.</p>
```

Markdown was originally developed for wikis, and its markup infrastructure reflects that. For use in issue tracking systems, we need to also reference to the program code, other issues, other developers, commits, etc.

GitHub flavored markdown: Tracker Specific Extensions

▷ *Remark 15.4.8.* Source code hosting systems offer special extensions for referencing their components.

▷ **Definition 15.4.9.** GitHub flavored markdown (GFM) is a markdown dialect extended for the use in GIT-based issue tracking systems; see [Gfm] for the specification.

▷ **Example 15.4.10.** GitHub/GitLab recognize most of GFM, most usefully

- ▷ @foo for team members (@all for all project members), e.g. *cc:* @miko
- ▷ #123 for issues, e.g. *depends on* #4711
- ▷ !123 for merge requests, e.g. *but merge* #19 *first*
- ▷ $123 for code snippets, e.g. *see* $123 *for an example usage*
- ▷ 1234567 for commits, e.g. *fixed by* 4c0decb *yesterday*.

- ▷ [file](path/to/file) for file references,
  e.g. *as we see in* [pre.tex](../lib/pre.tex)

▷ **Observation 15.4.11.** *Very useful for project planning and reporting in GitLab and GitHub.*

The anatomy of an issue only enables/restricts the form of an issue, not what would help the project along. We will explore that – to get you thinking – in a counter-example and the show what would have helped the developers.

Issues – How to Write a Good One

▷ The descriptions or issues should be concise, but describe all pertinent aspects of the situation leading to the unexpected behavior.

▷ **Example 15.4.12 (A bad bug report description).** *My browser crashed. I think I was on foo.com. I think that this is a really bad problem and you should fix it or else nobody will use your browser.*

▷ **Example 15.4.13 (A good one).** *I crash each time I go to foo.com (Mozilla build 20000609, Win NT 4.0SP5). This link will crash Firefox reproducibly unless you remove the* `border=0` *attribute:*

```
<IMG SRC="http://foo.com/topicfoos.gif" width=34 border=0 alt="News">
```

▷ **Remember:** Developers are also human (try to minimize their work)
Think about what would help you understand and reproduce the problem.

Let us now survey the typical workflow supported by a issue tracking systems by presenting the typical life-cycle of an issue.

Bugtracker Workflow

▷ **Definition 15.4.14 (Typical Workflow).** supported by all bugtrackers

- ▷ user reports issue (files report in the system)
- ▷ other users extend/discuss/up/downvote issue
- ▷ QA engineer triages issues by classification, remove duplicates, identify dependencies, tie to component, . . . and assign to developer.
- ▷ developer accept or reassigns issue (fixes who is responsible primarily)
- ▷ project planning by identification of sub-issues, dependencies (new issues)
- ▷ bug fixing (design, implementation, testing)
- ▷ issue landing (sign-off, integration into code base)
- ▷ release of the fix (in the next revision)
- ▷ QA engineer or developer closes issue

▷ **Observation 15.4.15.** *An issue tracker can serve as a full blown project planning system, if used accordingly.*

▷ **Definition 15.4.16.** For timing work plans, most issue trackers provide milestones that issues can be targeted to.

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 487 2025-02-06

The workflow presented on the last slide is supported by metadata recorded in the issue, most importantly some kind representation of a issue state.

Administrative Metadata for Issues

▷ To make the issue based workflows work we need data.

▷ **Definition 15.4.17 (Administrative Metadata).** Issue metadata can specify

- ▷ issue number: for referencing with e.g. #15
- ▷ an assignee: a developer currently responsible
- ▷ participants: people who get notified of changes/comments
- ▷ labels: for specializing bug search
- ▷ a state: e.g. one of new, assigned, fixed/closed, reopened.
- ▷ a resolution for fixed bugs, e.g.
  - ▷ FIXED: source updated and tested
  - ▷ INVALID: not a bug in the code
  - ▷ WONTFIX: "feature", not a bug
  - ▷ DUPLICATE: already reported elsewhere; include reference
  - ▷ WORKSFORME: couldn't reproduce issue
- ▷ dependencies: which issues does this one depend on/block?

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 488 2025-02-06

The resolutions can be realized in different ways in different bugtrackers. The ones shown here are hard coded in bugzilla. GitHub and GitLab use a system of developer-definable labels and a set of issue boards which are inspired by Kanban boards to assign and move between states and resolutions.

15.5 Exercises

Problem 5.1 (Make a GitLab Account)
We will use the GitLab instance at `http://gitlab.cs.fau.de` to manage your GIT repositories. Make an account there with your FAU Single Sign On, and set a password on the account.

Problem 5.2 (Make an IWGS Homework Repository)
Make a private repository in your personal group on your FAU GitLab from **??**, clone it on your machine, move all your IWGS homework submissions there, adds them, commits, and pushs.

It is a good idea to keep all the IWGS homework submissions there. We are going to re-use them in the IWGS project.

Problem 5.3 (IWGS-II Project)
During the remainder of the semester, you will be developing an information system for the "Farmer-Fair Pictures" data set.

1. find two other people you want to work with, form a working group (WG), and give the group a name.

2. Make a GitLab repository in the personal group of one of the WG members, and give the other WG members developer or admin permission.

3. add a file README.md that shortly explains the purpose of the repository – two sentences are enough.

Problem 5.4 (Using GitLab)

The aim of this problem is to get your hands dirty in using a revision control system, and to show you that this is actually quite simple. Here are your tasks:

1. Using your account from **??**, clone the repository `https://gitlab.cs.fau.de/IWGS-SS19/collaboration` to get a local working copy.

2. We want to collaboratively build a membership file for the IWGS course, you can find it as `users.txt`. Add your personal information (account name, real name, and e-mail) and commit it. If there are conflicts, you need to resolve them (make sure that you do not delete the information of your peers). Do not forget to give a meaningful commit message and to push your changes.

3. Add a file ⟪account⟫.txt with a friendly note about your experience with IWGS, where ⟪account⟫ is your account name to the users directory and commit/push it.

Problem 5.5 (Issues in GitLab)

The aim of this problem is to get your hands dirty in using a bugtracker, and to show you that this is actually quite simple. Here are your tasks:

1. On the collaboration project from **??** you can find the IWGS course notes as the file notes.pdf for reference. Report an issue about the slides, this can be a bug report or a feature request. Label it appropriately and assign it to one of your peers.

2. Comment on one of the existing issues. If you have been assigned an issue, be sure to comment on that as well.

3. Then add the number of the issue you created and the issues you comment upon, in the line with your data in the file `users.txt`, commit it and push.

Problem 5.6 (Bug Report in GitLab)

In his free time Dr. Leonardo is developing a system for automatic detection of famous buildings in paintings. To aid with the image processing, he is using an open source framework available on Gitlab. Unfortunately for certain images the framework sometimes causes his program to crash when loading images from the hard drive. Therefore Dr. Leonardo decides to file a bug report on the Gitlab page, hoping that the developers fix the problem. By doing some research, he gathers the following information:

1. The function responsible for image loading crashes for images which are higher than wide (i.e. upright images).

2. The bug only seems to appear on Windows 10. On Linux, Mac and older versions of Windows the function works as expected.

3. The problem only occurs for certain image file formats, `jpg` and `gif`, not for `png`.

4. The bug disappears, if he converts the image to grayscale in Photoshop beforehand [1].

5. Other image processing frameworks handle the problematic images just fine.

Help Dr. Leonardo formalize all these facts in a bug report.

Problem 5.7 (Markdown)

Being happy with his newly acquired bug report writing skills, Dr. Leonardo now wants to make it nicer in order to better catch the developer's attention. He found out that he can use a markdown language to achieve his goal!

Format your bug report from Problem 1.2 in **Markdown syntax**! Make sure to use at least one *heading*, two *sub-headings*, one *list*, and one element of *text formatting* (e.g., bold, italic, underline). Of course, you are free to use as many as you want!

[1] `https://en.wikipedia.org/wiki/Grayscale`

Chapter 16

What did we learn in IWGS?

Outline of IWGS 1:

- ▷ Programming in Python: (main tool in IWGS)
  - ▷ Systematics and culture of programming
  - ▷ Program and control structures
  - ▷ Basic data structures like numbers and wordsstring, character encodings, unicode, and regular expressions
- ▷ Electronic documents and document processing:
  - ▷ text files
  - ▷ markup systems, HTML, and CSS
  - ▷ XML: Documents are trees.
- ▷ Web technologies for interactive documents and web applications
  - ▷ internet infrastructure: web browsers and server
  - ▷ serverside computation: bottle routing and
  - ▷ clientside interaction: dynamic HTML, JavaScript, HTML forms
- ▷ Web application project (fill in the blanks to obtain a working web app)

Outline of IWGS-II:

- ▷ Databases
  - ▷ CRUD operations, querying, and python embedding
  - ▷ XML and JSON for file based data storage
- ▷ BooksApp: a Books Application with persistent storage
- ▷ Image processing
  - ▷ Basics

  - ▹ Image transformations, Image Understanding
- ▷ Ontologies, semantic web, and WissKI
  - ▹ Ontologies (inference $\rightsquigarrow$ get out more than you put in)
  - ▹ semantic web Technologies (standardize ontology formats and inference)
  - ▹ Using semantic web Tech for cultural heritage research data $\rightsquigarrow$ the WissKI System
- ▷ Legal Foundations of Information Systems
  - ▹ Copyright & Licensing
  - ▹ Data Protection (GDPR)

Bibliography

[All18] Jay Allen. *New User Tutorial: Basic Shell Commands*. 2018. URL: `https://www.liquidweb.com/kb/new-user-tutorial-basic-shell-commands/` (visited on 10/22/2018).

[BHK16] Jens Bove, Lutz Heusinger, and Angela Kailus. *Marburger Informations-, Dokumentations- und Administrations-System (MIDAS): Handbuch und CD*. 4th ed. K.G.Saur, 2016. DOI: `10.11588/artdok.00003770`.

[BLFM05] Tim Berners-Lee, Roy T. Fielding, and Larry Masinter. *Uniform Resource Identifier (URI): Generic Syntax*. RFC 3986. Internet Engineering Task Force (IETF), 2005. URL: `http://www.ietf.org/rfc/rfc3986.txt`.

[CC] *CIDOC CRM - The CIDOC Conceptual Reference Model*. URL: `http://www.cidoc-crm.org/` (visited on 07/13/2020).

[CQ69] Allan M. Collins and M. Ross Quillian. “Retrieval time from semantic memory”. In: *Journal of verbal learning and verbal behavior* 8.2 (1969), pp. 240–247. DOI: `10.1016/S0022-5371(69)80069-1`.

[CS14] Scott Chacon and Ben Straub. *Pro Git*. 2nd Edition. APress, 2014. ISBN: 978-1484200773. URL: `https://git-scm.com/book/en/v2`.

[CSSa] *All CSS Specifications*. URL: `https://www.w3.org/Style/CSS/specs.en.html` (visited on 01/12/2020).

[CSSb] *CSS Specificity*. URL: `https://en.wikipedia.org/wiki/Cascading_Style_Sheets#Specificity` (visited on 12/03/2018).

[CSSc] *CSS Tutorial*. URL: `https://www.w3schools.com/css/default.asp` (visited on 12/02/2018).

[DCM12] DCMI Usage Board. *DCMI Metadata Terms*. DCMI Recommendation. Dublin Core Metadata Initiative, June 14, 2012. URL: `http://dublincore.org/documents/2012/06/14/dcmi-terms/`.

[DH98] S. Deering and R. Hinden. *Internet Protocol, Version 6 (IPv6) Specification*. RFC 2460. Internet Engineering Task Force (IETF), 1998. URL: `http://www.ietf.org/rfc/rfc2460.txt`.

[Dri10] Vincent Driessen. *A successful Git branching model*. online at `http://nvie.com/posts/a-successful-git-branching-model/`. 2010. URL: `http://nvie.com/posts/a-successful-git-branching-model/` (visited on 03/19/2015).

[Dru] *Drupal.org – Community plumbing*. URL: `http://drupal.org` (visited on 02/14/2015).

[Ecm] *ECMAScript Language Specification*. ECMA Standard. 5th Edition. Dec. 2009.

[ECRMa] *erlangen-crm*. URL: `https://github.com/erlangen-crm` (visited on 07/13/2020).

[ECRMb] *Erlangen CRM/OWL - An OWL DL 1.0 implementation of the CIDOC Conceptual Reference Model (CIDOC CRM)*. URL: `http://erlangen-crm.org/` (visited on 07/13/2020).

[ET] *`xml.etree.ElementTree` – The ElementTree XML API.* URL: `https://docs.python.org/3/library/xml.etree.elementtree.html` (visited on 04/15/2021).

[FAIR18] European Commission Expert Group on FAIR Data. *Turning FAIR into reality.* 2018. DOI: `10.2777/1524`.

[Fie+99] R. Fielding et al. *Hypertext Transfer Protocol – HTTP/1.1.* RFC 2616. Internet Engineering Task Force (IETF), 1999. URL: `http://www.ietf.org/rfc/rfc2616.txt`.

[FOAF14] *FOAF Vocabulary Specification 0.99.* Namespace Document. The FOAF Project, Jan. 14, 2014. URL: `http://xmlns.com/foaf/spec/`.

[Gfm] *GitHub Flavored Markdown Spec.* URL: `https://github.github.com/gfm/` (visited on 05/10/2020).

[Gla17] Matt Glaman. *Drupal 8 Development Cookbook – Harness the power of Drupal 8 with this recipe-based practical guide.* 2nd ed. Packt Publishing, 2-17. ISBN: 9781788290401.

[GN] *Geonames.* URL: `https://www.geonames.org/` (visited on 07/29/2020).

[GND] *DNB – The Integrated Authority File (GND).* URL: `https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html` (visited on 07/29/2020).

[Her+13a] Ivan Herman et al. *RDF 1.1 Primer (Second Edition). Rich Structured Data Markup for Web Documents.* W3C Working Group Note. World Wide Web Consortium (W3C), 2013. URL: `http://www.w3.org/TR/rdfa-primer`.

[Her+13b] Ivan Herman et al. *RDFa 1.1 Primer – Second Edition. Rich Structured Data Markup for Web Documents.* W3C Working Goup Note. World Wide Web Consortium (W3C), Apr. 19, 2013. URL: `http://www.w3.org/TR/xhtml-rdfa-primer/`.

[Hic+14] Ian Hickson et al. *HTML5. A Vocabulary and Associated APIs for HTML and XHTML.* W3C Recommentation. World Wide Web Consortium (W3C), Oct. 28, 2014. URL: `http://www.w3.org/TR/html5/`.

[HiDa] *HiDa.* URL: `https://www.startext.de/produkte/hida` (visited on 07/12/2020).

[Hit+12] Pascal Hitzler et al. *OWL 2 Web Ontology Language Primer (Second Edition).* W3C Recommendation. World Wide Web Consortium (W3C), 2012. URL: `http://www.w3.org/TR/owl-primer`.

[HL11] Martin Hilbert and Priscila López. "The World's Technological Capacity to Store, Communicate, and Compute Information". In: *Science* 331 (2011). DOI: `10.1126/science.1200970`. URL: `http://www.sciencemag.org/content/331/6018/692.full.pdf`.

[HWC] *The Hello World Collection.* URL: `http://helloworldcollection.de/` (visited on 11/23/2018).

[JKI] Jonas Betzendahl. *`juptyter.kwarc.info` Documentation.* URL: `https://kwarc.info/teaching/IWGS/jupyter-documentation.pdf` (visited on 08/29/2020).

[JS] *`json` – JSON encoder and decoder.* URL: `https://docs.python.org/3/library/json.html` (visited on 04/16/2021).

[Kar] Folgert Karsdorp. *Python Programming for the Humanities.* URL: `http://www.karsdorp.io/python-course/` (visited on 10/14/2018).

[KC04] Graham Klyne and Jeremy J. Carroll. *Resource Description Framework (RDF): Concepts and Abstract Syntax.* W3C Recommendation. World Wide Web Consortium (W3C), Feb. 10, 2004. URL: `http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`.

[Koh06] Michael Kohlhase. *OMDoc – An open markup format for mathematical documents [Version 1.2].* LNAI 4180. Springer Verlag, Aug. 2006. URL: `http://omdoc.org/pubs/omdoc1.2.pdf`.

[Koh08] Michael Kohlhase. "Using LaTeX as a Semantic Markup Format". In: *Mathematics in Computer Science* 2.2 (2008), pp. 279–304. URL: `https://kwarc.info/kohlhase/papers/mcs08-stex.pdf`.

[LM] *LabelMe: the open annotation tool.* URL: `http://labelme.csail.mit.edu` (visited on 08/28/2020).

[LOD] *The Linked Open Data Cloud.* URL: `https://lod-cloud.net/` (visited on 08/19/2020).

[LP] *Learn Python – Free Interactive Python Tutorial.* URL: `https://www.learnpython.org/` (visited on 10/24/2018).

[LXMLa] *lxml – XML and HTML with Python.* URL: `https://lxml.de` (visited on 12/09/2019).

[LXMLb] *lxml API.* URL: `https://lxml.de/api/` (visited on 12/09/2019).

[LXMLc] *The lxml.etree Tutorial.* URL: `https://lxml.de/tutorial.html` (visited on 12/09/2019).

[Nor+18a] Emily Nordmann et al. *Lecture capture: Practical recommendations for students and lecturers.* 2018. URL: `https://osf.io/huydx/download`.

[Nor+18b] Emily Nordmann et al. *Vorlesungsaufzeichnungen nutzen: Eine Anleitung für Studierende.* 2018. URL: `https://osf.io/e6r7a/download`.

[ODC] *Open Data Commons – Legal Tools For Open Data.* URL: `https://opendatacommons.org/` (visited on 07/29/2020).

[OWL09] OWL Working Group. *OWL 2 Web Ontology Language: Document Overview.* W3C Recommendation. World Wide Web Consortium (W3C), Oct. 27, 2009. URL: `http://www.w3.org/TR/2009/REC-owl2-overview-20091027/`.

[P3D] *Python 3 Documentation.* URL: `https://docs.python.org/3/` (visited on 09/02/2014).

[PMDA] *Python – MySQL Database Access.* URL: `https://www.tutorialspoint.com/python/python_database_access.htm` (visited on 11/18/2018).

[Pro] *Protégé.* Project Home page at `http://protege.stanford.edu`. URL: `http://protege.stanford.edu`.

[PRR97] G. Probst, St. Raub, and Kai Romhardt. *Wissen managen.* 4 (2003). Gabler Verlag, 1997.

[PS08] Eric Prud'hommeaux and Andy Seaborne. *SPARQL Query Language for RDF.* W3C Recommendation. World Wide Web Consortium (W3C), Jan. 15, 2008. URL: `http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/`.

[PyRegex] Rodolfo Carvalho. *PyRegex - Your Python Regular Expression's Best Buddy.* URL: `http://www.pyregex.com/` (visited on 12/03/2018).

[Pyt] *`re` – Regular expression operations.* online manual at `https://docs.python.org/2/library/re.html`. URL: `https://docs.python.org/2/library/re.html`.

[Rfc] *DOD Standard Internet Protocol.* RFC. 1980. URL: `http://tools.ietf.org/rfc/rfc760.txt`.

[RHJ98] Dave Raggett, Arnaud Le Hors, and Ian Jacobs. *HTML 4.0 Specification.* W3C Recommendation REC-html40. World Wide Web Consortium (W3C), Apr. 1998. URL: `http://www.w3.org/TR/PR-xml.html`.

[She24] Esther Shein. 2024. URL: `https://cacm.acm.org/news/the-impact-of-ai-on-computer-science-education/`.

[Smi76] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations.* W. Strahan and T. Cadell, 1776.

[SR14] Guus Schreiber and Yves Raimond. *RDF 1.1 Primer.* W3C Working Group Note. World Wide Web Consortium (W3C), 2014. URL: `http://www.w3.org/TR/rdf-primer`.

[SSU04] Susan Schreibman, Ray Siemens, and John Unsworth, eds. *A Companion to Digital Humanities*. Wiley-Blackwell, 2004. ISBN: 978-1-405-10321-3. URL: `http://www.digitalhumanities.org/companion`.

[sTeX] *sTeX: A semantic Extension of TeX/LaTeX*. URL: `https://github.com/sLaTeX/sTeX` (visited on 05/11/2020).

[Sth] *A Beginner's Python Tutorial*. `http://www.sthurlow.com/python/`. seen 2014-09-02. URL: `http://www.sthurlow.com/python/`.

[STPL] *Simple Template Engine*. URL: `https://bottlepy.org/docs/dev/stpl.html` (visited on 12/08/2018).

[SUMO] *Suggested Upper Merged Ontology*. URL: `http://www.adampease.org/OP/` (visited on 01/25/2019).

[Swe13] Al Sweigart. *Invent with Python: Learn to program by making computer games*. 2nd ed. online at `http://inventwithpython.com`. 2013. ISBN: 978-0-9821060-1-3. URL: `http://inventwithpython.com`.

[Tom17] Todd Tomlinson. *Enterprise Drupal 8 Development – For Advanced Projects and Large Development Teams*. Apress, 2017. ISBN: 9781484202548.

[Tur95] Sherry Turkle. *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster, 1995.

[UL] *`urllib` – URL handling modules*. URL: `https://docs.python.org/3/library/urllib.html` (visited on 04/15/2021).

[WH] *WissKI Handbuch*. URL: `http://wiss-ki.eu/documentation/wisski_handbuch` (visited on 07/23/2020).

[Wil+16] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3 (2016). DOI: `10.1038/sdata.2016.18`.

[Xam] *apache friends - Xampp*. `http://www.apachefriends.org/en/xampp.html`. URL: `http://www.apachefriends.org/en/xampp.html`.

Part III

Excursions

As this course is predominantly an overview over (some) computer science tools useful in the humanities and social sciences and not about the theoretical underpinnings, we give the discussion about these as a “suggested readings” part here.

Appendix A

Internet Basics

We will show aspects of how the internet can cope with this enormous growth of numbers of computers, connections and services.

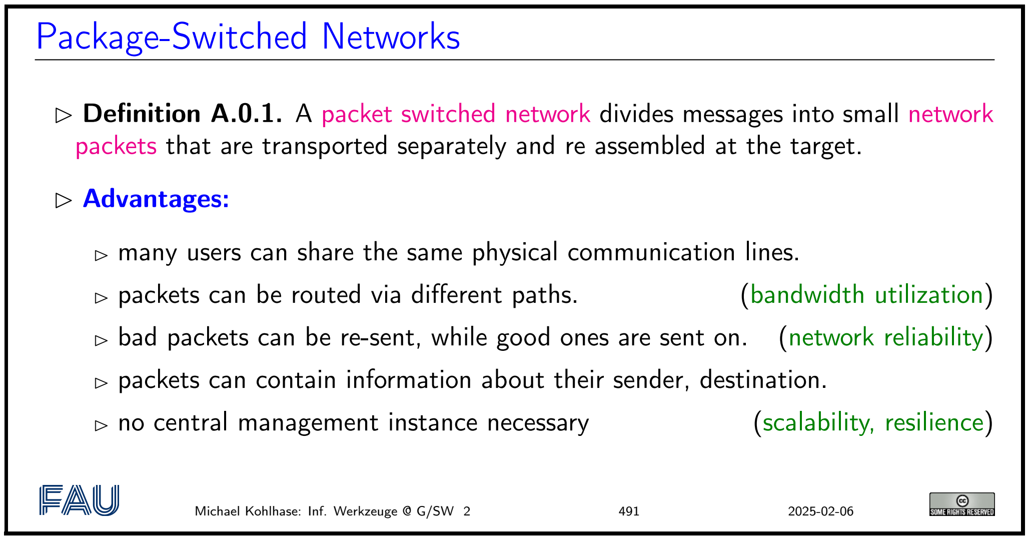
The growth of the internet rests on three design decisions taken very early on. The internet

1. is a packet-switched network rather than a network, where computers communicate via dedicated physical communication lines.

2. is a network, where control and administration are decentralized as much as possible.

3. is an infrastructure that only concentrates on transporting packets/datagrams between computers. It does not provide special treatment to any packets, or try to control the content of the packets.

The first design decision is a purely technical one that allows the existing communication lines to be shared by multiple users, and thus save on hardware resources. The second decision allows the administrative aspects of the internet to scale up. Both of these are crucial for the scalability of the internet. The third decision (often called "net neutrality") is hotly debated. The defenders cite that net neutrality keeps the Internet an open market that fosters innovation, where as the attackers say that some uses of the network (illegal file sharing) disproportionately consume resources.



Package-Switched Networks

- ▷ **Definition A.0.1.** A packet switched network divides messages into small network packets that are transported separately and re assembled at the target.
- ▷ **Advantages:**
  - ▷ many users can share the same physical communication lines.
  - ▷ packets can be routed via different paths. (bandwidth utilization)
  - ▷ bad packets can be re-sent, while good ones are sent on. (network reliability)
  - ▷ packets can contain information about their sender, destination.
  - ▷ no central management instance necessary (scalability, resilience)

Michael Kohlhase: Inf. Werkzeuge @ G/SW 2 491 2025-02-06

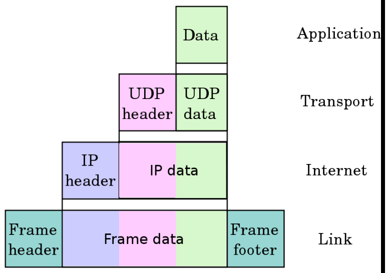
These ideas are implemented in the Internet Protocol Suite, which we will present in the rest of the chapter. A main idea of this set of protocols is its layered design that allows to separate concerns and implement functionality separately.

The Intenet Protocol Suite

▷ **Definition A.0.2.** The Internet Protocol Suite (commonly known as TCP/IP) is the set of communications protocols used for the internet and other similar networks. It structured into 4 layers.

Layer	e.g.
Application Layer	HTTP, SSH
Transport Layer	UDP, TCP
Internet Layer	IPv4, IPsec
Link Layer	Ethernet, DSL

▷ **Layers in TCP/IP:** TCP/IP uses encapsulation to provide abstraction of protocols and services.
An application (the highest level of the model) uses a set of protocols to send its data down the layers, being further encapsulated at each level.



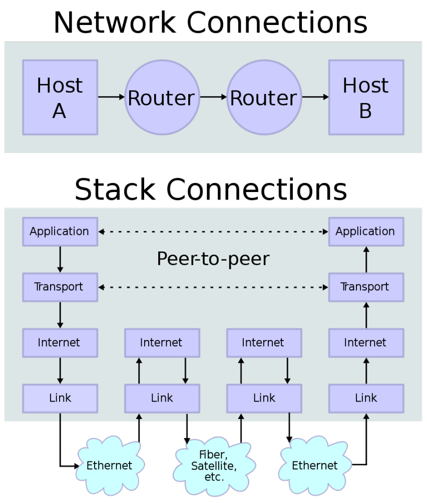
The Internet as a Network of Networks

▷ **Example A.0.3 (TCP/IP Scenario).** Consider a situation with two internet host computers communicate across local network boundaries.

▷ Network boundaries are constituted by internetworking gateways (routers).

▷ **Definition A.0.4.** A router is a purposely customized computer used to forward data among computer networks beyond directly connected devices.

▷ A router implements the link and internet layers only and has two network connections.



We will now take a closer look at each of the layers shown above, starting with the lowest one.
Instead of going into network topologies, protocols, and their implementation into physical signals that make up the link layer, we only discuss the devices that deal with them. Network Interface controllers are specialized hardware that encapsulate all aspects of link-level communication, and we take them as black boxes for the purposes of this course.

Network Interfaces

▷ The nodes in the internet are computers, the edges communication channels

▷ **Definition A.0.5.** A network interface controller (NIC) is a hardware device that handles an interface to a computer network and thus allows a network-capable device to access that network.

▷ **Definition A.0.6.** Each NIC contains a unique number, the media access control address (MAC address), identifies the device uniquely on the network.

▷ MAC addresses are usually 48-bit numbers issued by the manufacturer, they are usually displayed to humans as six groups of two hexadecimal digits, separated by hyphens (-) or colons (:), in transmission order, e.g. 01-23-45-67-89-AB, 01:23:45:67:89:AB.

▷ **Definition A.0.7.** A network interface is a software component in the operating system that implements the higher levels of the network protocol (the NIC handles the lower ones).

Layer	e.g.
Application Layer	HTTP, SSH
Transport Layer	TCP
Internet Layer	IPv4, IPsec
Link Layer	Ethernet, DSL

▷ A computer can have more than one network interface. (e.g. a router)

The next layer ist he Internet Layer, it performs two parts: addressing and packing packets.

Internet Protocol and IP Addresses

▷ **Definition A.0.8.** The Internet Protocol (IP) is a protocol used for communicating data across a packet-switched internetwork. The Internet Protocol defines addressing methods and structures for datagram encapsulation. The Internet Protocol also routes data packets between networks

▷ **Definition A.0.9.** An IP address is a numerical label that is assigned to devices participating in a computer network, that uses the Internet Protocol for communication between its nodes.

▷ An IP address serves two principal functions: host or network interface identification and location addressing.

▷ **Definition A.0.10.** The global IP address space allocations are managed by the Internet Assigned Numbers Authority (IANA), delegating allocate IP address blocks to five Regional Internet Registries (RIRs) and further to Internet service providers (ISPs).

Internet Protocol and IP Addresses

▷ **Definition A.0.11.** The internet mainly uses Internet Protocol Version 4 (IPv4) [Rfc], which uses 32 bit numbers (IPv4 addresses) for identification of network interfaces of computers.

▷ IPv4 was standardized in 1980, it provides 4,294,967,296 (2^{32}) possible unique addresses. With the enormous growth of the internet, we are fast running out of IPv4 addresses.

▷ **Definition A.0.12.** Internet Protocol Version 6 [DH98] (IPv6), which uses 128 bit numbers (IPv6 addresses) for identification.

▷ Although IP addresses are stored as binary numbers, they are usually displayed in human-readable notations, such as $208.77.188.166$ (for IPv4), and $2001{:}db8{:}0{:}1234{:}0{:}567{:}1{:}1$ (for IPv6).

The internet infrastructure is currently undergoing a dramatic retooling, because we are moving from IPv4 to IPv6 to counter the depletion of IP addresses. Note that this means that all routers and switches in the internet have to be upgraded. At first glance, it would seem that this problem could have been avoided if we had only anticipated the need for more the 4 million computers. But remember that TCP/IP was developed at a time, where the internet did not exist yet, and it's precursor had about 100 computers. Also note that the IP addresses are part of every packet, and thus reserving more space for them would have wasted bandwidth in a time when it was scarce.
We will now go into the detailed structure of the IP packets as an example of how a low-level protocol is structured. Basically, an IP packet has two parts: the “header”, whose sequence of bytes is strictly standardized, and the “payload”, a segment of bytes about which we only know the length, which is specified in the header.

The Structure of IP Packets

▷ **Definition A.0.13.** IP packets are composed of a $160\mathrm{b}$ header and a payload. The IPv4 packet header consists of:

b	name	comment
4	version	IPv4 or IPv6 packet
4	Header Length	in multiples 4 bytes (e.g., 5 means 20 bytes)
8	QoS	Quality of Service, i.e. priority
16	length	of the packet in bytes
16	fragid	to help reconstruct the packet from fragments,
3	fragmented	DF $\widehat{=}$ “Don't fragment”/MF $\widehat{=}$ “More Fragments”
13	fragment offset	to identify fragment position within packet
8	TTL	Time to live (router hops until discarded)
8	protocol	TCP, UDP, ICMP, etc.
16	Header Checksum	used in error detection,
32	Source IP	
32	target IP	
...	optional flags	according to header length

▷ Note that delivery of IP packets is not guaranteed by the IP protocol.

As the internet protocol only supports addressing, routing, and packaging of packets, we need another layer to get services like the transporting of files between specific computers. Note that the IP protocol does not guarantee that packets arrive in the right order or indeed arrive at all, so the transport layer protocols have to take the necessary measures, like packet re-sending or handshakes, ...

The Transport Layer

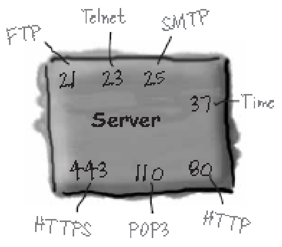
- ▷ **Definition A.0.14.** The transport layer is responsible for delivering data to the appropriate application process on the host computers by forming data packets, and adding source and destination port numbers in the header.
- ▷ **Definition A.0.15.** The internet protocol mainly uses suite the Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) protocols at the transport layer.
- ▷ TCP is used for communication, UDP for multicasting and broadcasting.
- ▷ TCP supports virtual circuits, i.e. provide connection oriented communication over an underlying packet oriented datagram network. (hide/reorder packets)
- ▷ TCP provides end-to-end reliable communication (error detection & automatic repeat)

We will see that there are quite a lot of services at the network application level. And indeed, many web-connected computers run a significant subset of them at any given time, which could lead to problems of determining which packets should be handled by which service. The answer to this problem is a system of “ports” (think pigeon holes) that support finer-grained addressing to the various services.

Ports

- ▷ **Definition A.0.16.** To separate the services and protocols of the network application layer, network interfaces assign them specific port, referenced by a number.
- ▷ **Example A.0.17.** We have the following ports in common use on the internet

Port	use	comment
22	SSH	remote shell
53	DNS	Domain Name System
80	HTTP	World Wide Web
443	HTTPS	HTTP over SSL



On top of the transport-layer services, we can define even more specific services. From the perspective of the internet protocol suite this layer is unregulated, and application-specific. From a user perspective, many useful services are just “applications” and live at the application layer.

The Application Layer

▷ **Definition A.0.18.** The application layer of the internet protocol suite contains all protocols and methods that fall into the realm of process-to-process communications via an Internet Protocol (IP) network using the Transport Layer protocols to establish underlying host-to-host connections.

▷ **Example A.0.19 (Some Application Layer Protocols and Services).**

BitTorrent	Peer-to-peer	Atom	Syndication
DHCP	Dynamic Host Configuration	DNS	Domain Name System
FTP	File Transfer Protocol	HTTP	HyperText Transfer
IMAP	Internet Message Access	IRCP	Internet Relay Chat
NFS	Network File System	NNTP	Network News Transfer
NTP	Network Time Protocol	POP	Post Office Protocol
RPC	Remote Procedure Call	SMB	Server Message Block
SMTP	Simple Mail Transfer	SSH	Secure Shell
TELNET	Terminal Emulation	WebDAV	Write-enabled Web

The domain name system is a sort of telephone book of the internet that allows us to use symbolic names for hosts like `kwarc.info` instead of the IP number 212.201.49.189.

Domain Names

▷ **Definition A.0.20.** The DNS (Domain Name System) is a distributed set of servers that provides the mapping between (static) IP addresses and domain names.

▷ **Example A.0.21.** e.g. `www.kwarc.info` stands for the IP address 212.201.49.189.

▷ **Definition A.0.22.** Domain names are hierarchically organized, with the most significant part (the top level domain TLD) last.

▷ networked computers can have more than one DNS name. (virtual servers)

▷ Domain names must be registered to ensure uniqueness (registration fees vary, cybersquatting)

▷ **Definition A.0.23.** ICANN is a non profit organization was established to regulate human friendly domain names. It approves top-level domains, and corresponding domain name registrars and delegates the actual registration to them.

Let us have a look at a selection of the top-level domains in use today.

Domain Name Top-Level Domains

▷ .com ("commercial") is a generic top-level domain. It was one of the original top-level domains, and has grown to be the largest in use.

▷ .org ("organization") is a generic top-level domain, and is mostly associated with non-profit organizations. It is also used in the charitable field, and used by the open-source movement. Government sites and Political parties in the US have domain names ending in .org

▷ .net ("network") is a generic top-level domain and is one of the original top-level domains. Initially intended to be used only for network providers (such as Internet service providers). It is still popular with network operators, it is often treated as a second .com. It is currently the third most popular top-level domain.

▷ .edu ("education") is the generic top-level domain for educational institutions, primarily those in the United States. One of the first top-level domains, .edu was originally intended for educational institutions anywhere in the world. Only post-secondary institutions that are accredited by an agency on the U.S. Department of Education's list of nationally recognized accrediting agencies are eligible to apply for a .edu domain.

SOME RIGHTS RESERVED

Domain Name Top-Level Domains

▷ .info ("information") is a generic top-level domain intended for informative website's, although its use is not restricted. It is an unrestricted domain, meaning that anyone can obtain a second-level domain under .info. The .info was one of many extension(s) that was meant to take the pressure off the overcrowded .com domain.

▷ .gov ("government") a generic top-level domain used by government entities in the United States. Other countries typically use a second-level domain for this purpose, e.g., .gov.uk for the United Kingdom. Since the United States controls the .gov Top Level Domain, it would be impossible for another country to create a domain ending in .gov.

▷ .biz ("business") the name is a phonetic spelling of the first syllable of "business". A generic top-level domain to be used by businesses. It was created due to the demand for good domain names available in the .com top-level domain, and to provide an alternative to businesses whose preferred .com domain name which had already been registered by another.

▷ .xxx ("porn") the name is a play on the verdict "X-rated" for movies. A generic top-level domain to be used for sexually explicit material. It was created in 2011 in the hope to move sexually explicit material from the "normal web". But there is no mandate for porn to be restricted to the .xxx domain, this would be difficult due to problems of definition, different jurisdictions, and free speech issues.

SOME RIGHTS RESERVED

Note: Anybody can register a domain name from a registrar against a small yearly fee. Domain names are given out on a first-come-first-serve basis by the domain name registrars, which usually also offer services like domain name parking, DNS management, URL forwarding, etc.

The telnet Protocol

▷ **Problem:** We need a way to remotely operate networked computers via a shell.

▷ **Idea:**

Send shell instructions and responses as text messages between a terminal client (a program on the local host) and a terminal server (a program on the remote host).

- ▷ **Definition A.0.24.** The telnet protocol uses TCP directly to send text based messages two networked computers. It customarily uses port 25.
- ▷ **Remark:**

  telnet is one of the oldest protocols in the TCP/IP protocol suite. It is no longer used much by itself (it is superseded by rsh and ssh), but still serves as a basis for other protocols, e.g. HTTP.

The next application-level service is the SMTP protocol used for sending e-mail. It is based on the telnet protocol for remote terminal emulation which we do not discuss here.

A Protocol Example: SMTP over telnet

- ▷ **Definition A.0.25.** The Simple Mail Transfer Protocol (SMTP) is a communication protocol for electronic mail transmission based on telnet.
- ▷ **Example A.0.26.** The SMTP protocol starts out by establishing identity
  - ▷ We call up the telnet service on the Jacobs mail server

    ```
    telnet exchange.jacobs-university.de 25
    ```
  - ▷ it identifies itself (have some patience, it is very busy)

    ```

    Trying 10.70.0.128...
    Connected to exchange.jacobs-university.de.
    Escape character is '^]'.
    220 SHUBCAS01.jacobs.jacobs-university.de
    Microsoft ESMTP MAIL Service ready at Tue, 3 May 2011 13:51:23 +0200
    ```
  - ▷ We introduce ourselves politely (but we lie about our identity)

    ```
    helo mailhost.domain.tld
    ```
  - ▷ It is really very polite.

    ```
    250 SHUBCAS04.jacobs.jacobs-university.de Hello [10.222.1.5]
    ```

SMTP over telnet: The e mail itself

- ▷ **Example A.0.27 (Continued).** After identity is established, the e-mail is specified.
  - ▷ We start addressing an e-mail (again, we lie about our identity)

    ```
    mail from: user@domain.tld
    ```
  - ▷ this is acknowledged

    ```
    250 2.1.0 Sender OK
    ```
  - ▷ We set the recipient (the real one, so that we really get the e-mail)

    ```
    rcpt to: m.kohlhase@jacobs-university.de
    ```
  - ▷ this is acknowledged

    ```
    250 2.1.0 Recipient OK
    ```
  - ▷ we tell the mail server that the mail data comes next

    ```
    data
    ```
  - ▷ this is acknowledged

    ```
    354 Start mail input; end with <CRLF>.<CRLF>
    ```

- ▷ Now we can just type the a-mail, optionally with Subject, date,...

```
Subject: Test via SMTP

and now the mail body itself
.
```

- ▷ And a dot on a line by itself sends the e mail off

```
250 2.6.0 <ed73c3f3-f876-4d03-98f2-e5ad5bbb6255@SHUBCAS04.jacobs.jacobs-university.de>
[InternalId=965770] Queued mail for delivery
```

SMTP over telnet: Disconnecting

▷ **Example A.0.28 (Continued).**

- ▷ That was almost all, but we close the connection (this is a telnet command)

```
quit
```

- ▷ our terminal server (the telnet program) tells us

```
221 2.0.0 Service closing transmission channel
Connection closed by foreign host.
```

Essentially, the SMTP protocol mimics a conversation of polite computers that exchange messages by reading them out loud to each other (including the addressing information). We could go on for quite a while with understanding one Internet protocol after each other, but this is beyond the scope of this course (indeed there are specific courses that do just that). Here we only answer the question where these protocols come from, and where we can find out more about them.

Internet Standardization

- ▷ **Question:** Where do all the protocols come from?(someone has to manage that)
- ▷ **Definition A.0.29.** The Internet Engineering Task Force (IETF) is an open standards organization that develops and standardizes internet standards, in particular the TCP/IP and Internet protocol suite.
- ▷ All participants in the IETF are volunteers (usually paid by their employers)
- ▷ **Rough Consensus and Running Code:** Standards are determined by the "rough consensus method" (consensus preferred, but not all members need agree) IETF is interested in practical, working systems that can be quickly implemented.
- ▷ **Idea:** running code leads to rough consensus or vice versa.
- ▷ **Definition A.0.30.** The standards documents of the IETF are called Request for Comments (RFC). (more than 6300 so far; see `http://www.rfceditor.org/`)