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

2024-02-08

Chapter 1
Preliminaries

1.1 Administrativa

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)

Assessment, Grades

- **Grading Background/Theory:** Only modules are graded! (by the law)
  - Module "DH-Einführung" (DHE) $\widehat{=}$ courses IWGS1/2, DH-Einführung.
  - DHE module grade $\leadsto$ pass/fail determined by "portfolio" $\widehat{=}$ collection of contributions/assessments.
- **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. 2024.
- **Retake Exam:** 60 min exam at the end of the exam break. ($\sim$ May 4. 2024)

IWGS Homework Assignments

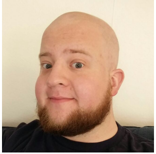
- **Homeworks:** will be small individual problem/programming/system assignments
  - but take time to solve (at least read them directly $\leadsto$ questions)
  - group submission if and only if explicitly permitted.
- ⚠ Without trying the homework assignments you are unlikely to pass the exam.
- **Admin:** To keep things running smoothly
  - Homeworks will be posted on StudOn.
  - Sign up for IWGS under `https://www.studon.fau.de/crs5323051.html`.
  - Homeworks are handed in electronically there. (plain text, program files, PDF)
  - Go to the tutorials, discuss with your TA! (they are there for you!)
- **Homework Discipline:**
  - Start early! (many assignments need more than one evening's work)
  - Don't start by sitting at a blank screen (talking & study group help)
  - Humans will be trying to understand the text/code/math when grading it.

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)



- **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 slides 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

Textbook, Handouts and Information, Forums, Videos

- **No Textbook:** but lots of online python tutorials on the web.
- Course notes will be posted at `http://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 WS 2020/21 are at `https://www.fau.tv/course/id/1923` (not much changed)
- Matrix chat at `#iwgs:fau.de` (via IDM) (instructions)
- **StudOn Forum:** `https://www.studon.fau.de/crs5323051.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 forum turns into a valuable resource!

Experiment: Learning Support with KWARC Technologies

- **My research area:** Deep representation formats for (mathematical) knowledge
- **One Application:** Learning support systems (represent knowledge to transport it)
- **Experiment:** Start with this course (Drink my own medicine)
  1. Re-represent the slide materials in OMDoc (Open Mathematical Documents)
  2. Feed it into the ALeA system (`http://courses.voll-ki.fau.de`)
  3. Try it on you all (to get feedback from you)
- Research tasks
  - help me complete the material on the slides (what is missing/would help?)
  - I need to remember "what I say", examples on the board. (take notes)
- Benefits for you (so why should you help?)
  - you will be mentioned in the acknowledgements (for all that is worth)
  - you will help build better course materials (think of next-year's students)

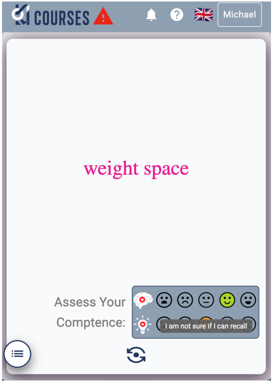
VoLL-KI Portal at https://courses.voll-ki.fau.de

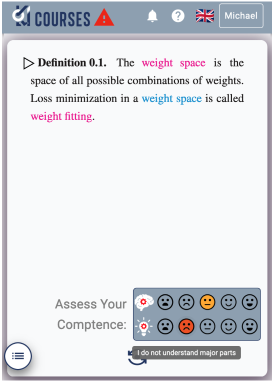
- **Portal for ALeA Courses:** https://courses.voll-ki.fau.de

- **AI-1 in ALeA:** https://courses.voll-ki.fau.de/course-home/ai-1
  - All details for the course.
  - recorded syllabus (keep track of material covered in course)
  - syllabus of the last semester (for over/preview)
- **ALeA Status:** The ALeA system is deployed at FAU for over 1000 students taking six courses
  - (some) students use the system actively (our logs tell us)
  - reviews are mostly positive/enthusiastic (error reports pour in)

New Feature: Drilling with Flashcards

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



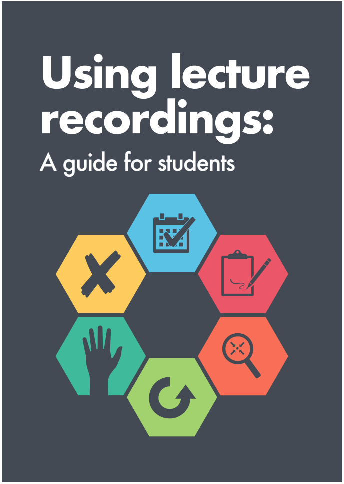


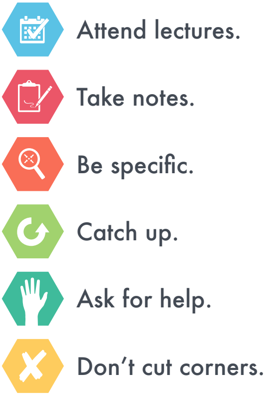
...and the definition/answer is on the back.

- 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)

Practical recommendations on Lecture Videos

- **Excellent Guide:** [Nor+18a] (german Version at [Nor+18b])





- Normally intended for "offline students" $\widehat{=}$ everyone during Corona times.

Software/Hardware tools

- You will need computer access for this course
- we recommend the use of standard software tools
  - find a text editor you are comfortable with (get good with it) A text editor is a program you can use to write text files. (not `MSWord`)
  - any operating system you like (I can only help with `UNIX`)
  - Any browser you like (I use `FireFox`: less spying)
- **Advice:** learn how to touch-type NOW (reap the benefits earlier, not later)
  - you will be typing multiple hours/week in the next decades
  - touch-typing is about twice as fast as "system eagle".
  - you can learn it in two weeks (good programs)

1.2 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 and tools that can become *useful in the short term*
  - generate immediate success and gratification,
  - 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.

Academic Culture in Computer Science

- **Definition 2.1.** The academic culture is the overall style of working, research, and discussion in an academic field.
- **Observation 2.2.** *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)

Outline of IWGS 1:

- Programming in Python: (main tool in IWGS)
  - Systematics and culture of programming
  - Program and control structures
  - Basic data strutures like numbers and strings, character encodings, unicode, and regular expressions
- Digital 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 servers
  - serverside computing: bottle routing and
  - client-side interaction: dynamic HTML, JavaScript, HTML forms
- Web application project (fill in the blanks to obtain a working web app)

Do I need to attend the lectures

- Attendance is not mandatory for the IWGS lecture

Do I need to attend the lectures

- Attendance is not mandatory for the IWGS lecture
- There are two ways of learning IWGS: (both are OK, your mileage may vary)
  - Approach B: Read a Book
  - Approach I: come to the lectures, be involved, interrupt me whenever you have a question.

The only advantage of I over B is that books do not answer questions (yet! ⇝ we are working on this in AI research)

Do I need to attend the lectures

- Attendance is not mandatory for the IWGS lecture
- There are two ways of learning IWGS: (both are OK, your mileage may vary)
  - Approach B: Read a Book
  - Approach I: come to the lectures, be involved, interrupt me whenever you have a question.

  The only advantage of I over B is that books do not answer questions (yet! ⇝ we are working on this in AI research)
- Approach S: come to the lectures and sleep does not work!

Do I need to attend the lectures

- Attendance is not mandatory for the IWGS lecture
- There are two ways of learning IWGS: (both are OK, your mileage may vary)
  - Approach B: Read a Book
  - Approach I: come to the lectures, be involved, interrupt me whenever you have a question.

  The only advantage of I over B is that books do not answer questions (yet! ⇝ we are working on this in AI research)
- Approach S: come to the lectures and sleep does not work!
- **I really mean it:** If you come to class, be involved, ask questions, challenge me with comments, tell me about errors, ...

Do I need to attend the lectures

- Attendance is not mandatory for the IWGS lecture
- There are two ways of learning IWGS: (both are OK, your mileage may vary)
  - Approach B: Read a Book
  - Approach I: come to the lectures, be involved, interrupt me whenever you have a question.

  The only advantage of I over B is that books do not answer questions (yet! ⇝ we are working on this in AI research)
- Approach S: come to the lectures and sleep does not work!
- **I really mean it:** If you come to class, be involved, ask questions, challenge me with comments, tell me about errors, ...
  - I would much rather have a lively discussion than get through all the slides

Do I need to attend the lectures

- Attendance is not mandatory for the IWGS lecture
- There are two ways of learning IWGS: (both are OK, your mileage may vary)
  - Approach B: Read a Book
  - Approach I: come to the lectures, be involved, interrupt me whenever you have a question.

  The only advantage of I over B is that books do not answer questions (yet! ⇝ we are working on this in AI research)
- Approach S: come to the lectures and sleep does not work!
- **I really mean it:** If you come to class, be involved, ask questions, challenge me with comments, tell me about errors, ...
  - I would much rather have a lively discussion than get through all the slides
  - You learn more, I have more fun (Approach B serves as a backup)

Do I need to attend the lectures

- Attendance is not mandatory for the IWGS lecture
- There are two ways of learning IWGS: (both are OK, your mileage may vary)
  - Approach B: Read a Book
  - Approach I: come to the lectures, be involved, interrupt me whenever you have a question.

  The only advantage of I over B is that books do not answer questions (yet! ⇝ we are working on this in AI research)
- Approach S: come to the lectures and sleep does not work!
- **I really mean it:** If you come to class, be involved, ask questions, challenge me with comments, tell me about errors, ...
  - I would much rather have a lively discussion than get through all the slides
  - You learn more, I have more fun (Approach B serves as a backup)
  - You may have to change your habits, overcome shyness, ... (please do!)

Do I need to attend the lectures

- Attendance is not mandatory for the IWGS lecture
- There are two ways of learning IWGS: (both are OK, your mileage may vary)
  - Approach B: Read a Book
  - Approach I: come to the lectures, be involved, interrupt me whenever you have a question.

  The only advantage of I over B is that books do not answer questions (yet! ⇝ we are working on this in AI research)
- Approach S: come to the lectures and sleep does not work!
- **I really mean it:** If you come to class, be involved, ask questions, challenge me with comments, tell me about errors, ...
  - I would much rather have a lively discussion than get through all the slides
  - You learn more, I have more fun (Approach B serves as a backup)
  - You may have to change your habits, overcome shyness, ... (please do!)
- This is what I get paid for, and I am more expensive than most books (get your money's worth)

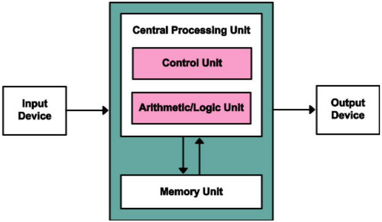
Part 1
IWGS-1: Programming, Documents, Web Applications

Chapter 2
Introduction to Programming

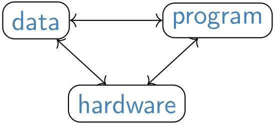
2.1 What is Programming?

Computer Hardware/Software & Programming

- **Definition 1.1.** Computers consist of hardware and software.
- **Definition 1.2.** Hardware consists of
  - 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 1.3.** Software consists of
  - data that represents objects and their relationships in the world
  - programs that inputs, manipulates, outputs data



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

Programming Languages

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

Programming Languages

- Programming $\widehat{=}$ writing programs (Telling the computer what to do)
- *Remark 1.6.* The computer does exactly as told
  - extremely fast extremely reliable
  - completely stupid: will not do what you mean unless you tell it exactly
- Programming can be extremely fun/frustrating/addictive (try it)

Programming Languages

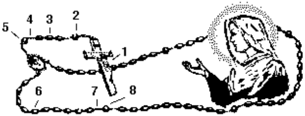
- Programming $\widehat{=}$ writing programs (Telling the computer what to do)
- *Remark 1.8.* The computer does exactly as told
  - extremely fast extremely reliable
  - completely stupid: will not do what you mean unless you tell it exactly
- Programming can be extremely fun/frustrating/addictive (try it)
- **Definition 1.9.** 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)

Programming Languages

- Programming $\widehat{=}$ writing programs (Telling the computer what to do)
- *Remark 1.10.* The computer does exactly as told
  - extremely fast extremely reliable
  - completely stupid: will not do what you mean unless you tell it exactly
- Programming can be extremely fun/frustrating/addictive (try it)
- **Definition 1.11.** 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.

Program Execution

- **Definition 1.12.** Algorithm: informal description of what to do (good enough for humans)
- **Example 1.13.** 



- **Example 1.14.** Program: computer processable version, e.g. in Python.

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

- **Definition 1.15.** Interpreter: reads a program and executes it directly
  - special case: interactive interpretation (lets you experiment easily)
- **Definition 1.16.** Compiler: translates a program (the source) into another program (the binary) in a much simpler programming language for optimized execution on hardware directly.
- *Remark 1.17.* Compilers are efficient, but more cumbersome for development.

2.2 Programming in IWGS

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)

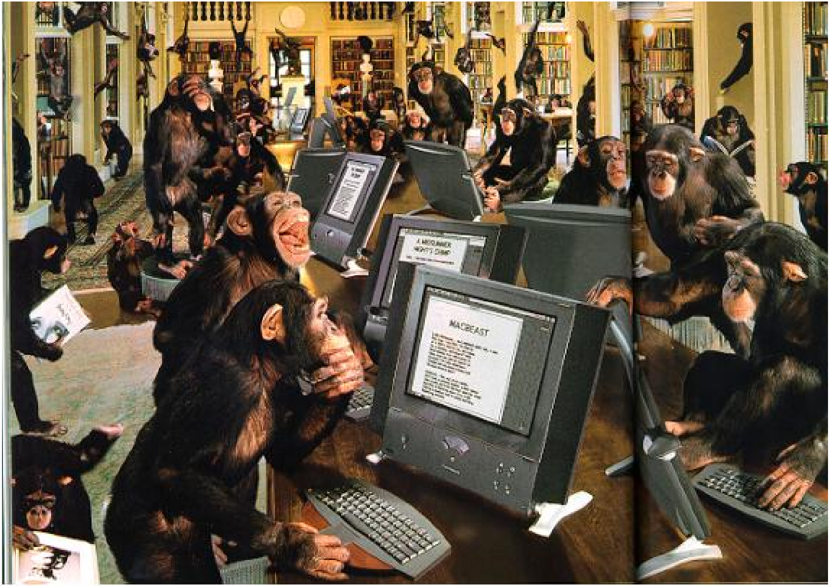
$$\text{RTFM } (\widehat{=} \text{ “read those fine manuals”})$$

- **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

But Seriously... Learning programming in IWGS

- The IWGS lecture 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

2am in the Kollegienhaus CIP Pool



no, let's think

- We have to fully understand the problem, our tools, and the solution space first (That is what the IWGS lecture is for)
  - read Richard Stallman's quote carefully $\leadsto$ 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

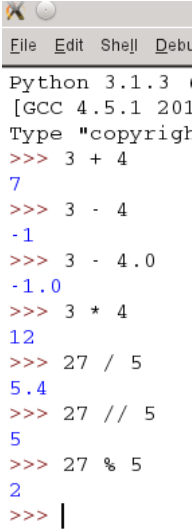
2.3.1 Hello IWGS

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)

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 3.1.** Expressions are created from values (and other expressions) via operators.
  - **Observation:** The Python interpreter simplifies expressions to values by computation.



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

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.
- 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 ''')

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.
- 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

JupyterLab A Cloud IDE for Python

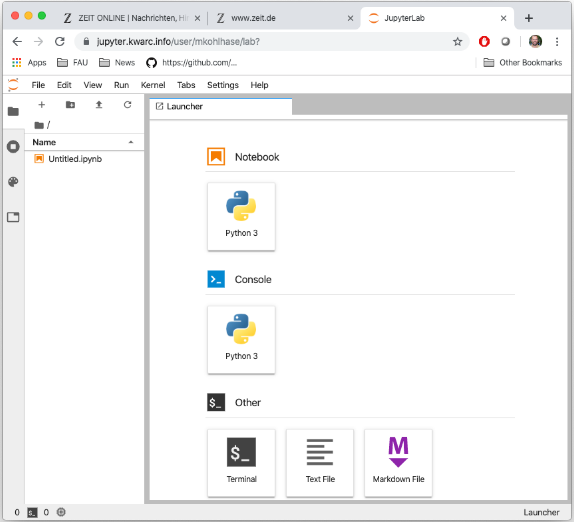
- **For helping you** it would be good if the TAs could access to your code
- **Idea:** Use a web IDE (a web based integrated development environment): JupyterLab, which you can use for interacting with the interpreter.

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`

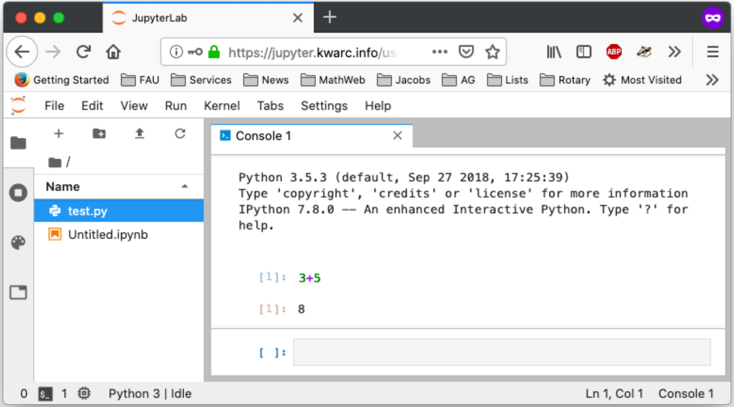
JupyterLab Components

- **Definition 3.2.** The JupyterLab dashboard gives you access to all components.



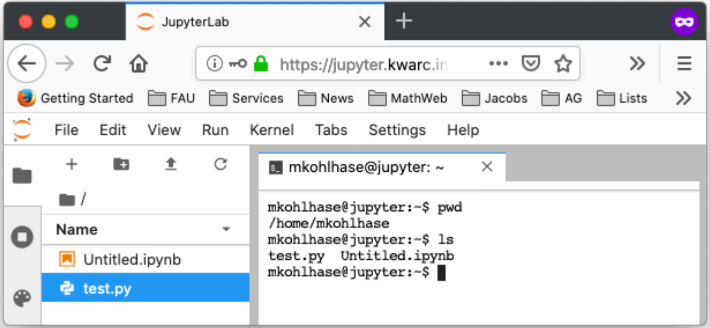
JupyterLab Components

- **Definition 3.6.** The JupyterLab dashboard gives you access to all components.
- **Definition 3.7.** The JupyterLab python console, i.e. a Python interpreter in your browser. (use this for Python interaction and testing.)



JupyterLab Components

- **Definition 3.10.** The JupyterLab dashboard gives you access to all components.
- **Definition 3.11.** The JupyterLab python console, i.e. a Python interpreter in your browser. (use this for Python interaction and testing.)
- **Definition 3.12.** The JupyterLab terminal, i.e. a `UNIX` shell in your browser.(use this for managing files)

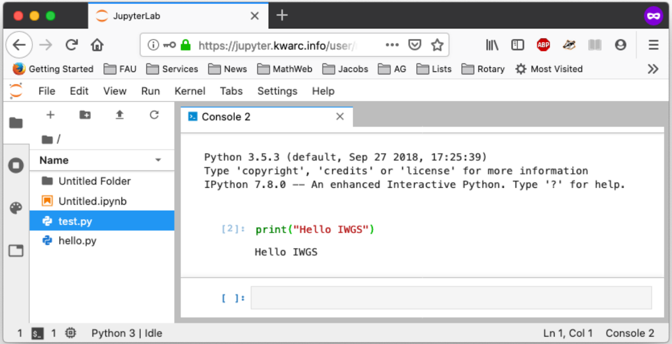


JupyterLab Components

- **Definition 3.14.** The JupyterLab dashboard gives you access to all components.
- **Definition 3.15.** The JupyterLab python console, i.e. a Python interpreter in your browser. (use this for Python interaction and testing.)
- **Definition 3.16.** The JupyterLab terminal, i.e. a `UNIX` shell in your browser.(use this for managing files)
- **Definition 3.17.** 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 ...

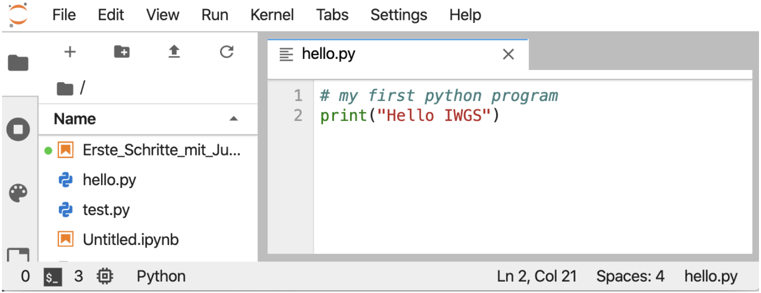
A first program in Python

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



A first program in Python

- **A classic "Hello World" program:** start your python console, type `print("Hello␣IWGS")`. (print a string)
- **Alternatively:**
  1. got to the JupyterLab dashboard select "Text File",
  2. Type your program,



  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)

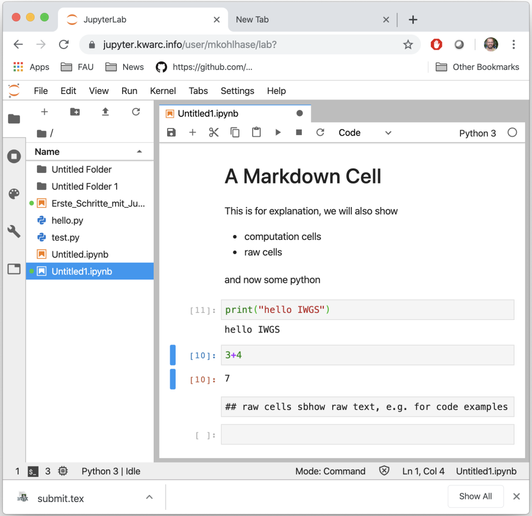
```
import hello
```

jupyter Notebooks

- **Definition 3.18.** Jupyter notebooks are documents that combine live runnable code with rich, narrative text (for comments and explanations).
- **Definition 3.19.** 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 3.20 (Showing off Cells in a Notebook).**



Markdown a simple Markup Format Generating HTML

.

- **Idea:** We can translate between markup formats.
- **Definition 3.21.** 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 3.22.** 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

Variables in Python

- **Idea:** Values (of expressions) can be given a name for later reference.
- **Definition 3.23.** A variable is an (the variable name) that references a memory location which contains a .
- **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 3.24 (in Python).** A variable assignment ⟪var⟫=⟪val⟫ assigns a new value to a variable.
- **Example 3.25 (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
>>> |
```

Variables in Python: Extended Example

- **Example 3.26 (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 3.27 (Variables for Storing Intermediate Variables).**

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

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 3.28.** 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.

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 3.29.** *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 3.30 (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

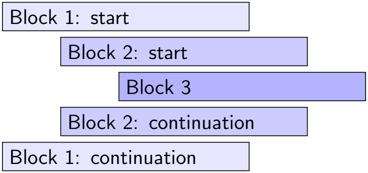
Conditionals and Loops

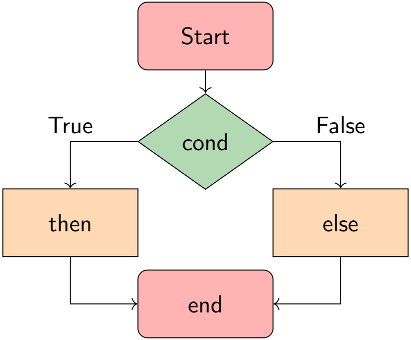
- **Problem:** Up to now programs seem to execute all the instructions in sequence, from the first to the last. (a linear program)
- **Definition 3.31.** The control flow of a program is the sequence of execution of the program instructions. It is specified via special program instructions called control structures.
- **Definition 3.32.** 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 3.33.** A condition is a Boolean expression in a control structure.
- **Definition 3.34.** 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.
- **Example 3.35.** 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

Conditionals in Python

- **Definition 3.36.** 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.

Conditional Execution Example

- **Example 3.37 (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.

Variant: Multiple Branches

- Making multiple branches is similar

  if $\langle\langle \text{condition} \rangle\rangle$:
  $\quad\langle\langle \text{then} - \text{part} \rangle\rangle$
  elif $\langle\langle \text{condition} \rangle\rangle$:
  $\quad\langle\langle \text{otherthen} - \text{part} \rangle\rangle$
  else :
  $\quad\langle\langle \text{else} - \text{part} \rangle\rangle$

  - The there can be more than one **elif** clause.
  - The conditions 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 3.38 (Better Empathy in Python).** In 3.37 we print Good! even if the input is e.g. I feel terrible, so extend **if**/**else** by

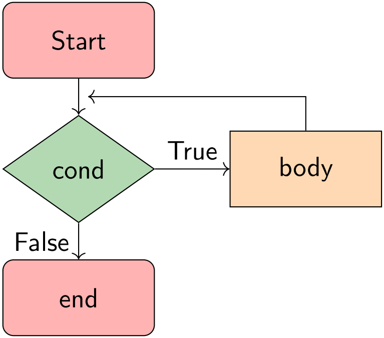
```
elif answer == 'Yes' or answer == 'yes' :
    print("Good!")
else :
    print("I␣do␣not␣understand␣your␣answer")
```

Loops in Python

- **Definition 3.39.** 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!



Examples of Loops

- **Example 3.40 (Counting in python).**

```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 3.41 (Breaking an unbounded Loop).**

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

Examples of Loops

- **Example 3.42 (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

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)

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 4.5 (Deep Fundamental Result).* There are things no computer can compute.
- **Example 4.6.** There cannot be a program that decides whether another program will terminate in finite time.

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 4.9 (Deep Fundamental Result).* There are things no computer can compute.
- **Example 4.10.** There cannot be a program that decides whether another program will terminate in finite time.
- *Remark 4.11 (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 4.12 (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".

Top Principle of Programming: Compositionality

- **Observation 4.13.** *Modern programming languages compose various primitives and give them a pleasing, concise, and uniform syntax.*
- **Question:** What does all of this even mean?
- **Definition 4.14.** 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 4.15 (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 4.16.** *The semantics of a programming language, is determined by the meaning of its primitives and composition principles.*
- **Definition 4.17.** 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.

Consequences of Compositionality

- **Observation 4.18.** *To understand a programming language, we (only) have to understand its primitives, composition principles, and their syntax.*
- **Definition 4.19.** The "art of programming" consists of composing the primitives of a programming language.
- **Observation 4.20.** *We only need very few – about half a dozen – primitives to obtain a Turing complete programming language.*
- **Observation 4.21.** *The space of program behaviors we can achieve by programming is infinites large nonetheless.*
- *Remark 4.22.* More primitives make programming more convenient.
- *Remark 4.23.* Primitives in one language can be composed in others.

A note on Programming: Little vs. Large Languages

- **Observation 4.24.** *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

2.5.1 Sequences and Iteration

Lists in Python

- **Definition 5.1.** A list is a finite sequence of objects, its element.
- 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 5.2.** We call [⟪seq⟫] the list constructor.
- **Example 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 5.4.** List elements can be accessed by specifying ranges

```
>>> list1[0]
'physics'
```

```
>>> list1[-2]
1997
```

```
>>> list2[1:4]
[2, 3, 4]
```

Sequences in Python

- **Definition 5.5.** 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 5.6.** A range is a finite sequence of numbers it can conveniently be constructed by the range function: range(⟪start⟫,⟪stop⟫,⟪step⟫) construts a range from ⟪start⟫ (inclusive) to ⟪stop⟫ (exclusive) with step size ⟪step⟫.
- **Example 5.7.** 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.

Iterating over Sequences in Python

- **Definition 5.8.** 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 5.9.** 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 5.10.** 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)
```

Python Dictionaries

- **Definition 5.11.** 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 5.12.** Dictionaries can be used for various purposes,

```
painting = {
    "artist": "Rembrandt",
    "title": "The Night Watch",
    "year": 1642
}
```

```
dict_de_en = {
  "Maus": "mouse",
  "Ast": "branch",
  "Klavier": "piano"
}
```

```
enum = {
  1: "copy",
  2: "paste",
  3: "adapt"
}
```

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

Interacting with Dictionaries

- **Example 5.13 (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 5.14 (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.

2.5.2 Input and Output

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)

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

- **Definition 5.15.** A file is a resource for recording data in a storage device. File size is measured in bit.
- **Definition 5.16.** 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 folder 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 5.17.** 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.

Disk Input/Output in Python

- **Definition 5.18.** 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 5.19 (Duplicating the contents of a file).**

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

Disk Input/Output in Python (continued)

- **Example 5.20 (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)

2.5.3 Functions and Libraries in Python

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 5.21.** 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 5.22 (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.

Functions in Python (Definition)

- **Definition 5.23.** 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.

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 3.28 are classes, ...
- **Definition 5.24.** 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:
  $\langle\langle \mathrm{obj}\rangle\rangle.\langle\langle \mathrm{meth}\rangle\rangle(\langle\langle \mathrm{args}\rangle\rangle)$ corresponds to $\langle\langle \mathrm{meth}\rangle\rangle(\langle\langle \mathrm{obj}\rangle\rangle,\langle\langle \mathrm{args}\rangle\rangle)$.
- **Example 5.25.** 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 5.26 (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).

Python Libraries

- **Idea:** Functions, classes, and methods are re usable, so why not package them up for others to use.
- **Definition 5.27.** 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 $\geq$ 150.000 libraries for Python ($\widehat{=}$ 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)

2.5.4 A Final word on Programming in IWGS

For more information on Python

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

Chapter 3
Numbers, Characters, and Strings

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).

3.1 Representing and Manipulating Numbers

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)

Unary Natural Numbers on the Computer

- **Definition 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 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	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 /)

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 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 $\mathcal{N}$.
  - $\varphi\colon D\rightarrow[0,b-1]$ 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 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

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

Commonly Used Positional Number Systems

- **Definition 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)
  - $110001101011100_2 = \underbrace{0110_2}_{6_{16}}\underbrace{0011_2}_{3_{16}}\underbrace{0101_2}_{5_{16}}\underbrace{1100_2}_{C_{16}} = 635C_{16}$
  - $110001101011100_2 = \underbrace{110_2}_{6_8}\underbrace{001_2}_{1_8}\underbrace{101_2}_{5_8}\underbrace{011_2}_{3_8}\underbrace{100_2}_{4_8} = 61534_8$
  - $F3A_{16} = \underbrace{F_{16}}_{1111_2}\underbrace{3_{16}}_{0011_2}\underbrace{A_{16}}_{1010_2} = 111100111010_2$, $4721_8 = \underbrace{4_8}_{100_2}\underbrace{7_8}_{111_2}\underbrace{2_8}_{010_2}\underbrace{1_8}_{001_2} = 100111010001_2$

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 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 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 1.8 (Convert 456 to base 8).** Result: 710_8

$$\begin{array}{ll} 456 \operatorname{div} 8 = 57 & 456 \bmod 8 = 0 \\ 57 \operatorname{div} 8 = 7 & 57 \bmod 8 = 1 \\ 7 \operatorname{div} 8 = 0 & 7 \bmod 8 = 7 \end{array}$$

3.2 Characters and their Encodings: ASCII and UniCode

The ASCII Character Code

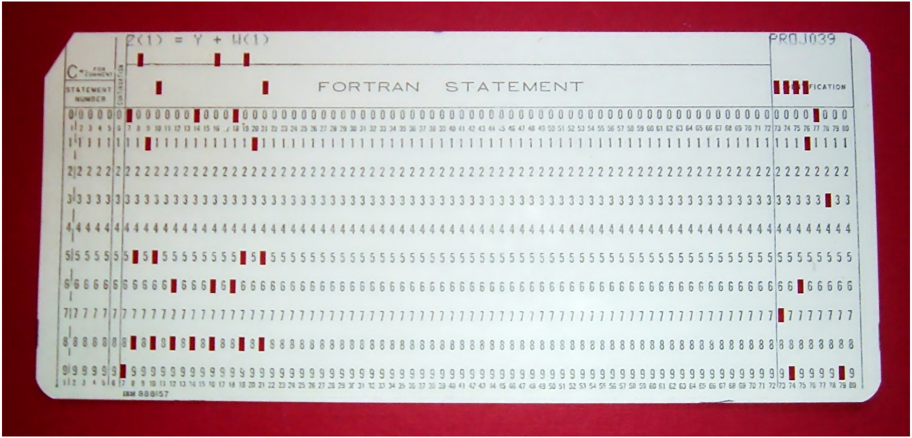
- **Definition 2.1.** The American Standard Code for Information Interchange (ASCII) is a character encoding that assigns characters to numbers 0 127.

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	[	\	]	^	_
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)

A Punchcard

- **Definition 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 2.3.** This punch card encodes the FORTRAN statement
  Z(1) = Y + W(1)



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 2.4 (German ASCII).** Remap e.g. [$\mapsto$Ä,]$\mapsto$Ü in German ASCII ("Apple][" comes out as "`Apple ÜÄ`")
- **Definition 2.5 (ISO-Latin (ISO/IEC 8859)).** 16 Extensions of ASCII to 8-bit (256 characters) ISO Latin 1 $\hat{=}$ "Western European", ISO Latin 6 $\hat{=}$ "Arabic", ISO Latin 7 $\hat{=}$ "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)

Unicode and the Universal Character Set

- **Definition 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 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 2.8.** Most (non-Chinese) characters have code points in $[1,65536]$: the basic multilingual plane (BMP).
- **Definition 2.9 (Notation).** For code points in the (BMP), four hexadecimal digits are used, e.g. $U+0058$ for the character $\mathrm{LATINCAPITALLETTERX}$;

Character Encodings in Unicode

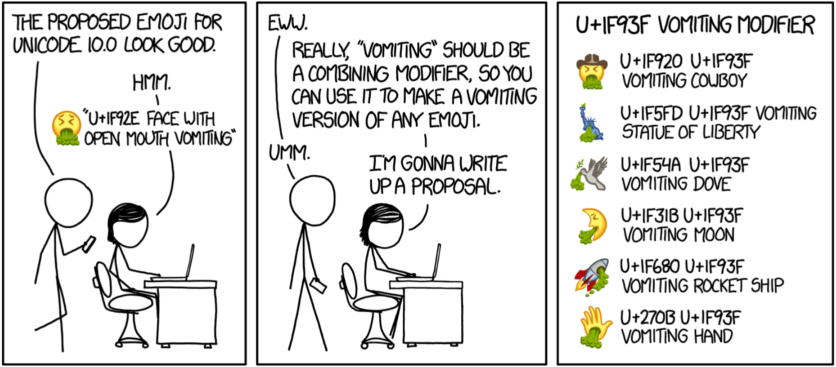
- **Definition 2.10.** A character encoding is a mapping from bit strings to UCS code points.
- **Idea:** Unicode supports multiple character encodings (but not character sets) for efficiency.
- **Definition 2.11 (Unicode Transformation Format).**
  - $\mathrm{UTF}-8$, 8-bit, variable width character encoding, which maximizes compatibility with ASCII.
  - $\mathrm{UTF}-16$, 16-bit, variable width character encoding (popular in Asia)
  - $\mathrm{UTF}-32$, a 32-bit, fixed width character encoding (as a fallback)
- **Definition 2.12.** The $\mathrm{UTF}-8$ 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 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)

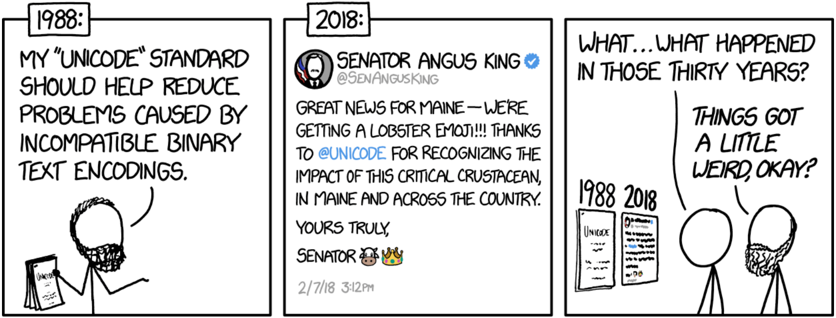
XKCD's Take on Recent Unicode Extensions

- UniCode 6.0 adopted hundreds of emoji characters in 2010 (2666 in July 2017)
- Modifying characters (https://xkcd.com/1813/)



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

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



3.3 More on Computing with Strings

Playing with Strings and Characters in Python

- **Definition 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 3.2 (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 $[i{:}j]$ $([i] \mathrel{\widehat{=}} [i{:}i])$
- **Example 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')
```

String Literals in Python

- **Problem:** How to write strings including special characters?
- **Definition 3.4.** A literal is a notation for representing a fixed value for a data structure in source code.
- **Definition 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 3.6.** 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 3.7.** 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.

Unicode in Python

- *Remark 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'
```

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 3.9.** 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.

  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 3.10 (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 3.11 (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

Anonymous Functions (lambda)

- **Observation 4.1.** *A Python function definition combines making a function object with giving it a name.*
- **Definition 4.2.** Python also allows to make anonymous functions via the function literal lambda for function objects:

  lambda $p_1, \ldots, p_n$: $\langle\langle\mathrm{expr}\rangle\rangle$

- **Example 4.3.** The following two Python fragments are equivalent:

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

```
cube = lambda 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)

Higher-Order Functions in Python

- **Definition 4.4.** We call a function a higher order function, iff it takes a function as argument.
- **Definition 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 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]
```

Argument Passing in Python: Keyword Arguments

- **Definition 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 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")
```

Argument Passing in Python: Flexible Arity

- **Definition 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 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 4.11.** The star operator unpacks a list into an argument sequence.
- **Example 4.12 (Passing a starred list).**

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

Argument Passing in Python: Flexible Keyword Arguments

- **Definition 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 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
```

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

- **Definition 4.15.**3 The double star operator unpacks a dictionary into a sequence of keyword arguments.
- **Example 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 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
```

3.5 Regular Expressions: Patterns in Strings

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: 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)

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 ⇝ regular expressions.

Regular Expressions, see [Pyt]

- **Definition 5.1.** A regular expression (also called regex) is a formal expression that specifies a set of strings.
- **Definition 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 \: \?.

Regular Expression Examples

- **Example 5.3 (Regular Expressions and their Values).**

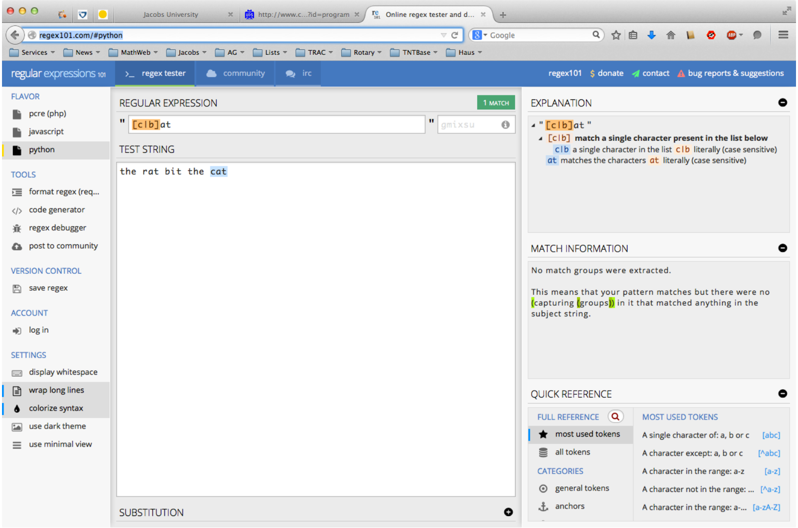
regexp	values
car	car
.at	cat, hat, mat, ...
[hc]at	cat, hat
[^c]at	hat, mat, ... (but not cat)
^[hc]at	hat, cat, but only at the beginning of the line
[0-9]	Digits
[1-9][0-9]*	natural numbers
(.*)\1	mama, papa, wakawaka
cat|dog	cat, dog

- 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 5.4 (A more complex example).** The following regex 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)))?$
```

Playing with Regular Expressions

- If you want to play with regexs, go e.g. to `http://regex101.com`



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 *pmetavar pat*.

    ```
    >>> 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']
    ```

Example: Correcting and Anonymizing Documents

- **Example 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)
```

Example: Correcting and Anonymizing Documents (cont.)

- **Example 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?

Example: Correcting and Anonymizing Documents (all)

- **Example 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
```

Chapter 4
Documents as Digital Objects

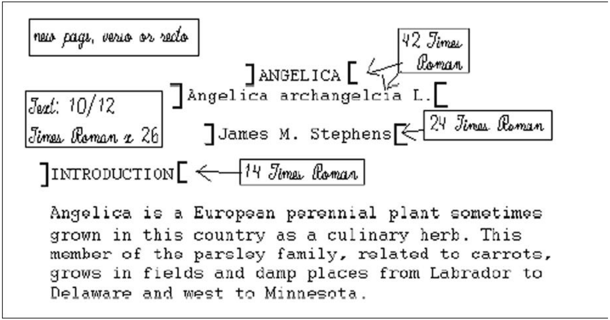
4.1 Representing & Manipulating Documents on a Computer

Electronic Documents

- **Definition 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 1.2.** PDFs, digital images, videos, audio recordings, web pages, ...
- **Definition 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 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 1.5.** Python programs are plain text, PDFs are formatted.

Document Markup

- **Definition 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 1.7.** A text with markup codes (for printing)



- **Definition 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 1.9.* Markup turns plain text into formatted text.

File Types

- **Observation 1.10.** *We mostly encounter electronic documents in the form of files on some storage medium.*
- **Definition 1.11.** A text file is a file that contains text data, a binary file one that contains binary data
- *Remark 1.12.* Text files are usually encoded with ASCII, ISO Latin, or increasingly $\mathrm{UniCode}$ encodings like $\mathrm{UTF}-8$.
- **Example 1.13.** 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 $\mathrm{U}+000\mathrm{A}$; $\mathrm{LINEFEED}(\mathrm{LF})$. The line number is just the position in the sequence.

Text Editors

- **Definition 1.14.** A text editor is a program used for rendering and manipulating text files.
- **Example 1.15.** Popular text editors include
  - `Notepad` is a simple editor distributed with `Windows`.
  - `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 1.16.** 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.

Word Processors and Formatted Text

- **Definition 1.17.** 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 1.18.** Popular word processors include
  - `MSWord`, an elaborated word processor for `Windows`, 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 `MacOSX` it uses a proprietary format.
  - `OfficeOnline` and `GoogleDocs` are browser-based real-time collaborative word processors.
- **Example 1.19.** Text editor are usually not considered to be word processors, even though they can sometimes be used to edit markup based formatted text.

4.2 Measuring Sizes of Documents/Units of Information

Units for Information

- **Observation:** The smallest unit of information is knowing the state of a system with only two states.
- **Definition 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 2.2.** The byte is a derived unit for information capacity: $1\mathrm{B} = 8\mathrm{b}$.

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 numbers, therefore the derived information units are prefixed by special prefixes that are based on powers of 2.
- **Definition 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 2.4.** You can buy hard-disks that say that their capacity is "one terabyte", but they actually have a capacity of one tebibyte.

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.

4.3 Hypertext Markup Language

4.3.1 Introduction

HTML: Hypertext Markup Language

- **Definition 3.1.** The HyperText Markup Labnguage (HTML), is a representation format for web pages [Hic+14].
- **Definition 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)

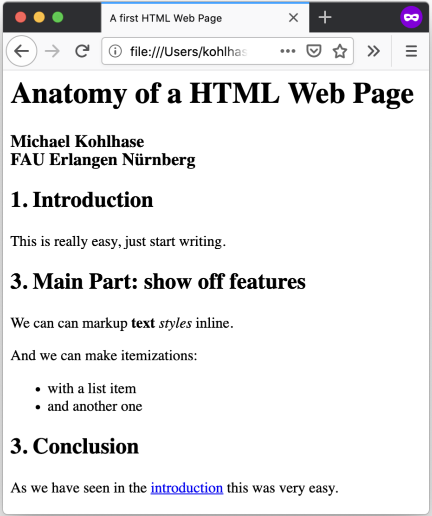
- **Example 3.3.** A (very simple) HTML file with a single paragraph.

```
<html>
  <body>
    <p>Hello IWGS students!</p>
  </body>
</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>
```

A very first HTML Example (Result)



4.3.2 Interacting with HTML in Web Broswers

Web Browsers

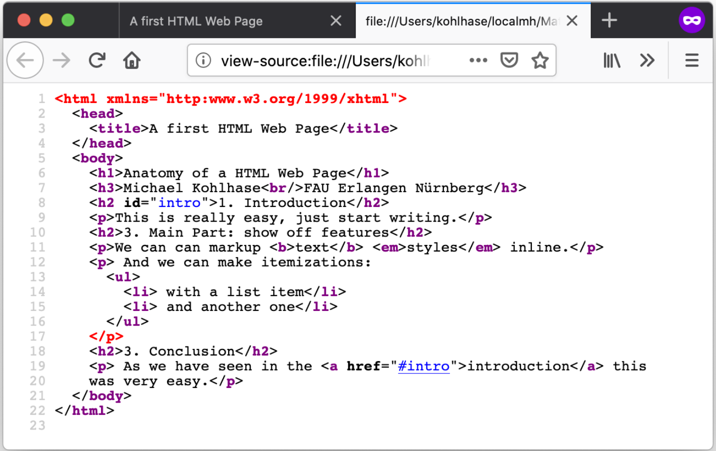
- **Definition 3.4.** A web browser is a software application for retrieving (via HTTP), presenting, and traversing information resources on the WWW, enabling users to view web pages and to jump from one page to another.

  Definition 3.5. A web browser usually supplies user tools like
  - history that gives the user access to the
  - an inspector to inspect the DOM

  Definition 3.6. A web browser usually supplies developer tools like
  - the console that logs system-level events in the browser
- **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 3.7 (Common Browsers).**
  - `MSInternetExplorer` is an once dominant, now obsolete browser for `Windows`.
  - `Edge` is provided by Microsoft for `Windows`. (replaces `MSInternetExplorer`)
  - `FireFox` is an open source browser for all platforms, it is known for its standards compliance.
  - `Safari` is provided by Apple for `MacOSX` and `Windows`.
  - `Chrome` is a lean and mean browser provided by Google Inc. (very cor

Browser Tools for dealing with HTML, e.g. in FireFox

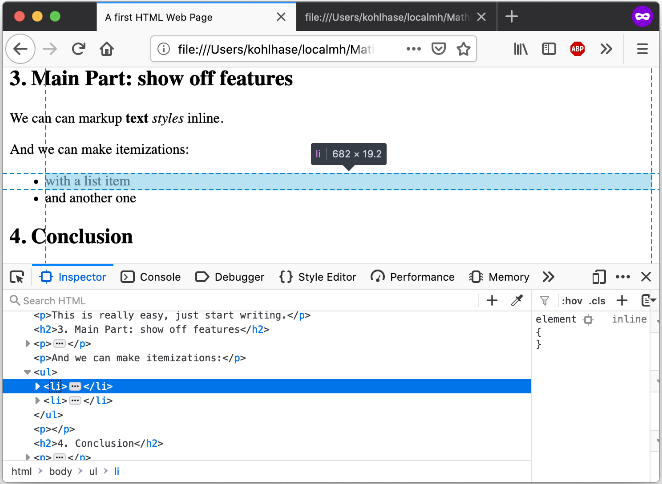
- Hit Control-U to see the page source in the browser




```html
<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>
```

Browser Tools for dealing with HTML, e.g. in FireFox

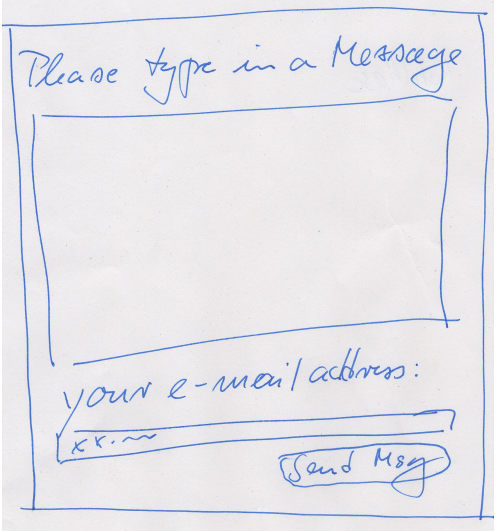
- Hit Control-U to see the page source in the browser
- go to an element and right-click ⇝ "Inspect element"



4.3.3 A Worked Example: The Contact Form

HTML in Practice: Worked Example

- Make a design and "paper prototype" of the page:



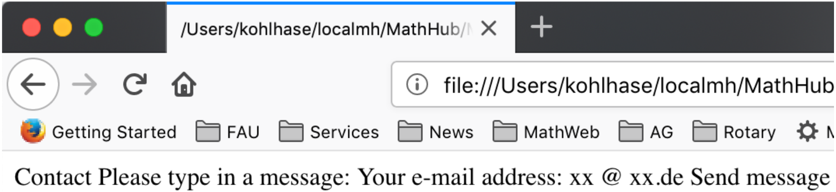
HTML in Practice: Worked Example

- 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
```

HTML in Practice: Worked Example

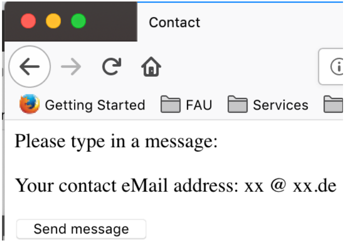
- Make a design and "paper prototype" of the page:
- Put the intended text into a file: contact.html:
- Load into your browser to check the state:



HTML in Practice: Worked Example

- Make a design and "paper prototype" of the page:
- Put the intended text into a file: contact.html:
- Load into your browser to check the state:
- Add title, paragraph and button markup:

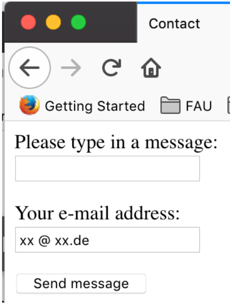
```
<title>Contact</title>
<h2>Please enter a message:</h2>
<h3>Your e—mail address: xx @ xx.de</h3>
<button>Send message</button>
```



HTML in Practice: Worked Example

- Make a design and "paper prototype" of the page:
- Put the intended text into a file: contact.html:
- Load into your browser to check the state:
- Add title, paragraph and button markup:
- 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>
```



HTML in Practice: Worked Example

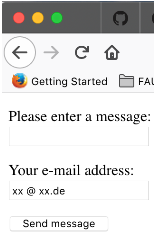
- Make a design and "paper prototype" of the page:
- Put the intended text into a file: contact.html:
- Load into your browser to check the state:
- Add title, paragraph and button markup:
- Add input fields and breaks:
- Convert into a HTML form with action (message receipt):

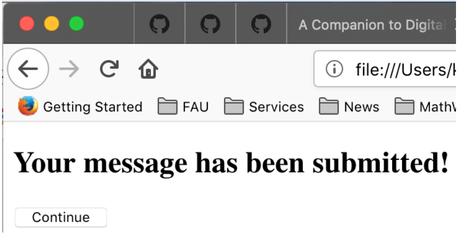
```
<title>Contact</title>
<form action="contact−after.html">
  <h2>Please enter a message:</h2>
    <input name="msg" type="text"/>
  <h3>Your e−mail address:</h3>
  <input name="addr" type="text"
          value="xx␣@␣xx.de"/>
  <br/>
  <input type="submit"
         value="Send␣message"/>
</form>
```

```
<title>
  Contact − Message Confirmed
</title>
<form action="contact4.html">
  <h2>
    Your message has been submitted!
  </h2>
  <input type="submit"
         value="Continue"/>
</form>
```

HTML in Practice: Worked Example

- Make a design and “paper prototype” of the page:
- Put the intended text into a file: contact.html:
- Load into your browser to check the state:
- Add title, paragraph and button markup:
- Add input fields and breaks:
- Convert into a HTML form with action (message receipt):





HTML in Practice: Worked Example

- Make a design and "paper prototype" of the page:
- Put the intended text into a file: contact.html:
- Load into your browser to check the state:
- Add title, paragraph and button markup:
- Add input fields and breaks:
- Convert into a HTML form with action (message receipt):
- That's as far as we will go, the rest is page layout and interaction. (up next)

HTML Forms

- **Question:** But how does the interaction with the contact form really work?
- **Definition 3.8.** The HTML form tags groups the layout and input elements:
  - <**form action**="⟪URI⟫"...> specifies the form action (as a web page address).
  - the input element <**input type**="submit"...> triggers the form action: it sends the form data to web page specified there.
- **Example 3.9 (In the Contact Form).** We send the request

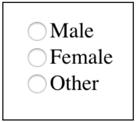
  ```
  GET contact−after.html?
      msg=Hi;addr=foo@bar.de
  ```

  We current ignore the form data (the part after the ?)
- We will come to the full story of processing actions later.

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
```



More useful types of Input fields

- Radio buttons: type="radio" (grouped by name attribute)
- 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
```

More useful types of Input fields

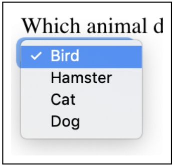
- Radio buttons: type="radio" (grouped by name attribute)
- Check boxes: type="checkbox"
- File selector dialogs (interaction is system specific here for MacOS Mojave)

```
<p> Upload your resume <input type="file" name="resume"/></p>
```

More useful types of Input fields

- Radio buttons: type="radio" (grouped by name attribute)
- Check boxes: type="checkbox"
- File selector dialogs (interaction is system specific here for MacOS Mojave)
- 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

Well-Bracketed Structures in Computer Science

- **Observation 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)*

Well-Bracketed Structures in Computer Science

- **Observation 4.2.** *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>
```

Well-Bracketed Structures in Computer Science

- **Observation 4.3.** *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:*
  - *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?")
```

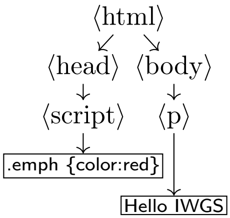
Well-Bracketed Structures in Computer Science

- **Observation 4.4.** *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:*
  - *Programming languages like python:*
- **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 4.5.** *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 4.6.** 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>
```



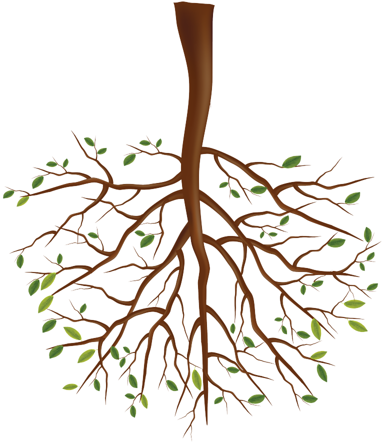
- **Definition 4.7.** We call such structures tree. (more on trees next)

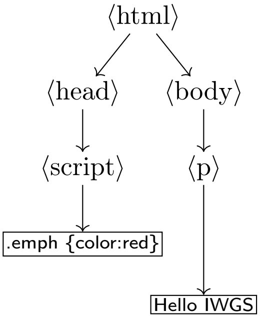
Well-Bracketed Structures: Tree Nomenclature

- **Definition 4.8.** In mathematics and CS, such well-bracketed structures are called trees (with root, branches, leaves, and height). (but written upside down)

Well-Bracketed Structures: Tree Nomenclature

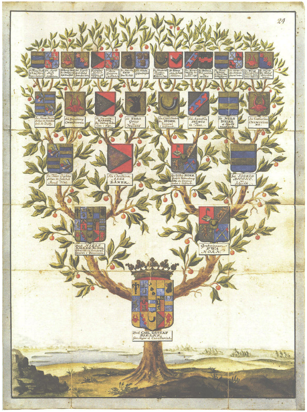
- **Definition 4.11.** In mathematics and CS, such well-bracketed structures are called trees (with root, branches, leaves, and height). (but written upside down)
- **Example 4.12.** In a tree, there is only one path from the root to the leaves

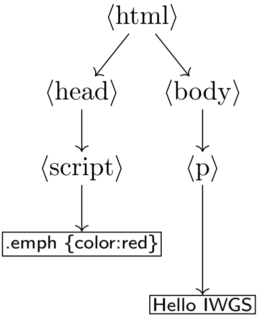




Well-Bracketed Structures: Tree Nomenclature

- **Definition 4.14.** In mathematics and CS, such well-bracketed structures are called trees (with root, branches, leaves, and height). (but written upside down)
- **Example 4.15.** In a tree, there is only one path from the root to the leaves
- **Definition 4.16.** We speak of parent, child, ancestor, and descendant nodes (genealogy nomenclature).





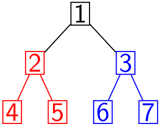
Upside Down Trees in Nature

- 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`

Computing with Trees in Python

- **Observation 4.17.** *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 4.18 (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:]))
```

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 4.19.** 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
```

The Document Object Model

- **Definition 4.20.** 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

4.5.1 Introduction to XML

XML (EXtensible Markup Language)

- **Definition 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)

XML is Everywhere (E.g. Web Pages)

- **Example 5.2.** Open web page file in FireFox, then click on *View* ↘ *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 5.3.** XHTML is the XML version of HTML. (just make it valid XML)

XML is Everywhere (E.g. Catalogs)

- **Example 5.4 (The NYC Galleries Catalog).** A public XML file at `https://data.cityofnewyork.us/download/kcrmj9hh/application/xml`

```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>
```

XML is Everywhere (E.g. Office Suites)

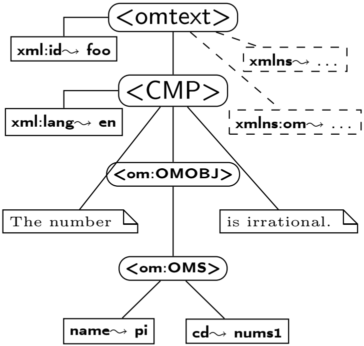
- **Example 5.5 (MS Office uses XML).** The `MSOffice` suite and `LibreOffice` use compressed XML as an electronic document format.
  1. Save a `MSOffice` file test.docx, add the extension .zip to obtain test.docx.zip.
  2. Uncompress with unzip (`UNIX`) or open File Explorer, right-click $\leadsto$ "Extract All" (`Windows`)
  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 5.6.** The XML document tree is made up of element nodes, attribute nodes, text nodes (and namespace declarations, comments,...)

XML Documents as Trees (continued)

- **Definition 5.7.** For communication this tree is serialized into a balanced bracketing structure, where
  - an inner element node 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 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

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)

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
```

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
```

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 node, you can use the append() method:

```
>>> root.append( etree.Element("child1") )
```

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 node, 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.

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.

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"/>'
```

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')]
```

Computing with XML in Python (Text nodes)

- 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>'
```

Case Study: Creating an HTML document

- We create nested html and body element

```
>>> 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>'
```

Computing with XML in Python (XML Literals)

- **Definition 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

XML is Everywhere (E.g. document metadata)

- **Example 5.9.** Open a PDF file in `AcrobatReader`, then click on

  $$\mathit{File} \searrow \mathit{DocumentProperties} \searrow \mathit{DocumentMetadata} \searrow \mathit{ViewSource}$$

  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 5.10.** 5.9 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

Mixing Vocabularies via XML Namespaces

- **Problem:** We would like to reuse elements from different XML vocabularies
  What happens if elements names coincide, but have different meanings?
- **Idea:** Disambiguate them by vocabulary name. (prefix)

Mixing Vocabularies via XML Namespaces

- **Problem:** We would like to reuse elements from different XML vocabularies
  What happens if elements 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

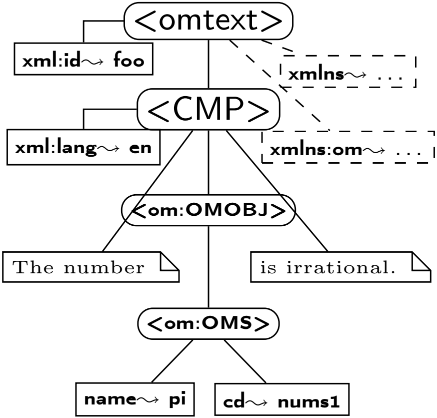
Mixing Vocabularies via XML Namespaces

- **Problem:** We would like to reuse elements from different XML vocabularies
  What happens if elements 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 5.15.** An XML namespace is a string that identifies an XML vocabulary. Every elements and attribute name in XML consists of a local name and a namespace.
- **Definition 5.16.** A namespace declaration is an attribute xmlns:prefix|=| whose value is an XML namespace *n* on an XML element *e*. The first associates the namepsace 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

XPath, A Language for talking about XML Tree Fragments

- **Definition 5.17.** The XML path language (XPath) is a language framework for specifying fragments of XML trees.
- **Intuition:**
  XPath is for trees what regular expressions are for strings.
- **Example 5.18.**



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

Computing with XML in Python (XPath)

- Say we have an XML tree:

```
>>> f = StringIO('<foo><bar></bar></foo>')
>>> tree = etree.parse(f)
```

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'
```

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'
```

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
```

XPath Example: Scraping Wikipedia

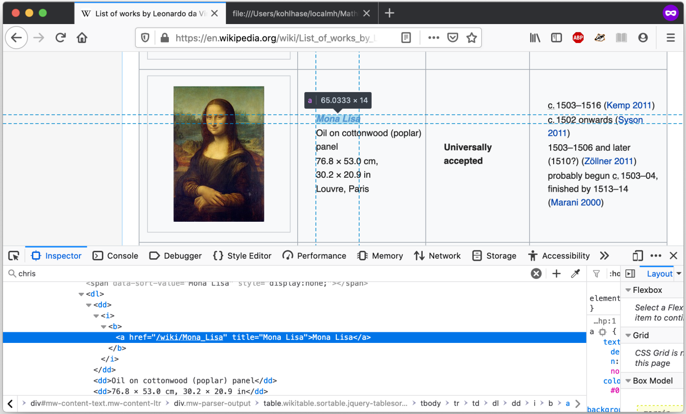
- **Example 5.19 (Extracting Information from HTML).**
  - We want a list of all titles of paintings by Leonardo da Vinci.

XPath Example: Scraping Wikipedia

- **Example 5.20 (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)

XPath Example: Scraping Wikipedia

- **Example 5.21 (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)



XPath Example: Scraping Wikipedia

- **Example 5.22 (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)
    The path is table > tbody > tr > td > dl > dd > i > b > a
    Alternatively: right-click on highlighted line, ⇝ "copy" ⇝ "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()')
```

Chapter 5
Web Applications

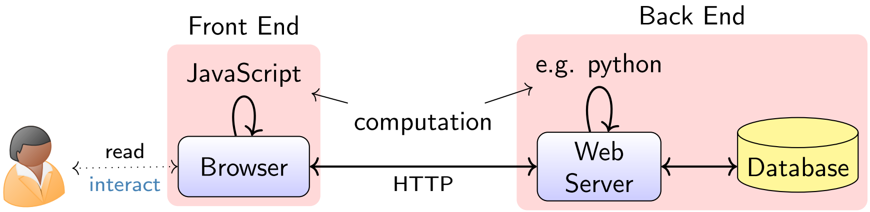
5.1 Web Applications: The Idea

Web Applications: Using Applications without Installing

- **Definition 1.1.** A web application 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 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.

Anatomy of a Web Application

- **Definition 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)

5.2 Basic Concepts of the World Wide Web

5.2.1 Preliminaries

The Internet and the Web

- **Definition 2.1.** The Internet is a global computer network that connects hundreds of thousands of smaller networks.
- **Definition 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)

Concepts of the World Wide Web

- **Definition 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 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 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 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

Uniform Resource Identifier (URI), Plumbing of the Web

- **Definition 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 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).

Relative URIs

- **Definition 2.9.** URIs can be abbreviated to relative URIs; missing parts are filled in from the context.
- **Example 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 2.11.** To distinguish them from relative URIs, we call URIs absolute URIs.

Uniform Resource Names and Locators

- **Definition 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 2.13.**
  The following URI is a URL (try it in your browser)
  http://kwarc.info/kohlhase/index.html
- **Example 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)

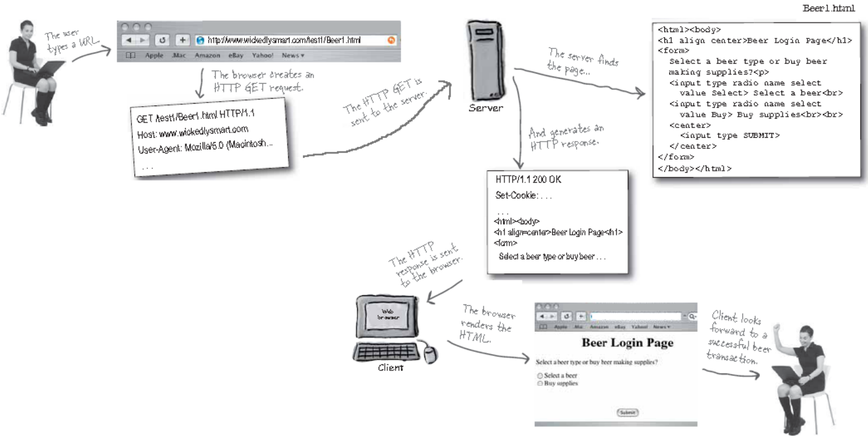
Internationalized Resource Identifiers

- *Remark 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 2.16.** Internationalized resource identifiers (IRIs) extend the ASCII-based URIs to the universal character set.
- **Definition 2.17.** URI encoding maps non-ASCII characters to ASCII strings:
  1. Map each character to its $\mathrm{UTF}-8$ representation.
  2. Represent each byte of the $\mathrm{UTF}-8$ 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 2.18.** The letter "ł" ($\mathrm{U}+142$) would be represented as `%C5%82`.
- **Example 2.19.** `http://www.Übergrößen.de` becomes `http://www.%C3%9Cbergr%C3%B6%C3%9Fen.de`
- *Remark 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 World Wide Web as a Client/Server System



HTTP: Hypertext Transfer Protocol

- **Definition 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 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 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 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.)

Web Servers

- **Definition 2.25.** Ein Web Server ist ein Netzwerk Programm (ein Server in der Client/Server Architektur des WWW) das über das Hypertext Transfer Protocol (HTTP) Web Resourcen an den Client ausliefert und Inhalte von ihm from erhält.
- **Example 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 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 `Windows` [Xam].

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

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 3.1.** Forms contain input fields and explanations.

```html
<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>
```

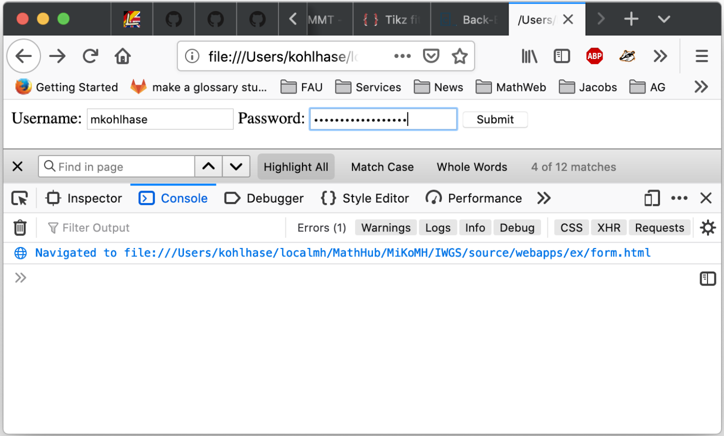
  yields the following in a web browser:

  Pressing the submit button activates a HTTP GET request to the URL login.html?user=$\langle\langle$name$\rangle\rangle$&pass=$\langle\langle$passwd$\rangle\rangle$

- ⚠ Never use the GET method for submitting passwords (see below)

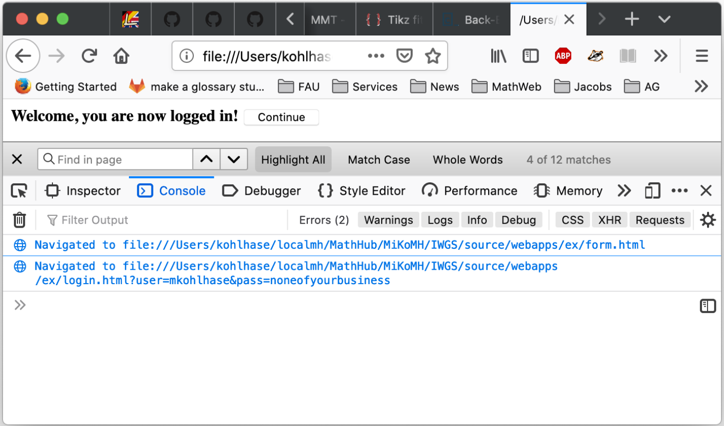
Checking up on the Transmission

- 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)



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)
- After submitting the form: (console logs the HTTP request)



HTML Forms and Form Data Transmission

- We specify the HTTP communication of HTML forms in detail.
- **Definition 3.2.** The HTML form element groups the layout and input elements:
  - <form action="⟪URI⟫" method="⟪req⟫"> specifies the form action in terms of a HTTP request ⟪req⟫ to the URI ⟪URI⟫.
  - The form data consists of a string ⟪data⟫ 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 ⟪req⟫ is get (the default), then the browser issues a GET request ⟪URI⟫?⟪data⟫.
    - If ⟪req⟫ is post, then the browser issues a POST request to ⟪URI⟫ with document content ⟪data⟫.
- We now also understand the form action, but should we use GET or POST.

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

Server-Side Scripting: Programming Web pages

- **Idea:** Why write HTML pages if we can also program them! (easy to do)
- **Definition 4.1.** A server-side scripting framework is a web server extension that generates web pages upon HTTP requests.
- **Example 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 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.

5.4.1 Routing and Argument Passing in Bottle

The Web Server and Routing in Bottle WSGI

- **Definition 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 given in ⟨⟨keys⟩⟩.
  - The @route decorator connects path components to Python function that return strings.
- **Example 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`

Dynamic Routes in Bottle

- **Definition 4.6.** A dynamic route is a route annotation that contains named wildcards, which can be picked up in the route function.
- **Example 4.7.** Multiple @route annotations per route function f are allowed $\leadsto$ 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)

Restricting Dynamic Routes

- **Definition 4.8.** A dynamic route can be restricted by a route filter to make it more selective.
- **Example 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 4.10 (A route with two wildcards).**

```
@route('/<action>/<user:re:[a-z]+>') # matches /follow/miko
def user_api(action, user):
    ...
```

Method-Specific Routes: HTTP GET and POST

- **Definition 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 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)

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 4.13 (Login 2).**
  Continuing from 4.12: 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.

5.4.2 Templating in Python via STPL

What would we do in Python

- **Example 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 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)

Template Processing for HTML

- **Definition 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 4.17.** PHP (originally "Programmable Home Page Tools") is a very successful server-side scripting framework following this model.

stpl: the "Simple Template Engine" from Bottle

- **Definition 4.18.** Bottle WSGI supplies the template engine stpl (Simple Template Engine). (documentation at [STPL])
- **Definition 4.19.** A template engine for a document format F is a program that transforms templates, i.e. strings or files with 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).
- 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'
```

stpl Syntax and Template Files

- **But what about. . . :** HTML files with embedded Python?
- stpl uses template files (extension .tpl) for that.
- **Definition 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 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>
```

Template Functions

- **Definition 4.22.** stpl python supplies the template functions
  1. include($\langle\langle$tpl$\rangle\rangle$,$\langle\langle$vars$\rangle\rangle$), where $\langle\langle$tpl$\rangle\rangle$ is another template file and $\langle\langle$vars$\rangle\rangle$ a set of variable declarations (for $\langle\langle$tpl$\rangle\rangle$).
  2. defined($\langle\langle$var$\rangle\rangle$) for checking definedness $\langle\langle$var$\rangle\rangle$
  3. get($\langle\langle$var$\rangle\rangle$,$\langle\langle$default$\rangle\rangle$): return the value of $\langle\langle$var$\rangle\rangle$, or $\langle\langle$default$\rangle\rangle$.
  4. setdefault($\langle\langle$name$\rangle\rangle$,$\langle\langle$val$\rangle\rangle$)

Template Functions

- **Definition 4.25.** 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 4.26 (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')
```

Template Functions

- **Definition 4.28.** 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 4.29 (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 4.30 (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
```

State in Web Applications and Cookies

- **Recall:** Web applications contain multiple pages, HTTP is a stateless protocol.
- **Problem:** How do we pass state between pages? (e.g. username, password)

State in Web Applications and Cookies

- **Recall:** Web applications contain multiple pages, HTTP is a stateless protocol.
- **Problem:** How do we pass state between pages? (e.g. username, password)
- **Simple Solution:** Pass information along in query part of page URLs.
- **Example 4.34 (HTTP GET for Single Login).** Since we are generating pages we can generated augmented links

```
<a href="http://example.org/more.html?user=joe,pass=hideme">... more</a>
```

- **Problem:** Only works for limited amounts of information and for a single session.

State in Web Applications and Cookies

- **Recall:** Web applications contain multiple pages, HTTP is a stateless protocol.
- **Problem:** How do we pass state between pages? (e.g. username, password)
- **Simple Solution:** Pass information along in query part of page URLs.
- **Example 4.37 (HTTP GET for Single Login).** Since we are generating pages we can generated augmented links
- **Problem:** Only works for limited amounts of information and for a single session.
- **Other Solution:** Store state persistently on the client hard disk.
- **Definition 4.38.** A cookie is a text file stored on the client hard disk by the web browser. Web servers can request the browser to store and send cookies.

State in Web Applications and Cookies

- **Recall:** Web applications contain multiple pages, HTTP is a stateless protocol.
- **Problem:** How do we pass state between pages? (e.g. username, password)
- **Simple Solution:** Pass information along in query part of page URLs.
- **Example 4.40 (HTTP GET for Single Login).** Since we are generating pages we can generated augmented links
- **Problem:** Only works for limited amounts of information and for a single session.
- **Other Solution:** Store state persistently on the client hard disk.
- **Definition 4.41.** A cookie is a text file stored on the client hard disk by the web browser. Web servers can request the browser to store and send cookies.
- **Note:** Cookies are data, not programs, they do not generate pop ups or behave like viruses, but they can include your log-in name and browser preferences.
- **Note:** Cookies can be convenient, but they can be used to gather information about you and your browsing habits.
- **Definition 4.42.** Third-party cookies are used by advertising companies to track users across multiple sites. (but you can turn off, and even delete cookies)

5.4.3 Completing the Contact Form

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>
```

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>
```

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
```

Back to our Contact Form (Current State)

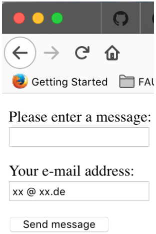
- A contact form and message receipt (communicate via HTTP requests)

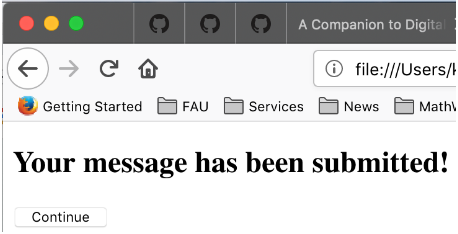
contact4.html

```
<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>
```

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>
```





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>
```

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>
```

- **Problem:** The answer is a static HTML document independent of form data.
- **Solution:** Generate the answer programmatically using the form data. (up next)

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 4.43 (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>
```

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 $\widehat{=}$ Simple Mail Transfer Protocol")
- **Example 4.44 (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")

Chapter 6
Frontend Technologies

6.1 Dynamic HTML: Client-side Manipulation of HTML Documents

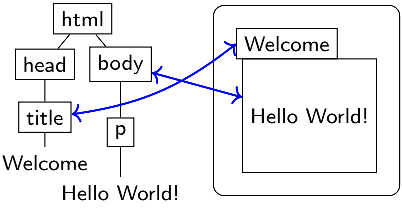
Background: Rendering Pipeline in browsers

- **Observation:** The nested markup codes turn HTML documents into trees.
- **Definition 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 Document **DOM** **Browser**

```
<html>
  <head>
    <title>Welcome</title>
  </head>
  <body>
    <p>Hello World!</p>
  </body>
</html>
```

parse →



The DOM is notified of any user events (resizing, clicks, hover,. . .)

6.1.1 JavaScript in HTML

Dynamic HTML

- **Idea:** generate parts of the web page dynamically by manipulating the DOM.
- **Definition 1.2.** 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 1.3.** We write the some text into a HTML document object (the document API)

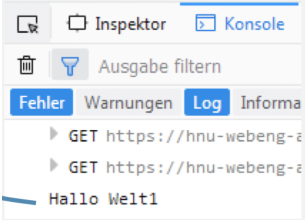
```
<html>
 <head>
  <script type="text/javascript">document.write("Dynamic␣HTML!");</script>
 </head>
 <body><!-- nothing here; will be added by the script later --></body>
</html>
```

- **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

Browser-level JavaScript functions: 1

- **Example 1.4 (Logging to the browser console).**

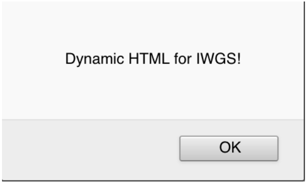
```
console.log("hello IWGS")
```



Browser-level JavaScript functions: 2

- **Example 1.6 (Raising a Popup).**

```
alert("Dynamic HTML for IWGS!")
```



Browser-level JavaScript functions: 3

- **Example 1.7 (Asking for Confirmation).**

```
var returnvalue = confirm("Dynamic HTML for IWGS!")
```

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

```
<script type="text/javascript" src="../js/foo.js"/>
```

  Advantage: HTML and JavaScript code are clearly separated

- In event attributes of various HTML elements, e.g.

```
<input type="button" value="Hallo" onclick="alert('Hello␣IWGS')"/>
```

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>
```

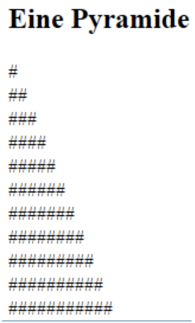
- JavaScript in an event handler attribute onclick, ondblclick, onmouseover, ...” whenever the corresponding event occurs.
- JavaScript in a “special link”: when the anchor is clicked

```
<a href="javascript:..."/>
```

Example: Changing Web Pages Programmatically

- **Example 1.9 (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>
```



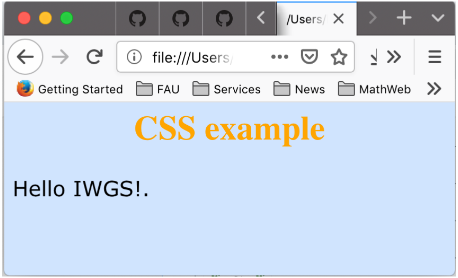
6.2 Cascading Stylesheets

6.2.1 Separating Content from Layout

CSS: Cascading Style Sheets

- **Idea:** Separate structure/function from appearance.
- **Definition 2.1.** Cascading Style Sheets (CSS) is a style sheet language that allows authors and users to attach style (e.g., fonts, colors, and spacing) to HTML and XML documents.
- **Example 2.2.** Our text file from 3.3 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: Rules, Selectors, and Declarations

- **Definition 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 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 2.5.** In 2.2 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.

A Styled HTML Title Box (Source)

- **Example 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;}
```

A Styled HTML Title Box (Result)

6.2.2 A small but useful Fragment of CSS

CSS Selectors

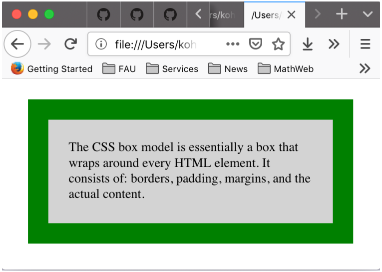
- **Question:** Which elements are affected by a CSS rule?
- Elements of a given name (optionally with given attributes)
  - Selectors: name $\widehat{=}$ ⟪elname⟫, attributes $\widehat{=}$ [⟪attname⟫=⟪attval⟫]
- **Example 2.7.** p[xml:lang='de'] applies to <**p** xml:**lang**="de">...</**p**>
- Any elements with a given class attributes
  - Selector: .⟪classname⟫
- **Example 2.8.** .important applies to <⟪el⟫ class='important'>...</⟪el⟫>
- The element with a given id attribute
  - Selector: #⟪id⟫
- **Example 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`.

The CSS Box Model

- **Definition 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:

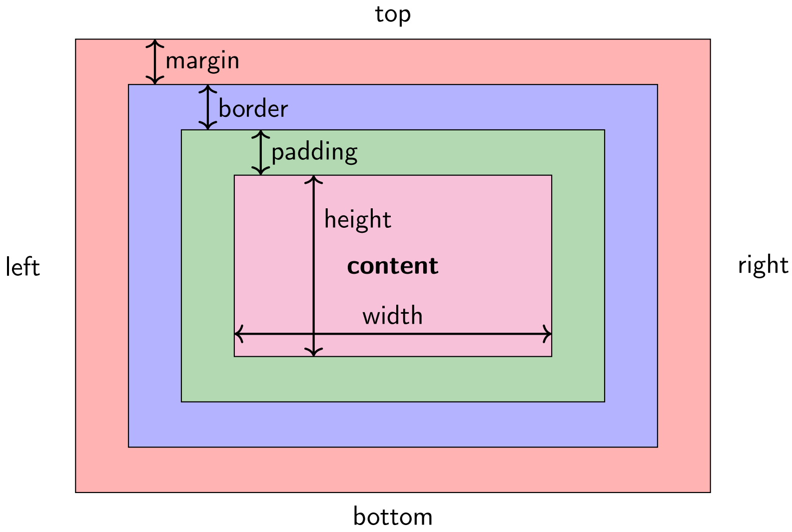
```
div {
    background—color: lightgrey;
    width: 300px;
    border: 25px solid green;
    padding: 25px;
    margin: 25px;
}
```



Note that the overall width of the CSS box is $300 + 2 \cdot 3 \cdot 25 = 450$ pixels.

The CSS Box Model: Diagram

- The following diagram summarizes the CSS box model



Cascading of selectors in CSS: Prioritization

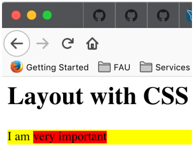
- Multiple CSS selectors apply with the following priorities:
  1. important (i.e. marked with !important) before unimportant
  2. inline (specified via the style attribute)
  3. media-specific rules before general ones
  4. user-defined CSS stylesheet (e.g. in the `FireFox` profile)
  5. specialized before general selectors (complicated; see e.g. [CSS])
  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

Cascading of selectors in CSS: Prioritization Example

- **Example 2.11.** 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}
```



Cascading in CSS: Inheritance

- **Definition 2.12.** If an element is fully contained in another, the inner inherits some properties (called inheritable) of the outer. 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 prioritization (recall case 7. above)
- Inheritance makes for consistent text properties and smaller CSS stylesheets.

CSS-Flow: How Boxes Flow to their Place

- CSS Flow describes how different elements are distributed in the visible area (how they flow; hence the name)
- **Example 2.13.** Block-level Boxes (here divs) flow 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;}
```

=

CSS-Flow: How Boxes Flow to their Place

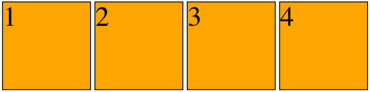
- CSS Flow describes how different elements are distributed in the visible area (how they flow; hence the name)
- **Example 2.17.** Block-level Boxes (here divs) flow to the left
- **Example 2.18.** float:left floats boxes as far as they will go (without overlap)

```html
<div class="square">1</div>
<div class="square">2</div>
<div class="square">3</div>
<div class="square">4</div>
```

+

```css
.square {font-size:200%;
        height:100px;
        width:100px;
        border:1px solid black;
        margin:2px;
        background-color:orange;
        float:left}
```

=



CSS-Flow: How Boxes Flow to their Place

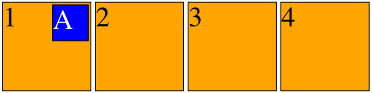
- CSS Flow describes how different elements are distributed in the visible area (how they flow; hence the name)
- **Example 2.21.** Block-level Boxes (here divs) flow to the left
- **Example 2.22.** float:left floats boxes as far as they will go (without overlap)
- **Example 2.23.** float:right in a 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}
```

=



CSS-Flow: How Boxes Flow to their Place

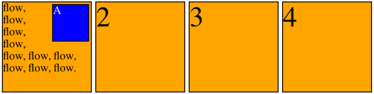
- CSS Flow describes how different elements are distributed in the visible area (how they flow; hence the name)
- **Example 2.25.** Block-level Boxes (here divs) flow to the left
- **Example 2.26.** float:left floats boxes as far as they will go (without overlap)
- **Example 2.27.** float:right in a div will float inside the corresponding box
- **Example 2.28.** 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}
```

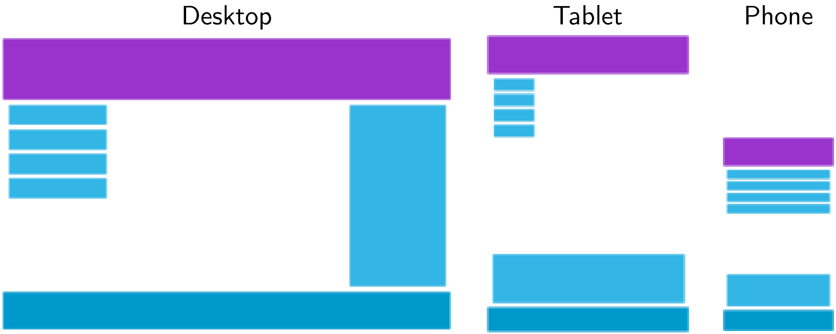
=



The large space (>2px) is caused because there is no linebreaking

CSS Application: Responsive Design

- **Problem:** What is the screen size/resolution of my device?
- **Definition 2.29.** 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 2.30.** A web page with content blocks

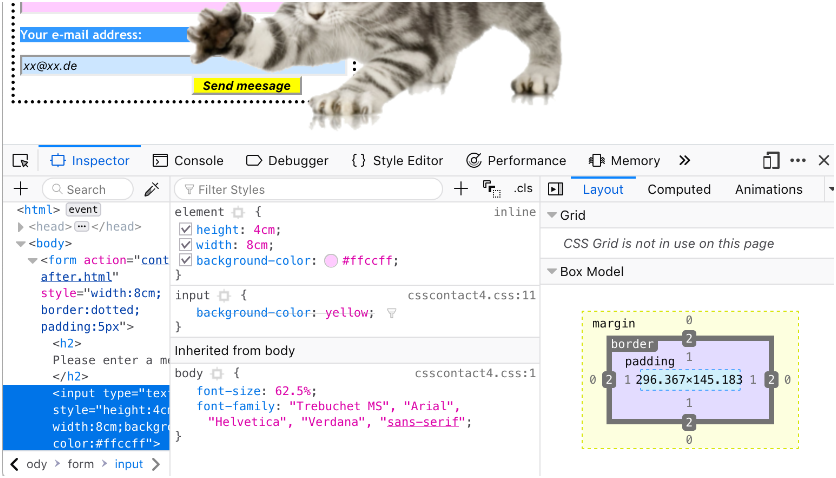


- **Implementation:** CSS based layout with relative sizes and media queries– CSS conditionals based on client screen size/resolution/...

6.2.3 CSS Tools

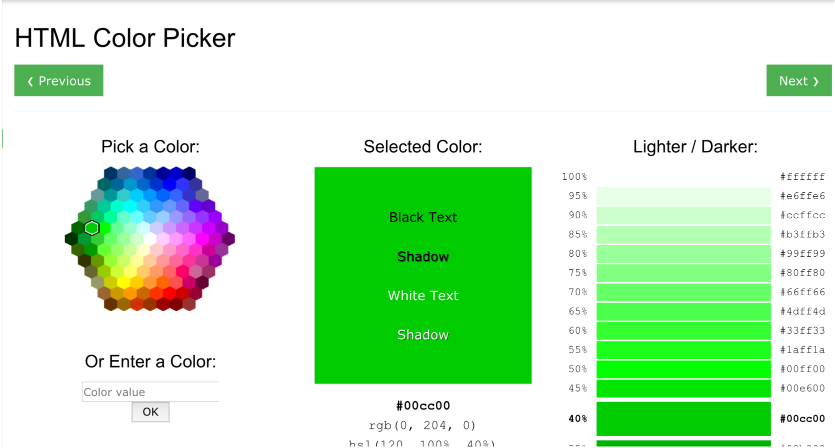
But how to find out what the browser really sees?

- CSS has many interesting inheritance rules
- **Definition 2.31.** The page inspector tool gives you an overview over the internal state of the browser.
- **Example 2.32.**



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`

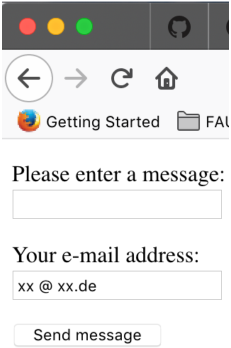


6.2.4 Worked Example: The Contact Form

CSS in Practice: The Contact Form Example (Continued)

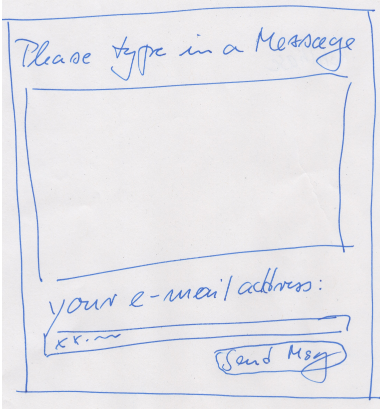
- Recap: The unstyled contact form –

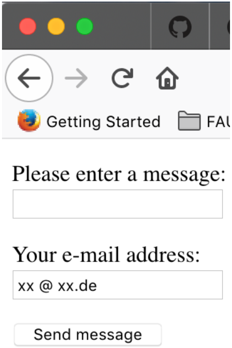
```
<title>Contact</title>
<form action="contact-after.html">
  <h2>Please enter a message:</h2>
    <input name="msg" type="text"/>
  <h3>Your e-mail address:</h3>
  <input name="addr" type="text"
             value="xx␣@␣xx.de"/>
  <br/>
  <input type="submit"
            value="Send␣message"/>
</form>
```



CSS in Practice: The Contact Form Example (Continued)

- Recap: The unstyled contact form – Dream vs. Reality

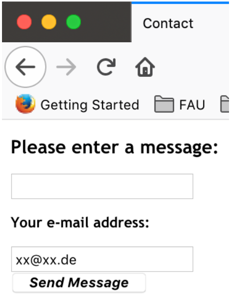




CSS in Practice: The Contact Form Example (Continued)

- Recap: The unstyled contact form – Dream vs. Reality
- Add a CSS file with font information

```
<link rel="stylesheet" type="text/css"
      href="csscontact1.css" />
  <input class="important" type="submit"
         value="Send␣Message"/>

body {font-size: 62.5%;
      font-family: "Trebuchet␣MS",
          "Arial", "Helvetica",
          "Verdana", "sans-serif"}
.important{font-style: italic;}
input[type="submit"]{font-weight: bold;}
```

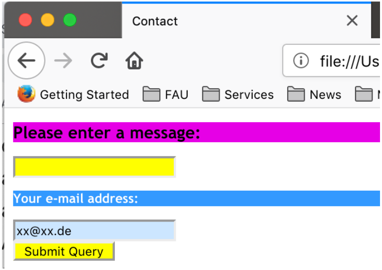


CSS in Practice: The Contact Form Example (Continued)

- Recap: The unstyled contact form – Dream vs. Reality
- Add a CSS file with font information
- 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"/>
```

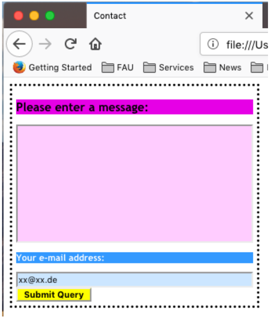
```
h2 {background−color: #e600e6;}
h3 {background−color: #3399ff;
    color: white;}
input{background−color:yellow}
```



CSS in Practice: The Contact Form Example (Continued)

- Recap: The unstyled contact form – Dream vs. Reality
- Add a CSS file with font information
- Add lots of color (ooops, what about the size)
- Add size information and a dotted frame

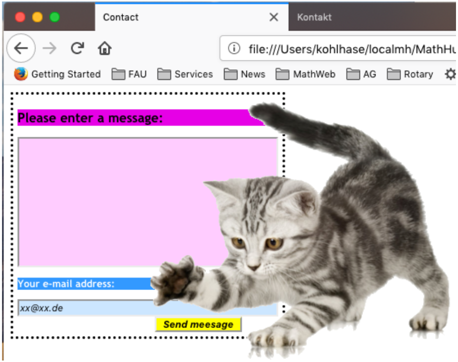
```
<form action="contact−after.html"
        style="width:8cm;border:dotted;padding:5px">
  <h2>Please enter a message:</h2>
  <input name="msg" type="text"
            style="height:4cm;width:8cm;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣background−color:#ffccff"/>
  <br/>
  <h3>Your e−mail address:</h3>
  <input class="important" name="addr"
            type="text"
            value="xx@xx.de" style="width:8cm;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣background−color:#cce6ff"/>
```



CSS in Practice: The Contact Form Example (Continued)

- Recap: The unstyled contact form – Dream vs. Reality
- Add a CSS file with font information
- Add lots of color (ooops, what about the size)
- Add size information and a dotted frame
- 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"/>
```



6.3 JQuery: Write Less, Do More

JQuery: Write Less, Do More

- **Definition 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" />
    ```

JQuery Philosophy and Layers

- **JQuery Philosophy:** Select an object from the DOM, and operate on it.
- **Syntax Convention:** JQuery instructions start with a $ to distinguish it from JavaScript.
- **Example 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 $\widehat{=}$ CSS selectors

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')
```

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 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 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 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.
  - first parameter of toggle() is a duration for the animation.
  - The second 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 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
    - plays it via the play() method.

6.4 Web Applications: Recap

What Tools have we seen so far?

- HTML (Hypertext Markup Language)
  - Text-based markup language for the web
  - tree structure (realized as the DOM in the browser)
    - easy search&find $\leadsto$ Selection
    - DOM changes easy by clear dependencies.

What Tools have we seen so far?

- HTML (Hypertext Markup Language)
  - Text-based markup language for the web
  - tree structure (realized as the DOM in the browser)
    - easy search&find $\leftrightsquigarrow$ Selection
    - DOM changes easy by clear dependencies.
- CSS (Cascading Stylesheets)
  - Language for specifying layout of HTML/DOM
  - CSS selection ties layout specifications into HTML/DOM

What Tools have we seen so far?

- HTML (Hypertext Markup Language)
  - Text-based markup language for the web
  - tree structure (realized as the DOM in the browser)
    - easy search&find $\leadsto$ Selection
    - DOM changes easy by clear dependencies.
- CSS (Cascading Stylesheets)
  - Language for specifying layout of HTML/DOM
  - CSS selection ties layout specifications into HTML/DOM
- Bottle (Server-Side web page generation via Python)
  - full programming language for comprehensive functionality
  - routes for complex but coherent web sites
  - template engine for HTML-centered web page design

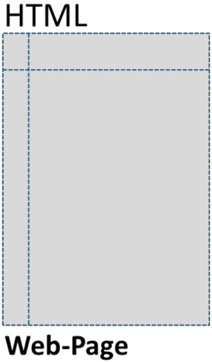
What Tools have we seen so far?

- HTML (Hypertext Markup Language)
  - Text-based markup language for the web
  - tree structure (realized as the DOM in the browser)
    - easy search&find $\leftrightsquigarrow$ Selection
    - DOM changes easy by clear dependencies.
- CSS (Cascading Stylesheets)
  - Language for specifying layout of HTML/DOM
  - CSS selection ties layout specifications into HTML/DOM
- Bottle (Server-Side web page generation via Python)
  - full programming language for comprehensive functionality
  - routes for complex but coherent web sites
  - template engine for HTML-centered web page design
- JavaScript (client-side scripting)
  - full programming language (Turing complete)
  - programmatic changes to the DOM $\rightsquigarrow$ dynamic HTML
    - navigating the DOM via JS-selection (relatively clumsy, but sufficient)
    - jQuery navigate the DOM via CSS-selection (reuses successful concepts)

Recap: Web Application Frontend

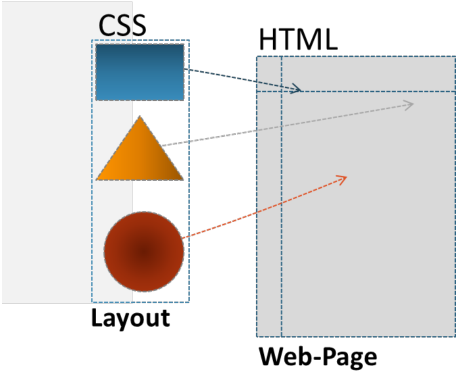
- **Recap: Web Application Frontend:**

  Web pages are just HTML files.



- **Recap: Web Application Frontend:**

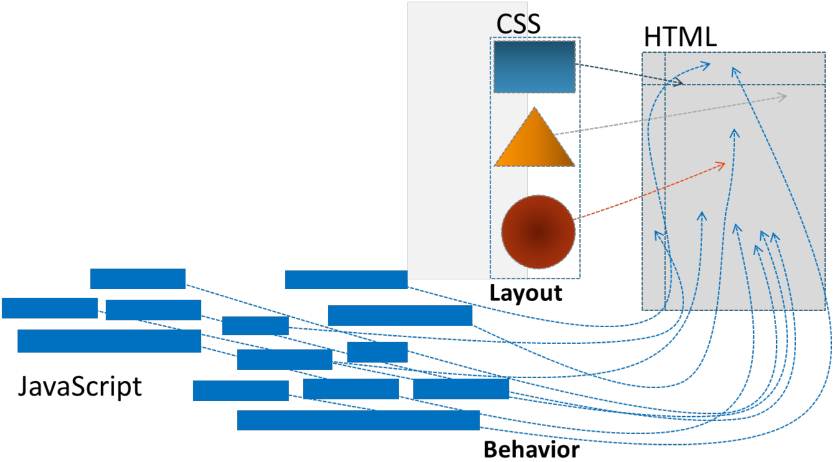
  Layout is specified by CSS instructions and selectors



Recap: Web Application Frontend

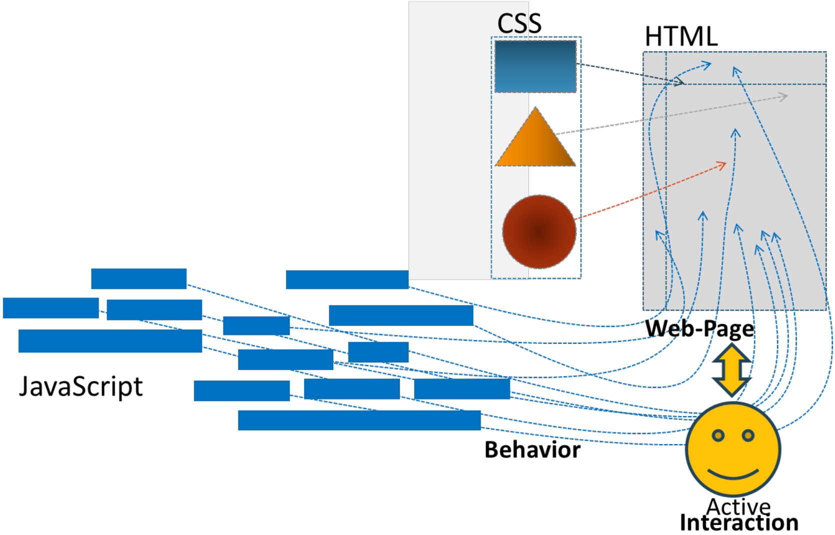
- **Recap: Web Application Frontend:**

Javascript specifies behavior



- **Recap: Web Application Frontend:**

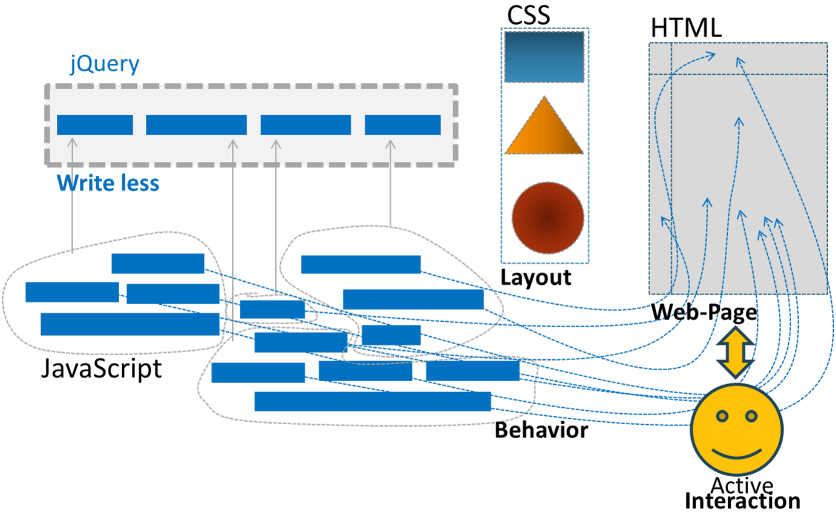
for interacting with the user



Recap: Web Application Frontend

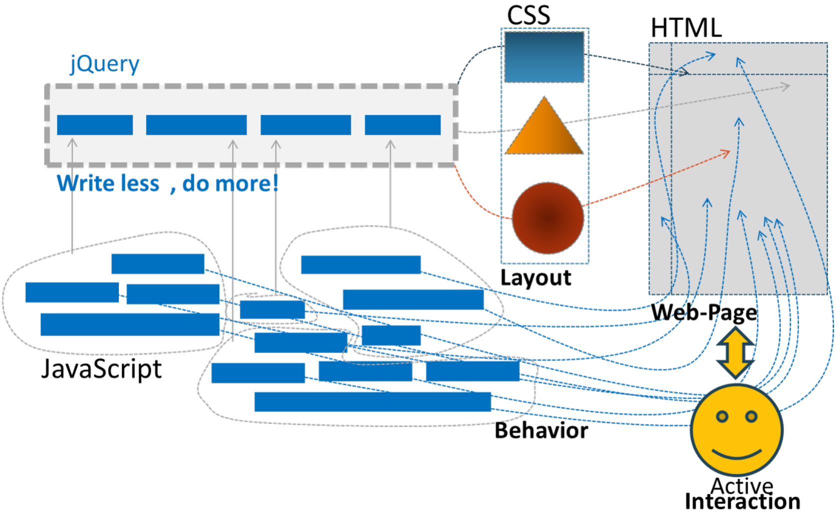
- **Recap: Web Application Frontend:**

  JQuery $\widehat{=}$ more succinct Javascript



Recap: Web Application Frontend

- **Recap: Web Application Frontend:**
  JQuery attaches behaviors to DOM elements via CSS selectors



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 strutures like numbers and strings, character encodings, unicode, and regular expressions
- Digital 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 servers
  - serverside computing: bottle routing and
  - client-side 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

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

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

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

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)

Informatische Werkzeuge in den Geistes- und Sozialwissenschaften 2

Prof. Dr. Michael Kohlhase

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

2024-02-08

Part 2
IWGS-II: DH Project Tools

Chapter 8
Semester Change-Over

8.1 Administrativa

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)

Assessment, Grades

- **Grading Background/Theory:** Only modules are graded! (by the law)
  - Module "DH-Einführung" (DHE) $\widehat{=}$ courses IWGS1/2, DH-Einführung.
  - DHE module grade $\rightsquigarrow$ pass/fail determined by "portfolio" $\widehat{=}$ collection of contributions/assessments.
- **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)

IWGS Homework Assignments

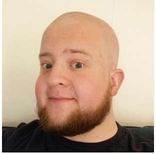
- **Homeworks:** will be small individual problem/programming/system assignments
  - but take time to solve (at least read them directly ⇝ questions)
  - group submission if and only if explicitly permitted.
- ⚠ Without trying the homework assignments you are unlikely to pass the exam.
- **Admin:** To keep things running smoothly
  - Homeworks will be posted on StudOn.
  - Sign up for IWGS under `https://www.studon.fau.de/frm5075965.html`.
  - Homeworks are handed in electronically there. (plain text, program files, PDF)
  - Go to the tutorials, discuss with your TA! (they are there for you!)
- **Homework Discipline:**
  - Start early! (many assignments need more than one evening's work)
  - Don't start by sitting at a blank screen (talking & study group help)
  - Humans will be trying to understand the text/code/math when grading it.

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)



- **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 slides 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

Textbook, Handouts and Information, Forums, Videos

- **No Textbook:** but lots of online python tutorials on the web.
- Course notes will be posted at `http://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 WS 2020/21 are at https://www.fau.tv/course/id/2350 (not much changed)
- Matrix chat at `#iwgs:fau.de` (via IDM) (instructions)
- **StudOn Forum:** `https://www.studon.fau.de/frm5075965.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 forum turns into a valuable resource!

Experiment: Learning Support with KWARC Technologies

- **My research area:** Deep representation formats for (mathematical) knowledge
- **One Application:** Learning support systems (represent knowledge to transport it)
- **Experiment:** Start with this course (Drink my own medicine)
  1. Re-represent the slide materials in OMDoc (Open Mathematical Documents)
  2. Feed it into the ALeA system (`http://courses.voll-ki.fau.de`)
  3. Try it on you all (to get feedback from you)
- Research tasks
  - help me complete the material on the slides (what is missing/would help?)
  - I need to remember "what I say", examples on the board. (take notes)
- Benefits for you (so why should you help?)
  - you will be mentioned in the acknowledgements (for all that is worth)
  - you will help build better course materials (think of next-year's students)

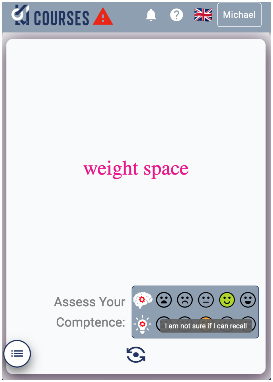
VoLL-KI Portal at https://courses.voll-ki.fau.de

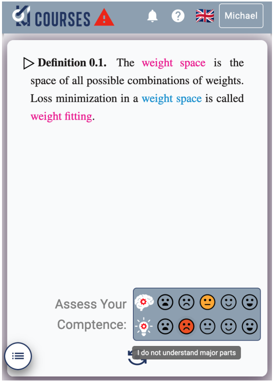
- **Portal for ALeA Courses:** https://courses.voll-ki.fau.de

- **AI-1 in ALeA:** https://courses.voll-ki.fau.de/course-home/ai-1
  - All details for the course.
  - recorded syllabus (keep track of material covered in course)
  - syllabus of the last semester (for over/preview)
- **ALeA Status:** The ALeA system is deployed at FAU for over 1000 students taking six courses
  - (some) students use the system actively (our logs tell us)
  - reviews are mostly positive/enthusiastic (error reports pour in)

New Feature: Drilling with Flashcards

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

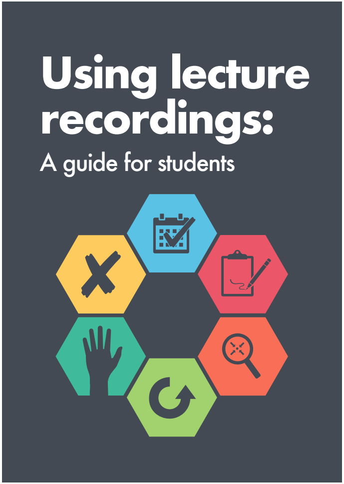




  ...and the definition/answer is on the back.
- Self-assessment updates the learner model (before/after)
- **Idea:** Challenge yourself to a card stack, keep drilling/assessing flashcards until the learner model eliminates all.
- **Bonus:** Flashcards can be generated from existing semantic markup (educational equivalent to free beer)

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.

- Normally intended for "offline students" $\widehat{=}$ everyone during Corona times.

Software/Hardware tools

- You will need computer access for this course
- we recommend the use of standard software tools
  - find a text editor you are comfortable with (get good with it) A text editor is a program you can use to write text files. (not `MSWord`)
  - any operating system you like (I can only help with `UNIX`)
  - Any browser you like (I use `FireFox`: less spying)
- **Advice:** learn how to touch-type NOW (reap the benefits earlier, not later)
  - you will be typing multiple hours/week in the next decades
  - touch-typing is about twice as fast as “system eagle”.
  - you can learn it in two weeks (good programs)

Outline of IWGS-II:

- Databases
  - CRUD operations, querying, and python embedding
  - XML and JSON for file based data storage

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

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

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

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)

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

IWGS-II Project

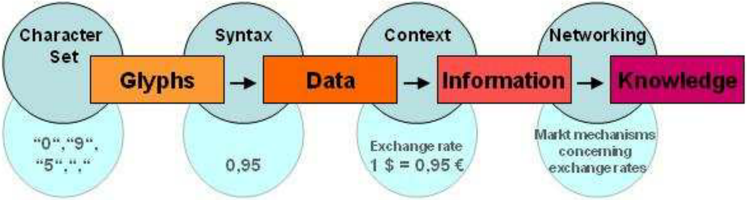
- **Idea:** Consolidate the techniques from IWGS-I and IWGS-II into a prototypical information system for Art History @ FAU. (Practical Digital Humanities)
- **A Running Example:** Research image + metadata collection “Bauernkirmes” provided by Prof. Peter Bell
- **What will you do?:** Build a web-based 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

9.1 Introduction

Databases, Data, Information, and Knowledge

- **Definition 1.1.** Discrete, objective facts or observations, which are unorganized and uninterpreted are called data (singular datum).
- According to Probst/Raub/Romhardt [PRR97]



- **Example 1.2.** The height of Mt. Everest (8.848 meters) is a datum.

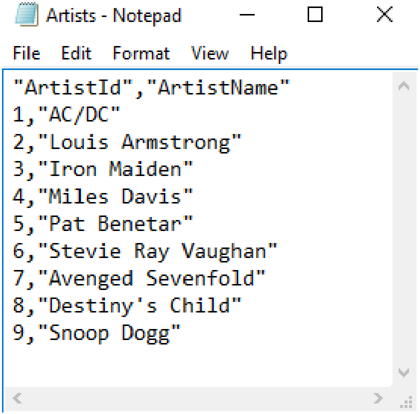
  Definition 1.3. A database is an organized collection of data, stored and accessed electronically from a computer system.
-

Storing Data Electronically

- Four conventional ways of storing data: (mileage varies)
  - In the computer's memory (RAM) (very fast (+), random access (+), but not persistent (-))

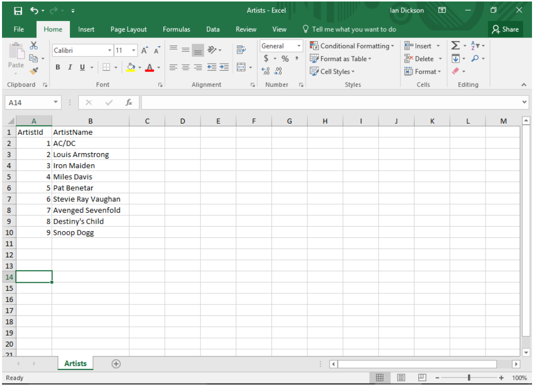
Storing Data Electronically

- 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 ( ))



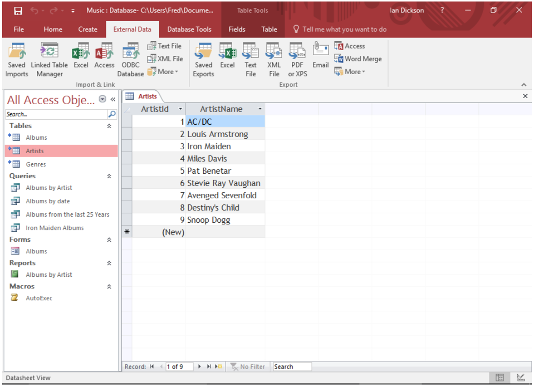
Storing Data Electronically

- 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 ( ))
  - In a spreadsheet (persistent (+), 2D-structured (+-), relations (+), slow (-))



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 ( ))
  - In a spreadsheet (persistent (+), 2D-structured (+-), relations (+), slow (-))
  - In a database (persistent (+), scalable (+), relations(+), managed (+), slow (-))



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 ( ))
  - In a spreadsheet (persistent (+), 2D-structured (+-), relations (+), slow (-))
  - In a database (persistent (+), scalable (+), relations(+), managed (+), slow (-))
- Databases constitute the most scalable, persistent solution.

9.2 Relational Databases

(Relational) Database Management Systems

- **Definition 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 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 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 Furry	1990	Random House	NY

- **Definition 2.4.** Tables are identified by table name and individual components of records by column name.

Open-Source Relational Database Management Systems

Definition 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 2.6. PostgreSQL is an open source RDBMS with an emphasis on extensibility, standards compliance, and scalability.

Definition 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.

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.

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
[...]
```

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".
  - If you have a database file books.db from 3.8, 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
```

Working with SQLite (via the SQLite shell)

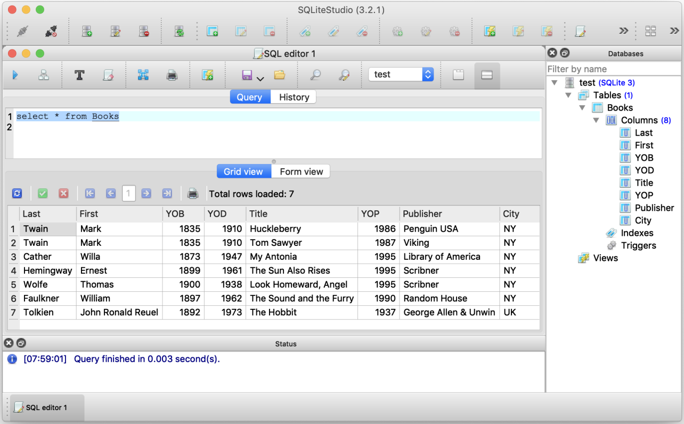
- In IWGS we will use SQLite, since it is very lightweight, easy to install, but feature complete, and widely used.
- Download SQLite at `https://www.sqlite.org/download.html`,
  - e.g. sqlite−dll−win64−x64−3280000.zip for windows.
  - unzip it into a suitable location, start sqlite3.exe there
  - this opens a command line interpreter: the SQLite shell. (all DBs have one)
    test it with .help that tells you about more "dot commands".
  - If you have a database file books.db from 3.8, use that.
  - .tables shows the available tables
    select * from Books is SQL (see below); it shows all entries of the Books table.

A Graphical User Interface for SQLite

- **Definition 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.

A Graphical User Interface for SQLite

- **Definition 2.9.** 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`



A Graphical User Interface for SQLite

- **Definition 2.10.** 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`
- 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

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 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.
-

DDL: Data Definition Language

- **Definition 3.2.** The data definition language (DDL) is a subset of SQL instructions that address the creation and deletion of database objects.
- **Definition 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 3.4 (Creating a Table).** The following SQL statement creates the table from 2.3

```
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 3.5.** The SQL statement **DROP** ⟪obj⟫ ⟪name⟫ deletes the database object of class ⟪obj⟫ with name ⟪name⟫.

SQL Data Types (for Column Specifications)

- **Definition 3.6.** SQL specifies data type for values including:
  - VARCHAR ($\langle\langle$length$\rangle\rangle$): character strings, including Unicode, of a variable length is up to the maximum length of $\langle\langle$length$\rangle\rangle$.
  - 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 ($\langle\langle$length$\rangle\rangle$) (character large object) up to (typically) 2GiB
  - BLOB ($\langle\langle$length$\rangle\rangle$) (binary large object) up to (typically) 2GiB

SQL: Adding Records to Tables

- **Definition 3.7.** SQL provides the **INSERT INTO** command for inserting records into a table. This comes in two forms:
  1. **INSERT INTO** $\langle\langle\text{table}\rangle\rangle$ **VALUES** ($\langle\langle\text{vals}\rangle\rangle$); where $\langle\langle\text{vals}\rangle\rangle$ is a comma-separated list of values given in the order the columns were declared in the **CREATE TABLE** instruction.
  2. **INSERT INTO** $\langle\langle\text{table}\rangle\rangle$ ($\langle\langle\text{cols}\rangle\rangle$) **VALUES** ($\langle\langle\text{vals}\rangle\rangle$) where $\langle\langle\text{vals}\rangle\rangle$ is a comma-separated list of values given in the order of $\langle\langle\text{cols}\rangle\rangle$ (a subset of columns) all other fields are filled with **NULL**

SQL: Adding Records to Tables

- **Definition 3.10.** 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 3.11 (Inserting into the Books Table).** The given the table Books from 3.4 we can add a record with

```
INSERT INTO Books
VALUES ('Tolkien', 'John␣Ronald␣Reuel', 1892, 1973, 'The␣Hobbit', 1937,
        'George␣Allen␣ Unwin', 'UK');
```

SQL: Adding Records to Tables

- **Definition 3.13.** 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 3.14 (Inserting into the Books Table).** The given the table Books from 3.4 we can add a record with

```
INSERT INTO Books
VALUES ('Tolkien', 'John␣Ronald␣Reuel', 1892, 1973, 'The␣Hobbit', 1937,
        'George␣Allen␣ Unwin', 'UK');
```

- **Example 3.15 (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');
```

SQL: Deleting Records from Tables

- **Definition 3.16.** The SQL delete statement allows to change existing records.

  DELETE FROM ⟪table⟫ **WHERE** ⟪condition⟫;

- **Example 3.17.** 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)

SQL: Updating Records in Tables

- **Definition 3.18.** The SQL update statement allows to change existing records.

  UPDATE $\langle\!\langle\text{table}\rangle\!\rangle$
  SET $\langle\!\langle\text{column}\rangle\!\rangle_1 = \langle\!\langle\text{value}\rangle\!\rangle_1$, $\langle\!\langle\text{column}\rangle\!\rangle_2 = \langle\!\langle\text{value}\rangle\!\rangle_2$, ...
  WHERE $\langle\!\langle\text{condition}\rangle\!\rangle$;

- **Example 3.19.** 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

Avoiding Redundancy in Databases

- Recall the books table from 2.3:

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 Furry	1990	Random House	NY

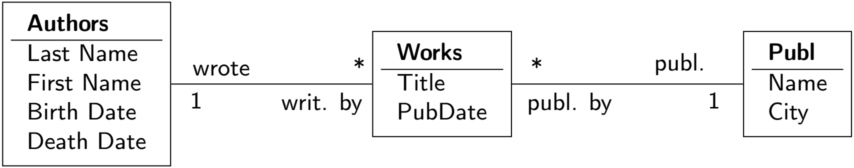
- **Observation:** Some of the fields appear multiple times, e.g. "Mark Twain".
- ⚠ When the database grows this can lead to scalability problems:
  - in querying: e.g. if we look for all works by Mark Twain
  - in maintenance: e.g. if we want to replace the pen name "Mark Twain" by the real name "Samuel 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

Entity Relationship Diagrams

- **Definition 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 4.2 (An ERD for Books).** Recall the Books table from 2.3:

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 Furry	1990	Random House	NY

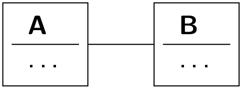
  - **Problem**: We have duplicate information in the authors and publishers
  - **Idea**: Spread the Books information over multiple tables.



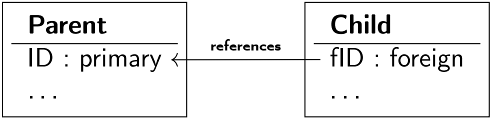
Linking Tables via Primary and Foreign Keys

- **Definition 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 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.

ERD



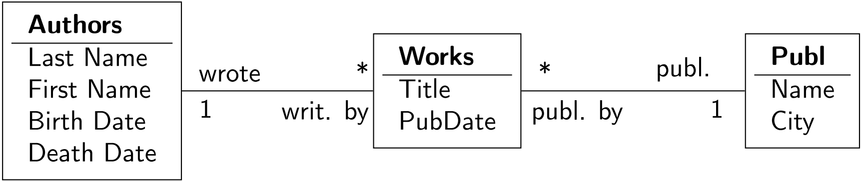
Implementation



- **BTW:** Primary keys are great for identification in the **WHERE** clauses of SQL instructions.

Linking Tables via Primary and Foreign Keys (Example)

- **Example 4.5.** Continuing 4.2, 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 4.6 (Inserting into the Works Table).** The given the tables Works Authors, and Publishers from 4.5 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);
```

9.5 RDBMS in Python

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(⟪host⟫,⟪user⟫,⟪pass⟫,⟪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.

Creating Tables in Python

▶ **Example 5.1.** Creating the table of 3.4

```
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
```

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)

9.6 Excursion: Programming with Exceptions in Python

How to deal with Errors in Python

- **Theorem 6.1 (Kohlhase's Law).**

How to deal with Errors in Python

- **Theorem 6.5 (Kohlhase's Law).** *I can be an idiot, and I do make mistakes!*
- **Corollary 6.6.** *Programming languages need a good way to deal with all kinds of errors!*

How to deal with Errors in Python

- **Theorem 6.9 (Kohlhase's Law).** *I can be an idiot, and I do make mistakes!*
- **Corollary 6.10.** *Programming languages need a good way to deal with all kinds of errors!*
- **Definition 6.11.** An exception is a special Python object. Raising an exception e terminates computation and passes e to the next higher level.

How to deal with Errors in Python

- **Theorem 6.13 (Kohlhase's Law).** *I can be an idiot, and I do make mistakes!*
- **Corollary 6.14.** *Programming languages need a good way to deal with all kinds of errors!*
- **Definition 6.15.** An exception is a special Python object. Raising an exception e terminates computation and passes e to the next higher level.
- **Example 6.16 (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
```

How to deal with Errors in Python

- **Theorem 6.17 (Kohlhase's Law).** *I can be an idiot, and I do make mistakes!*
- **Corollary 6.18.** *Programming languages need a good way to deal with all kinds of errors!*
- **Definition 6.19.** An exception is a special Python object. Raising an exception e terminates computation and passes e to the next higher level.
- **Example 6.20 (Division by Zero).** The Python interpreter reports unhandled exceptions.
- 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
```

How to deal with Errors in Python

- **Theorem 6.21 (Kohlhase's Law).** *I can be an idiot, and I do make mistakes!*
- **Corollary 6.22.** *Programming languages need a good way to deal with all kinds of errors!*
- **Definition 6.23.** An exception is a special Python object. Raising an exception e terminates computation and passes e to the next higher level.
- **Example 6.24 (Division by Zero).** The Python interpreter reports unhandled exceptions.
- 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)
  - can be raised when an abnormal condition appears

```
if denominator == 0 :
    raise DivByZero
else
    ⟨⟨computation⟩⟩
```

How to deal with Errors in Python

- **Theorem 6.25 (Kohlhase's Law).** *I can be an idiot, and I do make mistakes!*
- **Corollary 6.26.** *Programming languages need a good way to deal with all kinds of errors!*
- **Definition 6.27.** An exception is a special Python object. Raising an exception e terminates computation and passes e to the next higher level.
- **Example 6.28 (Division by Zero).** The Python interpreter reports unhandled exceptions.
- 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)
  - can be raised when an abnormal condition appears
  - can be handled in a **try**/**except** block (there can be multiple)

try:
    $\langle\langle$tentativecomputation$\rangle\rangle$
except : $\langle\langle$err$\rangle\rangle_1$, ..., $\langle\langle$err$\rangle\rangle_n$:
    $\langle\langle$errorhandling$\rangle\rangle$
finally :
    $\langle\langle$cleanup$\rangle\rangle$

Playing it Safe with Databases

- **Observation 6.29.** *Things can go wrong when connecting to a database!* (*e.g. missing file*)
- **Idea:** Raise exceptions and handle them.
- **Example 6.30.** 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

SQL Querying: The SELECT Statement

- SQL uses the **SELECT** instruction for retrieving data from a database.
- **SELECT** $\langle\langle$columns$\rangle\rangle$ **FROM** $\langle\langle$table$\rangle\rangle$ returns all records from $\langle\langle$table$\rangle\rangle$ restricted to the fields from $\langle\langle$columns$\rangle\rangle$.
- **Definition 7.1.** We call a **SELECT** instruction a query.

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 7.5.** We call a **SELECT** instruction a query.
- **Example 7.6.** **SELECT** Title, YOP **FROM** Books;
- **SELECT DISTINCT** removes duplicate values
- **SELECT** $*$ **FROM** ⟪table⟫ returns all records from ⟪table⟫.

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 7.9.** We call a **SELECT** instruction a query.
- **Example 7.10. SELECT** Title, YOP **FROM** Books;
- **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 7.11. SELECT** FirstN, LastN **FROM** Books **WHERE** YOP = 1995;

```
Willa|Cather
Ernest|Hemingway
Thomas|Wolfe
```

SQL Querying: The SELECT Statement

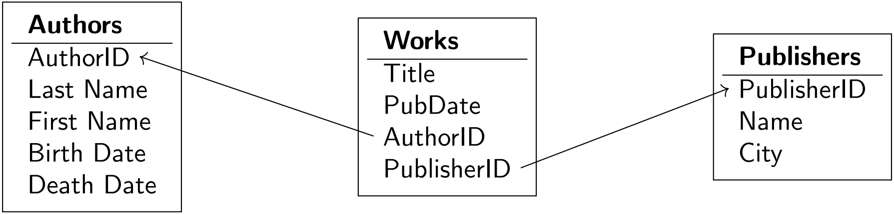
- SQL uses the **SELECT** instruction for retrieving data from a database.
- **SELECT** ⟪columns⟫ **FROM** ⟪table⟫ returns all records from ⟪table⟫ restricted to the fields from ⟪columns⟫.
- **Definition 7.13.** We call a **SELECT** instruction a query.
- **Example 7.14.** **SELECT** Title, YOP **FROM** Books;
- **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 7.15.** **SELECT** FirstN, LastN **FROM** Books **WHERE** YOP = 1995;
- **SELECT** ⟪columns⟫ **FROM** ⟪table⟫ **ORDER BY** ⟪colums⟫ orders the results by ⟪columns⟫

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 7.17.** We call a **SELECT** instruction a query.
- **Example 7.18.** **SELECT** Title, YOP **FROM** Books;
- **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 7.19.** **SELECT** FirstN, LastN **FROM** Books **WHERE** YOP = 1995;
- **SELECT** ⟪columns⟫ **FROM** ⟪table⟫ **ORDER BY** ⟪colums⟫ orders the results by ⟪columns⟫
- **Example 7.20.** Ordering can be ascending (**ASC**) or descending (**DESC**)
  SELECT FirstN, LastN **FROM** Books **ORDER BY** LastN **ASC**, YOP **DESC**;

Joining Tables in Queries

- **Problem:** We can query single tables, how cross-table queries? E.g. in

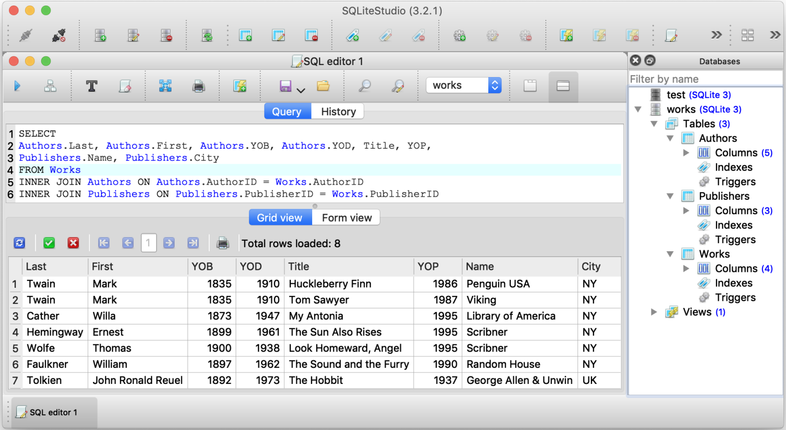


- **Idea:** Virtually join tables for the query! (as if we had the large books table)
- **Definition 7.21.** A table join (or simply join) is a means for combining columns from one (self join) or more tables by using values common to each.
- **Example 7.22.** Joining all three tables from 4.2.

```
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
```

Joining Tables in Queries (Result)

- **Example 7.23.**



Database Views: Persisting Queries

- **Observation:** Via the join in 7.22, the Works table queries like the original Books table.
- **Wouldn't it be nice** If we could also insert/update into that?
- **Definition 7.24.** 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 7.25.* 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)

Database Views: Persisting Queries (Books Example)

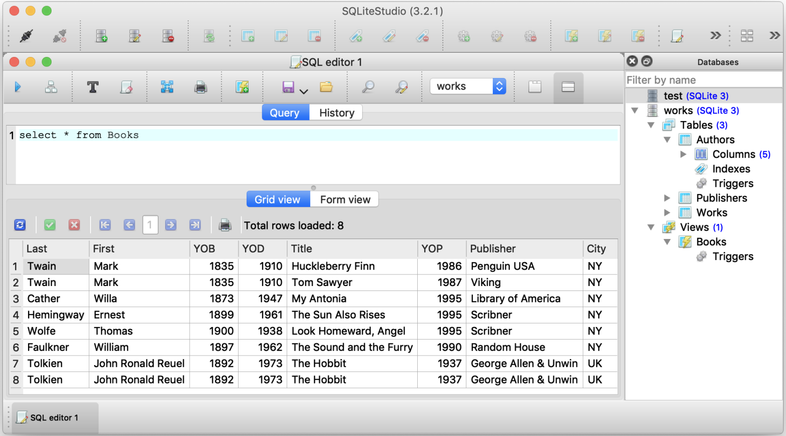
- **Example 7.26.** Use the query from 7.22 to define a view

```
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.

Database Views: Persisting Queries (Books Example)

- **Example 7.27.**



	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

Working with Cursors

- **Definition 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.

Extended Example: Listing Authors from the Books Table

- **Example 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!')
```

Inserting Multiple Records (Example)

- The cursor.executemany method takes an SQL instruction with parameters and a list of suitable tuples and executes them.
- **Example 8.3.** So the final form of insertion in 5.1 would be to define variable with a list of book tuples:

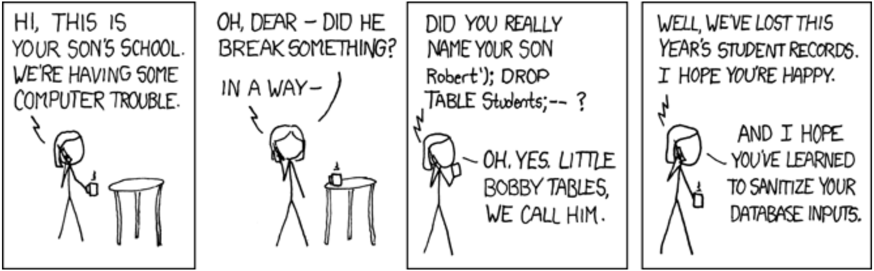
```
booklist = [
    ('Twain', 'Mark', 1835, 1910, 'Huckleberry␣Finn', 1986, 'Penguin␣USA', 'NY'),
    ('Twain', 'Mark', 1835, 1910, 'Tom␣Sawyer', 1987, 'Viking', 'NY'),
    ('Cather', 'Willa', 1873, 1947, 'My␣Antonia', 1995, 'Library␣of␣America', 'NY'),
    ('Hemingway', 'Ernest', 1899, 1961, 'The␣Sun␣Also␣Rises', 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␣', 'N
    ('Tolkien', 'John␣Ronald␣Reuel', 1892, 1973, 'The␣Hobbit', 1937,'George␣Allen␣ Unwin', 'UK
```

and then insert it via a call of cursor.executemany:

```
cursor.executemany('INSERT␣INTO␣Books␣VALUES␣(?,?,?,?,?,?,?,?)',booklist)
```

Beware of the Python/SQLite Interaction

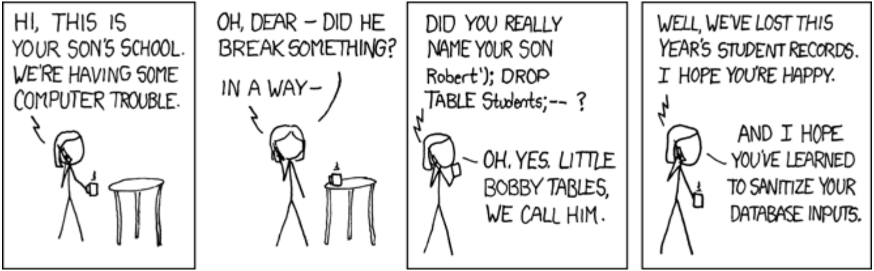
- **What have we learned?:** At least you now understand the following web comic: (https://xkcd.com/327/)



- **Definition 8.4.** We call this an SQL injection attack.

Beware of the Python/SQLite Interaction

- **What have we learned?:** At least you now understand the following web comic: (https://xkcd.com/327/)



- **Definition 8.5.** 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;
```

SQLite3 Parameter Substitution

- **Observation 8.6.** *We often need variables as parameters in* cursor.execute*.*
- **Example 8.7.** In 8.2 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 8.8.** 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

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 5.4, the XML file is online at
  `https://data.cityofnewyork.us/download/kcrm-j9hh/application/xml`

```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>
```

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 5.4, the XML file is online at `https://data.cityofnewyork.us/download/kcrm-j9hh/application/xml`
- **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.

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!

Filling a DB from via XML (Implementation)

- **Libraries:** urllib [UL] to retrieve the file and lxml [LXMLa] to parse it.
- 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 237.

Filling a DB from via XML (Implementation)

- **Libraries:** urllib [UL] to retrieve the file and lxml [LXMLa] to parse it.
- Collect all the XML tags in all the museums (for the column names)
- We create the SQLite database as discussed in slide 237.
- 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"
```

Filling a DB from via XML (Implementation)

- **Libraries:** urllib [UL] to retrieve the file and lxml [LXMLa] to parse it.
- Collect all the XML tags in all the museums (for the column names)
- We create the SQLite database as discussed in slide 237.
- Then we assemble a table specification in a string columns:
- 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});""")
```

Filling a DB from via XML (Implementation)

- **Libraries:** urllib [UL] to retrieve the file and lxml [LXMLa] to parse it.
- Collect all the XML tags in all the museums (for the column names)
- We create the SQLite database as discussed in slide 237.
- Then we assemble a table specification in a string columns:
- Create the Museums table from the specification in 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))
```

Filling a DB from via XML (Implementation)

- **Libraries:** urllib [UL] to retrieve the file and lxml [LXMLa] to parse it.
- Collect all the XML tags in all the museums (for the column names)
- We create the SQLite database as discussed in slide 237.
- Then we assemble a table specification in a string columns:
- Create the Museums table from the specification in columns
- Now the most important part: We fill the database
- We finalize the transaction as discussed in slide 237.

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
```

JSON — JavaScript Object Notation

- **Definition 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.

JSON — JavaScript Object Notation

- **Definition 9.3.** 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.
  - $\leadsto$ 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.

JSON — JavaScript Object Notation

- **Definition 9.5.** 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.
  - $\leadsto$ 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 9.6.** mongoDB is the most popular NoSQL database system. (no SQL inside)

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].

JSON Output for the NYC Museums DB

- **Libraries:** json for JSON [JS] and sqlite3 for the database.

```
import json
import sqlite3
```

JSON Output for the NYC Museums DB

- **Libraries:** json for JSON [JS] and sqlite3 for the database.
- 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;")
```

JSON Output for the NYC Museums DB

- **Libraries:** json for JSON [JS] and sqlite3 for the database.
- Connect to the SQLite database as usual and query the database for everything
- Initialize a dictionary and the list of Museums column names

```
data = {}
data['museums'] = []
columns = ['name', 'phone', 'address', 'closing', 'rates', 'specials']
```

JSON Output for the NYC Museums DB

- **Libraries:** json for JSON [JS] and sqlite3 for the database.
- Connect to the SQLite database as usual and query the database for everything
- Initialize a dictionary and the list of Museums column names
- 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)
```

JSON Output for the NYC Museums DB

- **Libraries:** json for JSON [JS] and sqlite3 for the database.
- Connect to the SQLite database as usual and query the database for everything
- Initialize a dictionary and the list of Museums column names
- For each of the rows in the Museums table build a row dictionary
- 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.

JSON Output for the NYC Museums DB I

```
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 Output for the NYC Museums DB II

```
    json.dump(data, outfile)

# Close database
db.close()
```

JSON Example (NYC Museums)

- **Example 9.7.** The NYC museums data from 5.4 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"
    }
   }
```

Chapter 10
Project: A Web GUI for a Books Database

10.1 A Basic Web Application

Building a full Web Application with Database Backend

- **Observation 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/MiKoMH/IWGS/blob/master/source/booksapp/code/books-app.py`.

The Books Application: Setup

- We have already seen how to set up the database in slide 249.

```
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
```

The Books Application: Setup

- We have already seen how to set up the database in slide 249.

```
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()
```

The Books Application: Setup

- We have already seen how to set up the database in slide 249.

```
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 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 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 Furry', 1990, 'Random House ', 'N
    ('Tolkien', 'John Ronald Reuel', 1892, 1973, 'The Hobbit', 1937,'George Allen  Unwin', 'UK
```

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␣Furry', 1990, 'Random␣House␣', 'N
    ('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.

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 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

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 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)
```

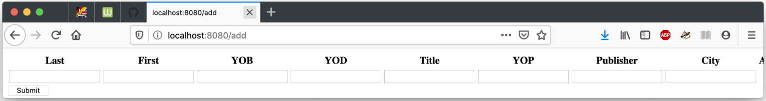
  Note that this is the route for the GET method on the path /add.

- 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 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 Books Application Routes: Adding Book Records

- This uses the function parseResponse, which we will reuse later.

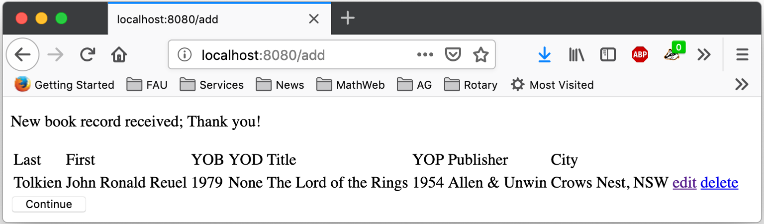
```
def parseResponse ():
    data = {'Last': request.forms.get('Last'),
            'First': request.forms.get('First'),
            'YOB': request.forms.get('YOB'),
            'YOD': request.forms.get('YOD'),
            'Title': request.forms.get('Title'),
            'YOP': request.forms.get('YOP'),
             'Publisher': request.forms.get('Publisher'),
            'City': request.forms.get('City')}
    return data
```

- and the template repsone.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 Books Application Routes: Adding Book Records

- Here is the result after filling in Tolkien's "*Lord of the Rings*":



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>
```

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)
```

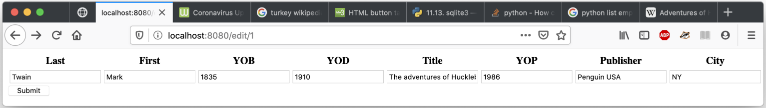
Books Application Routes: Editing Book Records (cont.)

- 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



- Again, we use the template response.tpl, which we fill with a different message.

10.2 Access Control and Management

Access Control and Management

- **Problem:** Anyone can write, edit, and delete records from the books database.
- **Solution:** Implement a password-based log in procedure and restrict write/edit/delete access to logged-in agents.
- Let's fix some terminology before we continue

Access Control and Management

- **Problem:** Anyone can write, edit, and delete records from the books database.
- **Solution:** Implement a password-based log in procedure and restrict write/edit/delete access to logged-in agents.
- Let's fix some terminology before we continue
- **Definition 2.3.** Access control is the selective restriction of access to a resource, access management describes the corresponding process.
- Access management usually comprises both authentication and authorization.
- **Definition 2.4.** Authorization refers to a set of rules that determine who is allowed to do what with a collection of resources.

Access Control and Management

- **Problem:** Anyone can write, edit, and delete records from the books database.
- **Solution:** Implement a password-based log in procedure and restrict write/edit/delete access to logged-in agents.
- Let's fix some terminology before we continue
- **Definition 2.5.** Access control is the selective restriction of access to a resource, access management describes the corresponding process.
- Access management usually comprises both authentication and authorization.
- **Definition 2.6.** Authorization refers to a set of rules that determine who is allowed to do what with a collection of resources.
- **For our books application** we need four things
  1. a browser interaction to query the user for username and password
  2. a way to transport them to the web application program
  3. a method for checking the username/password (authentication)
  4. a way the specify who can do what. (authorization)

  Realization: 1./2. via HTTP, 4. via bottle basic auth, implement 3. directly.

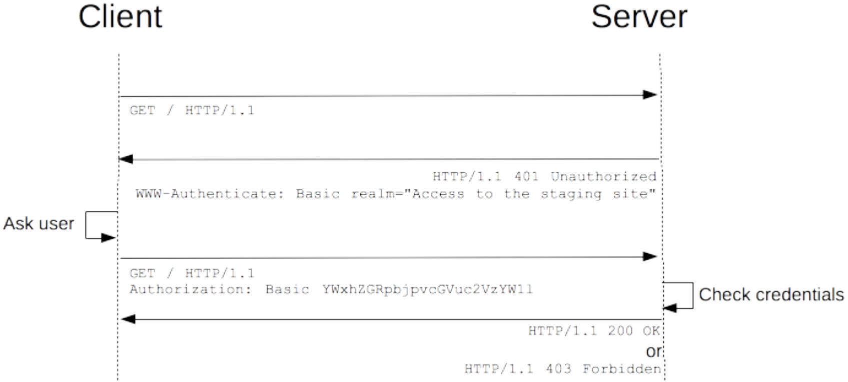
HTTP Basic Authentication

- **Recall** that HTTP is a plain text protocol that passes around headers like this

```
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)
```

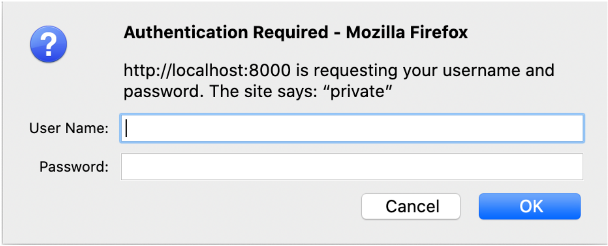
HTTP Basic Authentication

- **Recall** that HTTP is a plain text protocol that passes around headers like this
- **Idea:** For authentication extend the HTTP headers with support for username/password pairs.
- **Definition 2.8.** HTTP basic authentication introduces a HTTP header Authorization for base64 encoded pairs ⟪username⟫:⟪password⟫ and a couple of challenge/response messages.



HTTP Basic Authentication

- **Recall** that HTTP is a plain text protocol that passes around headers like this
- **Idea:** For authentication extend the HTTP headers with support for username/password pairs.
- **Definition 2.9.** HTTP basic authentication introduces a HTTP header Authorization for base64 encoded pairs $\langle\!\langle\mathrm{username}\rangle\!\rangle$:$\langle\!\langle\mathrm{password}\rangle\!\rangle$ and a couple of challenge/response messages.



HTTP Basic Authentication

- **Recall** that HTTP is a plain text protocol that passes around headers like this
- **Idea:** For authentication extend the HTTP headers with support for username/password pairs.
- **Definition 2.10.** HTTP basic authentication introduces a HTTP header Authorization for base64 encoded pairs $\langle\langle\text{username}\rangle\rangle$:$\langle\langle\text{password}\rangle\rangle$ and a couple of challenge/response messages.
- **Problem:** Base64 is very easy to decode, so usernames and passwords are communicated in the clear (very unsafe)
- Passwords are "binary data" (think special characters), encoding just keeps them unchanged over the network. (no encryption)

Basic Auth in Bottle

- **Idea:** Support the server side of HTTP basic authentication in bottle web-apps.
- **Implementation:** New decorator @auth_basic(⟪function⟫) to mark a route as password-protected.
- **Usage:** Decorate every route we want to restrict access of with @auth_basic(⟪function⟫), where ⟪function⟫ is a function that takes two string arguments (user name and password) and returns a Boolean for the authorization decision.

Basic Auth in Bottle: Minimal Viable Example

- **Example 2.11.** A web application with restricted route.

```
from bottle import run, get, auth_basic

def check(user, password):
    return user == "miko" and password == "test"

@get("/")
@auth_basic(check)
def protected():
    return "Authorized access granted!"

run(host="localhost", port=8000)
```

- **Idea:** Mix restricted and open routes in a partially restricted application.
- **Extension:** Use different check functions for different levels of restriction (user roles)

HTTPS: HTTP over TLS

- **Definition 2.12.** Hypertext Transfer Protocol Secure (HTTPS) is an extension of the Hypertext Transfer Protocol (HTTP) for secure communication over a computer network. HTTPS achieves this by running HTTP over a TLS connection.
- **Consequences for Web Applications:** We can use HTTP as usual, except
  - we gain communication privacy and server authentication,
  - server and browser need to speak HTTPS, (most do)
  - the server needs a public key certificate and a private key.

HTTPS: HTTP over TLS

- **Definition 2.13.** Hypertext Transfer Protocol Secure (HTTPS) is an extension of the Hypertext Transfer Protocol (HTTP) for secure communication over a computer network. HTTPS achieves this by running HTTP over a TLS connection.
- **Consequences for Web Applications:** We can use HTTP as usual, except
  - we gain communication privacy and server authentication,
  - server and browser need to speak HTTPS, (most do)
  - the server needs a public key certificate and a private key.
- In bottle, we can just swap out the HTTP server to one that can do HTTPS:

```
run(host='localhost',port='8888',
    server='gunicorn',keyfile='key.pem',certfile='cert.pem')
```

install it first with pip install gunicorn.

HTTPS: HTTP over TLS

- **Definition 2.14.** Hypertext Transfer Protocol Secure (HTTPS) is an extension of the Hypertext Transfer Protocol (HTTP) for secure communication over a computer network. HTTPS achieves this by running HTTP over a TLS connection.
- **Consequences for Web Applications:** We can use HTTP as usual, except
  - we gain communication privacy and server authentication,
  - server and browser need to speak HTTPS, (most do)
  - the server needs a public key certificate and a private key.
- In bottle, we can just swap out the HTTP server to one that can do HTTPS:

```
run(host='localhost',port='8888',
    server='gunicorn',keyfile='key.pem',certfile='cert.pem')
```

  install it first with pip install gunicorn.
- **Problem:** Where to get the certificate file cert.pem and private key key.pem?

HTTPS: HTTP over TLS

- **Definition 2.15.** Hypertext Transfer Protocol Secure (HTTPS) is an extension of the Hypertext Transfer Protocol (HTTP) for secure communication over a computer network. HTTPS achieves this by running HTTP over a TLS connection.
- **Consequences for Web Applications:** We can use HTTP as usual, except
  - we gain communication privacy and server authentication,
  - server and browser need to speak HTTPS, (most do)
  - the server needs a public key certificate and a private key.
- In bottle, we can just swap out the HTTP server to one that can do HTTPS:

```
run(host='localhost',port='8888',
    server='gunicorn',keyfile='key.pem',certfile='cert.pem')
```

  install it first with pip install gunicorn.
- **Problem:** Where to get the certificate file cert.pem and private key key.pem?
- **One Solution:** Self-sign one, e.g. using `https://www.selfsignedcertificate.com/` (adapt file names)
- **Remaining Problem:** Your browser forces you to specify an exception for `https://localhost:8888` (probably OK for development)

Getting a Real TLS Certificate via Let's-Encrypt

- **Intuition:** HTTPS is the new "regular HTTP" on the web!
- **Observation 2.16.** *A self-signed certificate gives communication privacy but not authentication ⇝ only you yourself vouch for the authenticity of the web site.*

Getting a Real TLS Certificate via Let's-Encrypt

- **Intuition:** HTTPS is the new "regular HTTP" on the web!
- **Observation 2.19.** *A self-signed certificate gives communication privacy but not authentication ⇜ only you yourself vouch for the authenticity of the web site.*
- **Definition 2.20.** In a public key infrastructure, the TLS certificate is issued by a certificate authority, an organization chartered to verify identity and issue TLS certificates.
- Commercial certificate authorities sell trust. (for a lot of money)
  They certify e.g. that the `https://bmw.com` is under control of BMW AG.

Getting a Real TLS Certificate via Let's-Encrypt

- **Intuition:** HTTPS is the new "regular HTTP" on the web!
- **Observation 2.22.** *A self-signed certificate gives communication privacy but not authentication* ⇜ *only you yourself vouch for the authenticity of the web site.*
- **Definition 2.23.** In a public key infrastructure, the TLS certificate is issued by a certificate authority, an organization chartered to verify identity and issue TLS certificates.
- Commercial certificate authorities sell trust. (for a lot of money)
  They certify e.g. that the `https://bmw.com` is under control of BMW AG.
- **Idea:** Finding out that you have control over a particular web site on the web can be automated, if you run a program on the server host.
- **Definition 2.24.** Let's Encrypt is a not for profit certificate authority that does this and issues free TLS certificates. (to encourage HTTPS adoption)

Getting a Real TLS Certificate via Let's-Encrypt

- **Intuition:** HTTPS is the new "regular HTTP" on the web!
- **Observation 2.25.** *A self-signed certificate gives communication privacy but not authentication ⇝ only you yourself vouch for the authenticity of the web site.*
- **Definition 2.26.** In a public key infrastructure, the TLS certificate is issued by a certificate authority, an organization chartered to verify identity and issue TLS certificates.
- Commercial certificate authorities sell trust. (for a lot of money)
  They certify e.g. that the `https://bmw.com` is under control of BMW AG.
- **Idea:** Finding out that you have control over a particular web site on the web can be automated, if you run a program on the server host.
- **Definition 2.27.** Let's Encrypt is a not for profit certificate authority that does this and issues free TLS certificates. (to encourage HTTPS adoption)
- **Concretely:** on a linux server you need two steps
  1. install certbot (usually via your package manager)
  2. then sudo /usr/local/bin/certbot certonly ——standalone will generate certs.

  Details at `https://letsencrypt.org`.
- **Success:** $\geq$ 1.000.000.000 TLS certificates, 200.000.000 sites since 2016

10.3 Asynchronous Loading in Modern Web Apps

AJAX for more responsive Web Pages

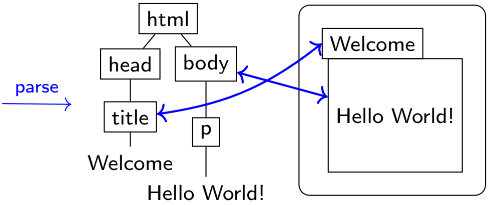
- **Definition 3.1.** Ajax, (also AJAX; short for "Asynchronous JavaScript and XML") is a set of client side techniques for creating asynchronous web applications.
- **Definition 3.2.** A process p is called asynchronous, iff the parent process (i.e. the one that spawned p) continues processing without 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.

Background: Rendering Pipeline in browsers

- **Observation:** The nested markup codes turn HTML documents into trees.
- **Definition 3.3.** 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 Document **DOM** **Browser**

```
<html>
  <head>
    <title>Welcome</title>
  </head>
  <body>
    <p>Hello World!</p>
  </body>
</html>
```



The DOM is notified of any user events (resizing, clicks, hover,. . .)

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.

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)

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)

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.
```

- 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)");
}
```

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.
- 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 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.
- 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.
- 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)");
  }
}
```

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:

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"><
       </td>
     </tr>
     <tr>
        <th>Publisher:</th>
        <td><span data−content="Publisher"></span>, <span data−content="YOP"
    </tr>
  </table>
</script>
```

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)
  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'}
```

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)
  2. a JavaScript data object: here the argument of the success function: the JSON record provided by the server under route /json/*i*
- 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>
```

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)
  2. a JavaScript data object: here the argument of the success function: the JSON record provided by the server under route /json/*i*
- The JQuery template processing places the value of the data−content attribute into the <span>. The resulting table constitutes the generated "detail view":
- **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.

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)
```

10.4 Deploying the Books Application as a Program

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.
- **Example 4.1.** 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 4.2.** 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
```

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 4.3 (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

11.1 Basics of Image Processing

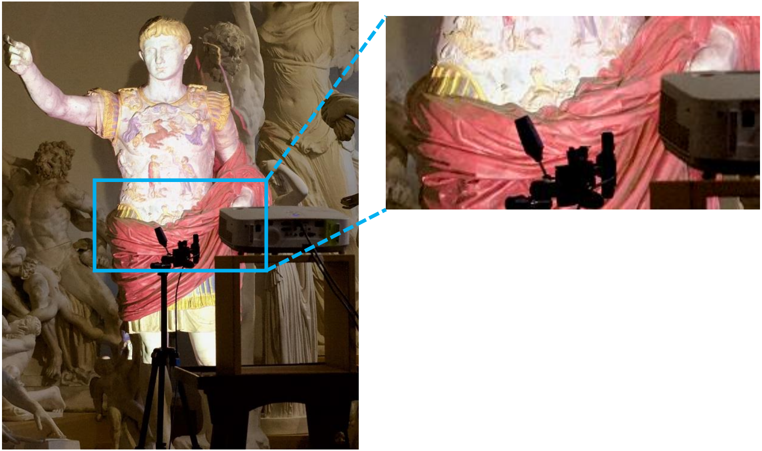
11.1.1 Image Representations

Images

- **Example 1.1 (Zooming in on Augustus).** A digital image taken by a standard DSLR camera. Let's zoom in on it!

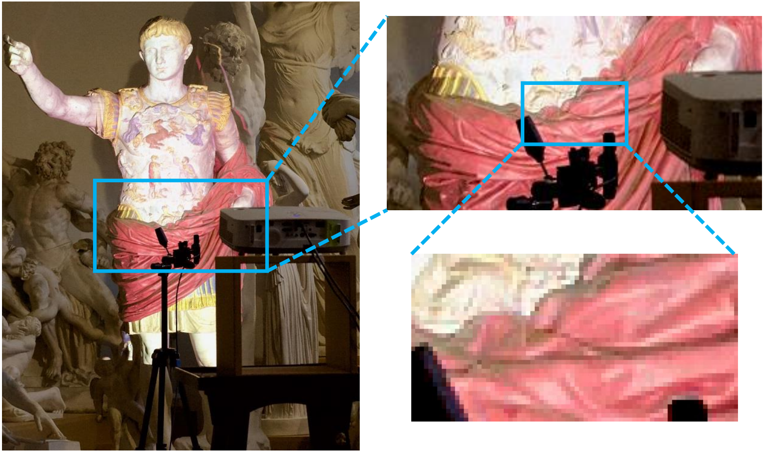
Images

- **Example 1.2 (Zooming in on Augustus).** And a bit more



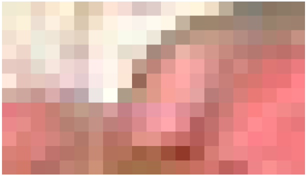
Images

- **Example 1.3 (Zooming in on Augustus).** When zooming in on an image, we start to see blocks of colors, which are organized in a regular grid.



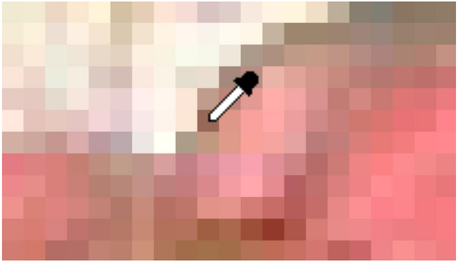
Images as Rasters of Pixels

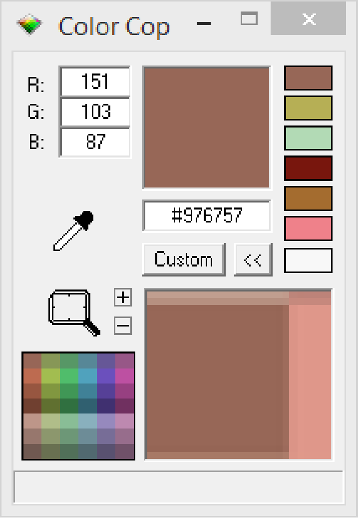
- If we zoom in quite a bit more, we see
- **Observation:** The colors are arranged in a two- dimensional grid (raster).



- **Definition 1.4.** We call the grid raster and each entry in it pixel (from "picture element").

Colors

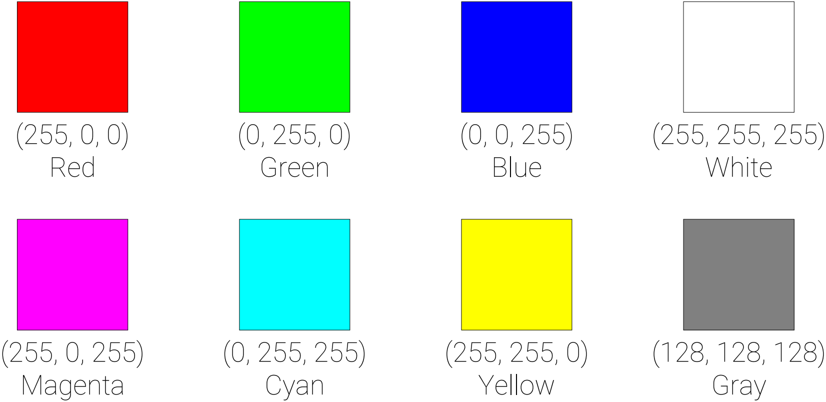




- **Definition 1.5.** Colors are usually represented in RGB format, i.e. as triples $\langle R, G, B\rangle$ with three channels (also called bands).
- $R, G, B \in [0{,}255]$ $\rightsquigarrow$ One Byte per channel per pixel.
- Images in this format can store $256 \cdot 256 \cdot 256 = 256^3$ (about 16 million) colors.

Color Examples

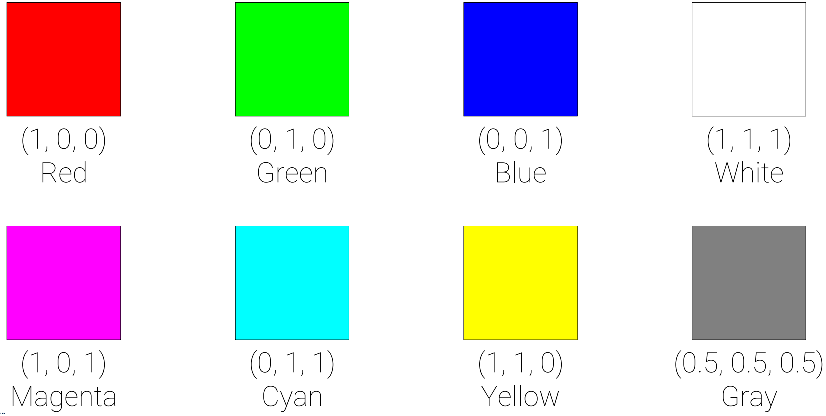
- **Example 1.6.** A color can be represented by three numbers.



- **Definition 1.7.** A color is called grayscale, iff $R = G = B$

Normalized Color Values

- **Observation 1.8.** *For color representations, only the relative contribution of the band is imporant.*
- **Definition 1.9.** Normalized colors use pixel values between 0 and 1.
- **Idea:** Values are still stored as Bytes, but normalized before use: $v' = v/255$
- **Example 1.10.**



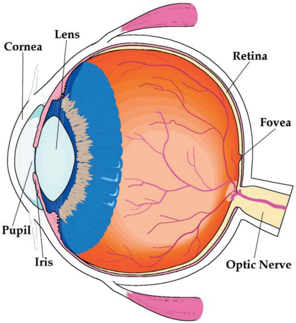
HTML Color Codes

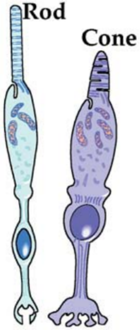
- HTML uses a shorthand notation for colors using hexadecimal numbers.
- **Example 1.11.**

#FF0000 Red	#00FF00 Green	#0000FF Blue	#FFFFFF White
#FF00FF Magenta	#00FFFF Cyan	#FFFF00 Yellow	#808080 Gray

The Human Eye

- **Definition 1.12 (The Human Eye).** Light from our surroundings enters our eye through the lens and then hits the retina on the back of our eye.



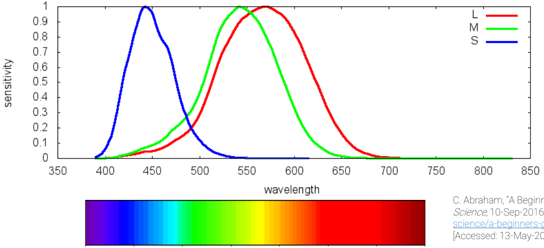


  The retina has cones and rods, which are responsible for color and brightness vision, respectively.

- Since we are interested in colors here, we will ignore the rods for the purpose of this lecture.

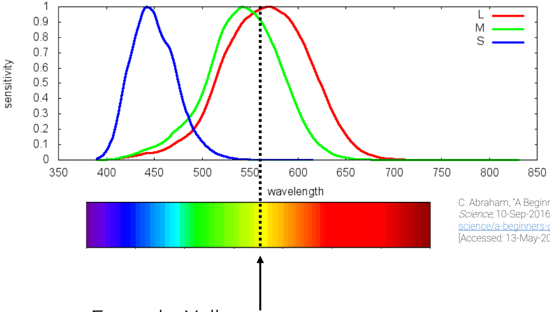
The Human Eye – Three Types of Cones

- **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].

- **Example 1.13 (We see Yellow).**



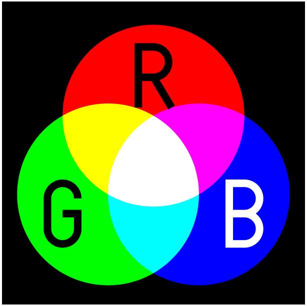
C. Abraham, "A Beginner's Guide to (CIE) Colorimetry," *Hipster Color Science*, 10-Sep-2016. Available: https://medium.com/hipster-color-science/a-beginners-guide-to-colorimetry-401f1830b65a. [Accessed: 13-May-2019].

Example: Yellow
Both "red" and "green" cone are stimulated.

- **Observation 1.14.** *We can create all (human-visible) colors as a mixture of red, green, and blue light.*

Monitors

- **Definition 1.15.** A computer monitor (or just monitor)is an output device for visual information.
- Monitors (usually) have pixels, too!
- **Definition 1.16.** 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.



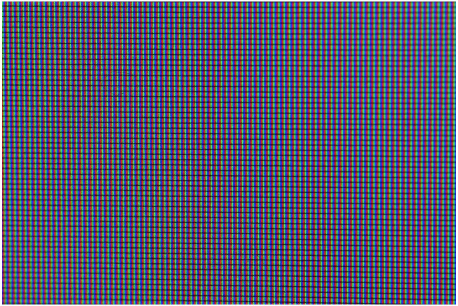


Image Size

- **Example 1.17 (Augustus again).**

  Image: 1440×746 pixels
  Expected file size:
  $\mathrm{Width} \cdot \mathrm{Height} \cdot \mathit{Channels}$
  $1440 \cdot 746 \cdot 3 = 3,222,720\mathrm{B} \approxeq 3\mathrm{MiB}$

- 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.

JPEG Compression Artefacts

- **Example 1.18 (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

11.1.2 Basic Image Processing in Python

The Pillow Library for Image Processing in Python

- We will use the Pillow library in IWGS.
- **Definition 1.19.** 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 1.20.** Determine the color of a particular pixel

```python
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))
```

The Pillow Library for Image Processing in Python

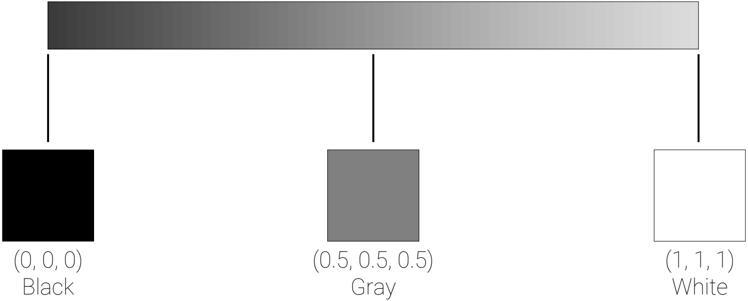
- We will use the Pillow library in IWGS.
- **Definition 1.22.** Pillow is a fork (a version) of the old Python library PIL (Python Image Library). (hence the name)
- Details at `https://pillow.readthedocs.io/slides/stable/`
- **Install:** pip install Pillow
- **Example 1.24.** Directly use the image 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.

Grayscale Images

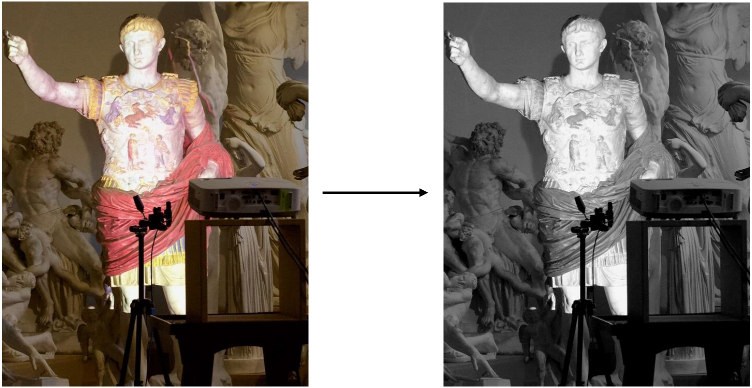
- **Recall:** A color is grayscale, iff R=G=B.



- **Idea:** If all channels have the same value, why store all three?
- Grayscale images usually have only one channel.

Grayscale Conversion

- **Observation 1.25.** *Humans are very sensitive to green, less to red, and least to blue.*
- **Definition 1.26.** To convert an image to an grayscale image (grayscale conversion), we compute $Gray = 0.21R + 0.71G + 0.08B$
- **Example 1.27 (Grayscale Conversion).**



More Image Operations

- **Example 1.28 (More Image Operations).**

Original

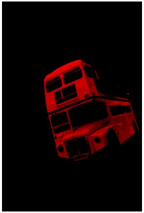
Grayscale

Sepia

Inverse

Each pixel is
processed separately!

Threshold



Red Channel
Extraction

- As for grayscale conversion of these process each pixel separately.

Image Operations in Pillow

- The pillow library supports many image operations out of the box.
- **Example 1.29 (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)
```

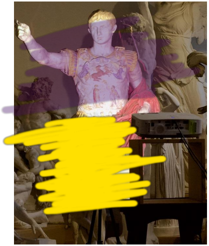
- Complete List:
  `https://pillow.readthedocs.io/en/stable/reference/ImageOps.html`

Transparency and Image Composition

- Sometimes we want to overlay images $\rightsquigarrow$ layers.
- We need a notion of how transparent a pixel is.
- **Definition 1.30.** 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 1.31 (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!

Transparency (continued)

▶ **Example 1.32 (Combining Images).**

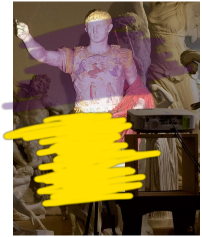
(R,G,B,A) = (0.6, 0.0, 1.0, 0.5)
Half transparent purple

(R,G,B,A) = (0, 0, 0, 0)
Full transparent

(R,G,B,A) = (1, 1, 0, 1)
Full yellow

+

=



$$\mathbf{R}_{\text{target}} = (1\text{-}A) \times \mathbf{R}_{\text{augustus}} + A \times \mathbf{R}_{\text{purple,yellow}}$$

$$\mathbf{G}_{\text{target}} = (1\text{-}A) \times \mathbf{G}_{\text{augustus}} + A \times \mathbf{G}_{\text{purple,yellow}}$$

$$\mathbf{B}_{\text{target}} = (1\text{-}A) \times \mathbf{B}_{\text{augustus}} + A \times \mathbf{B}_{\text{purple,yellow}}$$

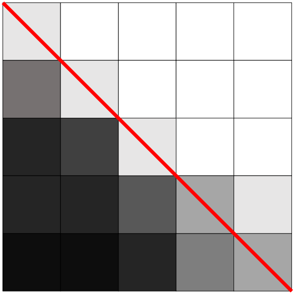
11.1.3 Edge Detection

Edge Detection

- **Goal:** Find interesting parts of image (features).

Edge Detection

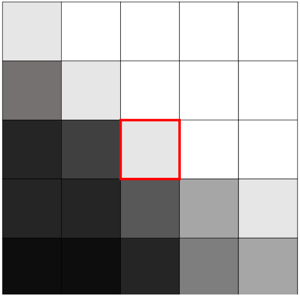
- **Goal:** Find interesting parts of image (features).
- **Example 1.35 (Edge Detection).** Find edges, i.e. image sections, where color changes rapidly.



Clearly there is an edge in this image. How do we detect it automatically?

Edge Detection

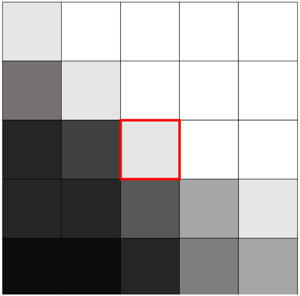
- **Goal:** Find interesting parts of image (features).
- **Example 1.37 (Edge Detection).** Find edges, i.e. image sections, where color changes rapidly.



Decide for each pixel, whether it is on an edge. Here: Is marked pixel an edge pixel?

Edge Detection

- **Goal:** Find interesting parts of image (features).
- **Example 1.39 (Edge Detection).** Find edges, i.e. image sections, where color changes rapidly.



Inspect neighbor pixels.

Edge Detection

- **Goal:** Find interesting parts of image (features).
- **Example 1.41 (Edge Detection).** Find edges, i.e. image sections, where color changes rapidly.
- **Definition 1.42.** We call a pixel a horizontal edge pixel, iff

$$I_B - I_T + I_{BL} - I_{TL} + I_{BR} - I_{TR} > \tau$$

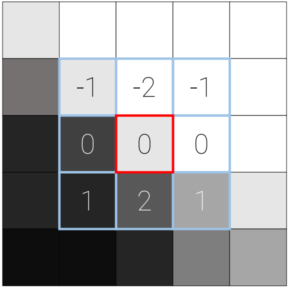
for some threshold τ and a vertical edge pixel, iff

$$I_R - I_L + I_{TR} - I_{TL} + I_{BR} - I_{BL} > \tau$$

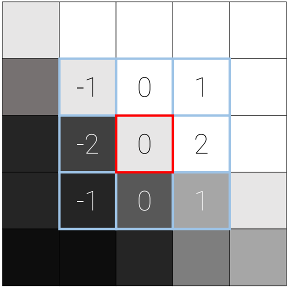
Algorithm: Sobel Filter

- **Idea:** There is a general algorithm that computes this.
- **Definition 1.43.** Given a 3×3 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 1.44.** *Given a suitable matrix M, the Sobel filter computes the quantities from 1.34.*
- **Example 1.45 (Edge Tests via Sobel Filters).**

Horizontal edge test:



Vertical edge test:

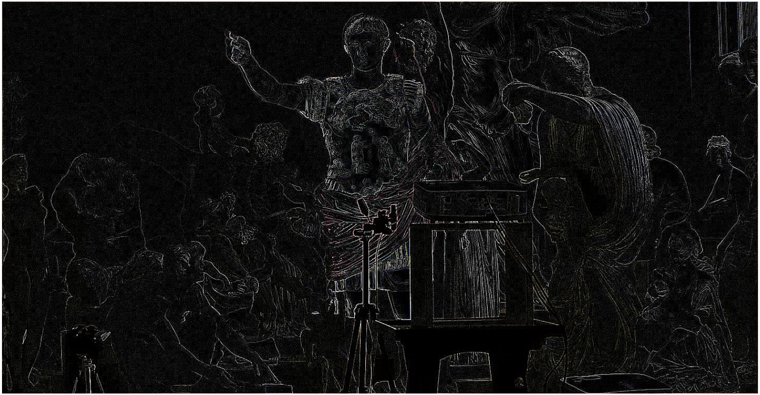


Edge-Detection in Pillow

- **Example 1.46 (Augustus and his Edges).**

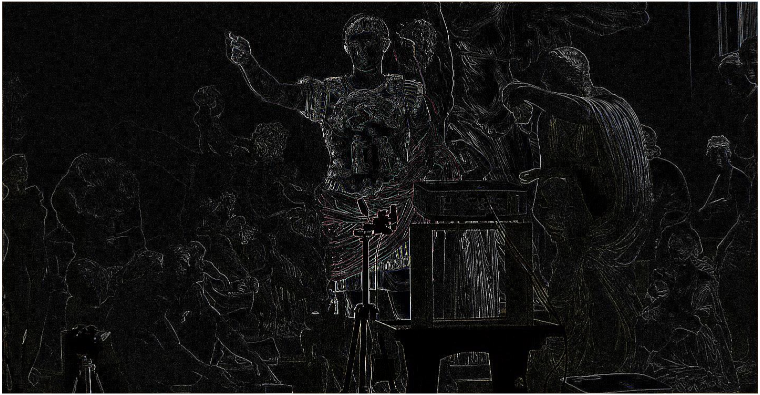
Edge-Detection in Pillow

- **Example 1.48 (Augustus and his Edges).**



Edge-Detection in Pillow

- **Example 1.50 (Augustus and his Edges).**



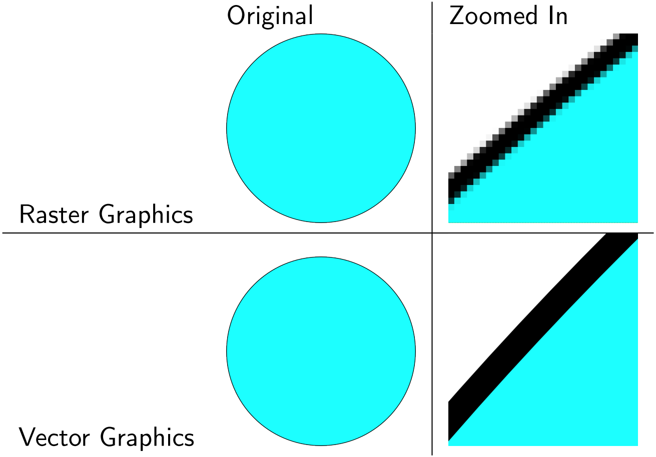
- **Example 1.51 (Edge Detection in Pillow).**

```
from PIL import Image, ImageFilter
im = Image.open('augustus.jpg')
edges = im.filter(ImageFilter.FIND_EDGES)
edges.show() # or just edges in Jupyter
```

11.1.4 Scalable Vector Graphics

Vector Graphics

- **Problem:** Raster images store colors in pixel grid. Quality deteriorates when image is zoomed into.
- Vector Graphics solve this problem!



Vector Graphics (Definition)

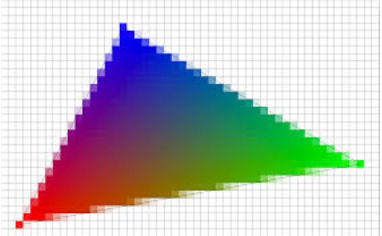
- **Definition 1.52.** Image representation formats that store shape information instead of individual pixels, are refered to as vector graphics.
- **Example 1.53.** For a circle, just store
  - center
  - radius
  - line width
  - line color
  - fill color
- **Example 1.54.** For a line, store
  - start and end point
  - line width
  - line color

Vector Graphics Display

- There are devices that directly display vector graphics.
- **Example 1.55.**

Vector Graphics Display

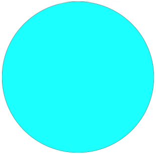
- There are devices that directly display vector graphics.
- **Example 1.58.**
- **Definition 1.59.** For monitors, vector graphics must be rasterized – i.e. converted into a raster image before display.
- **Example 1.60.**



Scalable Vector Graphics (SVG)

- **Definition 1.61.** Scalable Vector Graphics (SVG) is an XML-based markup format for vector graphics.
- **Example 1.62.**

```
<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 1.63 (Rectangle).**

```
<rect x="..." y="..." width="..." height="..." style="..." />
```

- **Example 1.64 (Ellipse).**

```
<ellipse cx="..." cy="..." rx="..." ry="..." style="..." />
```

- **Example 1.65 (Line).**

```
<line x1="..." y1="..." x2="..." y2="..." style="..." />
```

- **Example 1.66 (Text).**

```
<text x="..." y="..." style="...">This is my text!</text>
```

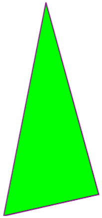
- **Example 1.67 (Image).**

```
<image xlink:href="..." x="..." y="..." width="..." height="..." />
```

SVG Polygons

- **Example 1.68 (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>
```



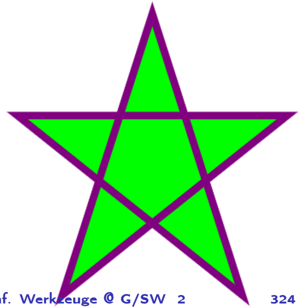
SVG Polygons

- **Example 1.70 (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 1.71 (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 in HTML

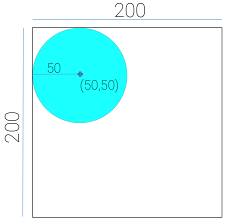
- SVG can be used in dedicated files (file ending .svg) and referenced in a <**img**> tag.
- It can however also be written directly in HTML files.
- **Example 1.72.** Triangle from 1.68 embedded in HTML file

```
<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>
```

The SVG viewBox Attribute

- **Idea:** The SVG viewBox attribute allows us to zoom into an image.
- **Example 1.73.**

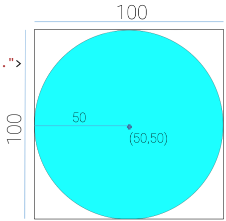
```
<svg width="200" height="200" xmlns="...">
  <circle cx="50" cy="50" r="50" style="..." />
</svg>
```



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 1.74.**

```
<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).

11.2 Project: An Image Annotation Tool

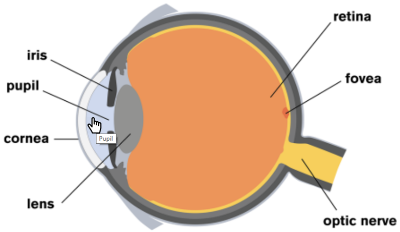
Project: Kirmes Image Annotation Tool

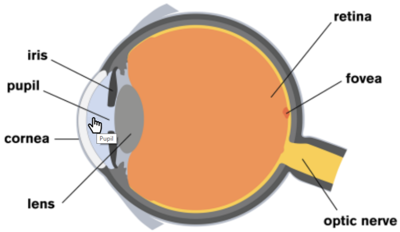
- **Problem:** Our Books-App project was a fully functional web application, but does not do anything useful for DigiHumS.
- **Idea:** Extend/Adapt it to a database for image annotation like LabelMe [LM].
- **Setting:** Prof. Peter Bell (formerly at FAU) conducts research on baroque paintings on parish 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

HTML Image Maps

- **Definition 2.1.** HTML image maps mark areas in an digital image and assign names and links to them.
- **Example 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 Image Maps

- **Definition 2.3.** HTML image maps mark areas in an digital image and assign names and links to them.
  - **Example 2.4.** An image map adds hover and on click behavior

```
<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/`

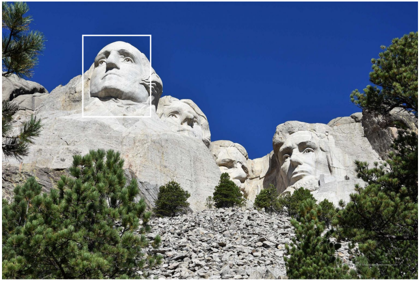
Problems of HTML Image Maps

- **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.

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 2.5 (Paper Prototype).** Highlight regions and display information on hover.



George Washington

Abraham Lincoln

SVG Annotation Implementation Areas

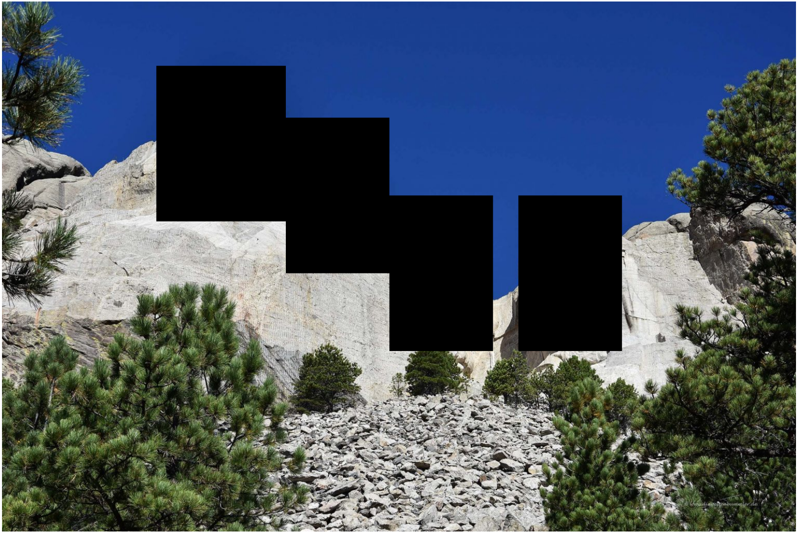
- **Implementing Areas as Rectangles:**

```
<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).

SVG Annotation Implementation Result

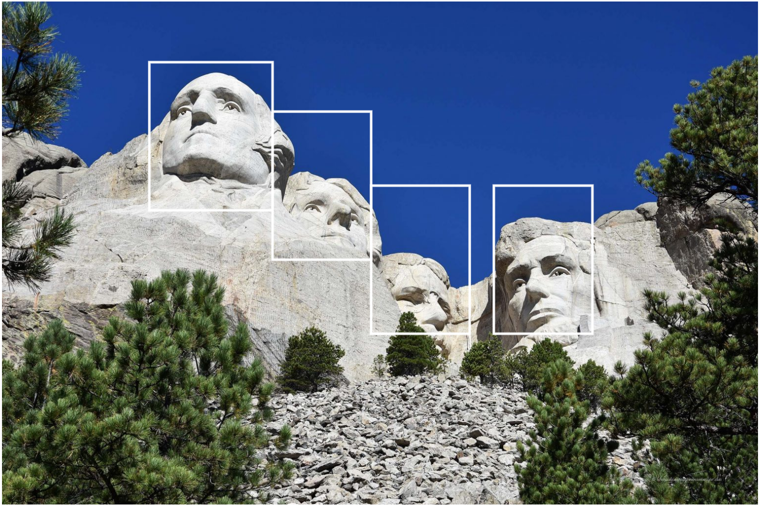
- **Areas as Rectangles – Result:** Now the rectangles are visible



Adding CSS for the Areas

- **Example 2.6 (Adding CSS).**

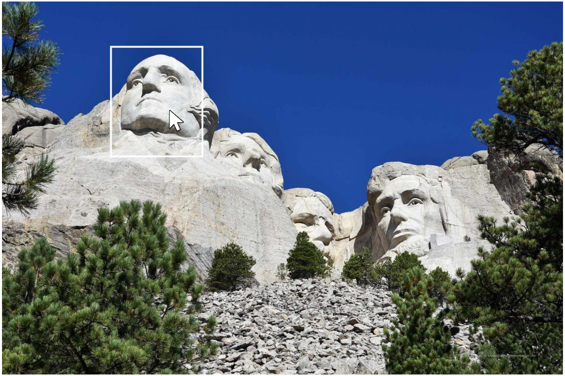
```
rect {fill−opacity:0; stroke:white; stroke−opacity:1; stroke−width:5px}
```



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}
```



Adding Annotation Text

- **Adding Annotation Text** and making space for it.

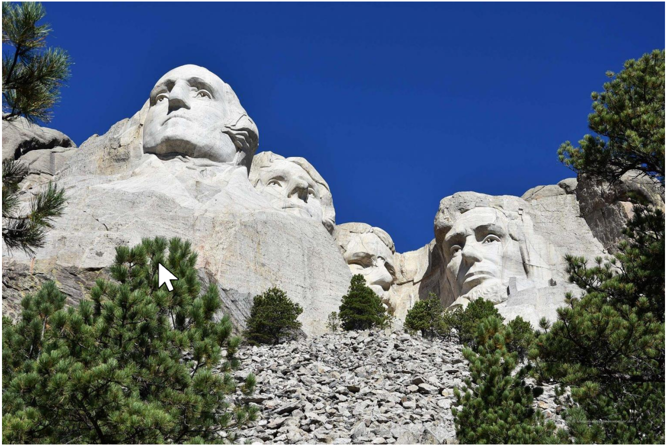
```
<svg xmlns="http://www.w3.org/2000/svg" width="1536" height="1224" >
  <!-- Image -->
  <image width="1536" height="1024" xlink:href="mount_rushmore.jpg" />
  <!-- Areas in image as rects, text below -->
  <rect x="300" y="125" width="250" height="300" />
  <text x="100" y="1200">George Washington</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}
```

Adding Annotation Text – Result

- **Adding Annotation Text – Result:**



Selectively Showing Annotations

- **Problem:** Now the annotations are always visible.
- **Idea:** Add CSS hover effect for <rect>s, which effects the |<text>|.
- **Definition 2.7.** The CSS sibling operator + modifies a selector so that it (only) affects following sibling elements (same level).
- **Example 2.8.** In the CSS directive

```
rect:hover + text {<rules>}
```

Selector   Sibling operator   Target

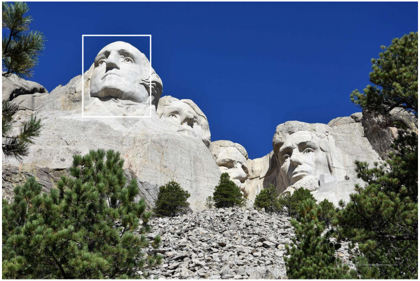
the rules affect the SVG <text> directly after the <rect> element.

- **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}
```

Image Annotation Tool – Final Result

- Now our annotation tool works as expected!
- **Example 2.9 (Final Result).** Highlight regions and display information on hover.



George Washington

Abraham Lincoln

11.3 Fun with Image Operations: CSS Filters

CSS Image Filters

- **Goal:** Apply image filters (grayscale etc.) directly in CSS.
- **Example 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!

Some more CSS Filters

- **Example 3.2 (Image Effects via CSS Style sheets).**

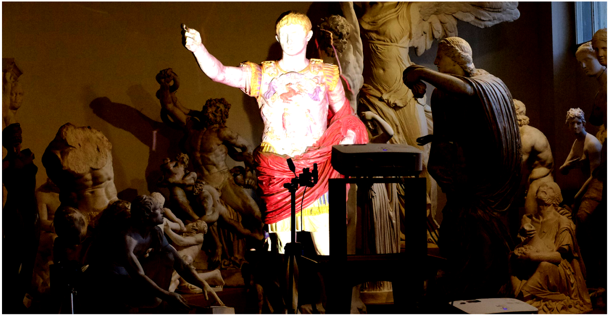
  ```
  <img style="filter: blur(4px" src="augustus.jpg" alt="no image"/>
  ```

Some more CSS Filters

- **Example 3.3 (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"/>
```



Some more CSS Filters

- **Example 3.4 (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"/>
```

Combining CSS Filters

- **Idea:** We can also combine image filters flexibly. The easist way is when we define CSS classes for that.
- **Example 3.5 (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>
```

Filtering Everyghing Else

- **Note:** CSS filters don't just apply to images! (Almost) everything can be filtered.
- **Example 3.6 (Filtering Text (Blurring)).**

```
<p style="filter:␣blur(3px)">A severely blurred Text</p>
```

CSS Animations

- **Definition 3.7.** CSS animations change state of an object over time.
- **Example 3.8 (Inverting an image).**

```
img {animation: invertAnimation 1s forwards}

@keyframes invertAnimation {
    from {filter: none}
    to {filter: invert(100%)}
  }
```

SVG Filters

- **Note:** Unfortunately in SVG the filtering works differently from CSS.
- **Example 3.9 (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 3.10 (SVG Filters can be combined).**

```
  <filter id="myCustomFilter">
    <feGaussianBlur stdDeviation="5" />
    <feColorMatrix type="saturate" values="0.1" />
  </filter>
```

Chapter 12
Ontologies, Semantic Web for Cultural Heritage

12.1 Documenting our Cultural Heritage

Documenting our Cultural Heritage

- **Definition 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?

Research Data in a Nutshell

- **Definition 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.

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)

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 1.4 (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].

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 1.5 (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?

Categories of Data in DigiHumS and their Formats

- We distinguish four broad categories of data in DigiHumS.
- **Definition 1.6.** 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)

Categories of Data in DigiHumS and their Formats

- We distinguish four broad categories of data in DigiHumS.
- **Definition 1.12.** 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 1.13.** Narrative data: documents and text fragments used for communicating knowledge to humans.
  - e.g. plain text and formatted text with markup code (see)

Categories of Data in DigiHumS and their Formats

- We distinguish four broad categories of data in DigiHumS.
- **Definition 1.18.** 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 1.19.** Narrative data: documents and text fragments used for communicating knowledge to humans.
  - e.g. plain text and formatted text with markup code (see)
- **Definition 1.20.** Symbolic data: descriptions of object and facts in a formal language
  - e.g. 3+5 in Python (see)

Categories of Data in DigiHumS and their Formats

- We distinguish four broad categories of data in DigiHumS.
- **Definition 1.24.** 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 1.25.** Narrative data: documents and text fragments used for communicating knowledge to humans.
  - e.g. plain text and formatted text with markup code (see)
- **Definition 1.26.** Symbolic data: descriptions of object and facts in a formal language
  - e.g. 3+5 in Python (see)
- **Definition 1.27.** 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 1.28.** *Metadata are the resources, DigiHumS results are made of ($\leadsto$ support that)*
  The other categories digitize artefacts and auxiliary data.

Categories of Data in DigiHumS and their Formats

- We distinguish four broad categories of data in DigiHumS.
- **Definition 1.30.** 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 1.31.** Narrative data: documents and text fragments used for communicating knowledge to humans.
  - e.g. plain text and formatted text with markup code (see)
- **Definition 1.32.** Symbolic data: descriptions of object and facts in a formal language
  - e.g. 3+5 in Python (see)
- **Definition 1.33.** 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 1.34.** *Metadata are the resources, DigiHumS results are made of (⇝ support that)*
  The other categories digitize artefacts and auxiliary data.
- **Observation 1.35.** *We will need all of these – and their combinations – to do DigiHumS.*

WissKI: a Virtual Research Env. for Cultural Heritage

- **Definition 1.36.** 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

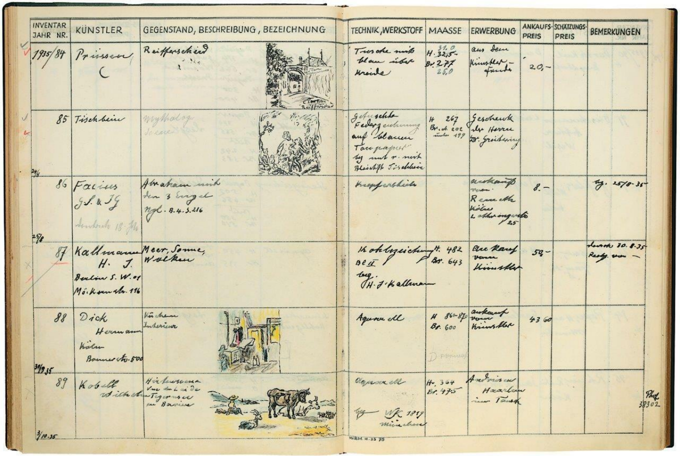
Documenting Cultural Heritage: Current State/Preview

- 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

Documenting Cultural Artefacts: Inventory Books

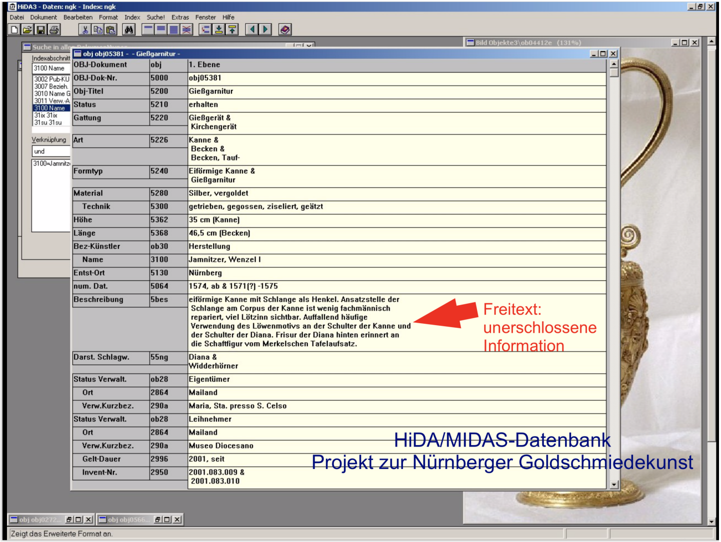
- **Definition 2.1.** An inventory book is a ledger that identifies, describes, and records provenance of the artefacts in the collection of a museum.
- **Example 2.2 (An Inventory Book).**



- **Problems:** non-digital, only single-user access, institution-local, no querying, …

Cultural Artefacts in Databases: Example

- **Example 2.3.** A typical database for cultural artefacts: (HiDa/MIDAS)

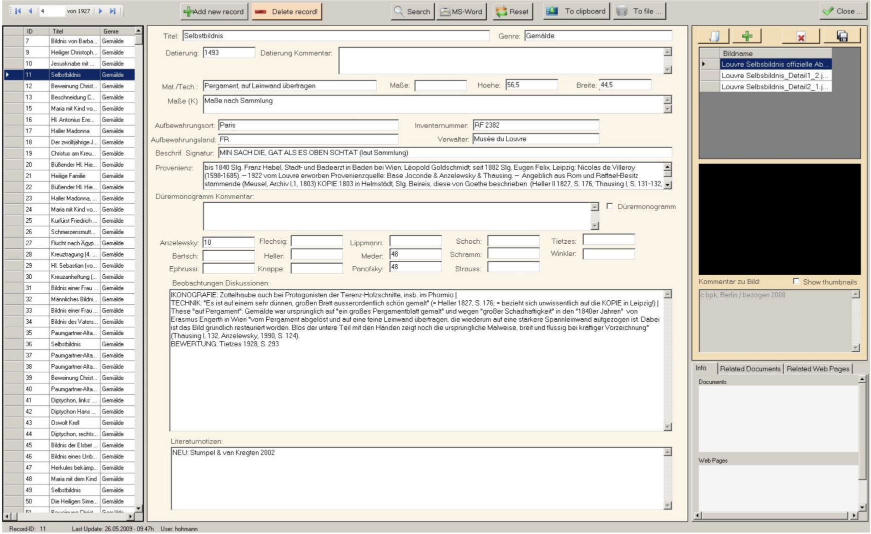


Cultural Artefacts in Databases: Pro/Con

- **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 $\leadsto$ no trans database relations,
  - database schemata are inflexible $\leftrightsquigarrow$ 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)

Cultural Artefacts in Databases II

- **Example 2.4.** Another database for cultural artefacts:



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)

Cultural Artefacts on the Web

- **Example 2.5.** A text about a cultural artefact (an etching by Dürer)

Not logged in Talk Contributions Create account Log in

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export

Article Talk Read Edit View history Search Wikipedia

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?

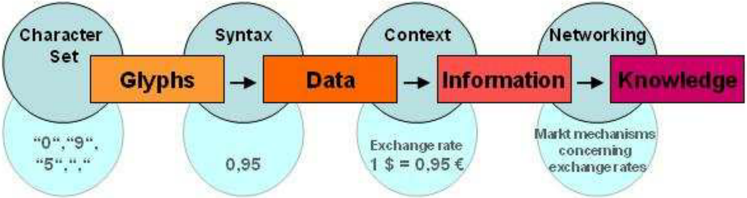
Using the Web for Cultural Heritage

- **Problems:** with using the Web as a resource
  - Information is often of dubious quality (imprecise, typos, incomplete, . . .)
  - Information is primarily written for human consumption
    - $\leadsto$ 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

The Semantic Web

- **Definition 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 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 $\leadsto$ semantic web technology.

What is the Information a User sees?

- **Example 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.

What the machine sees

- **Example 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∇⟩⌈⇔⊔⟨]\]§⊔}]\]∇⊣⊔⟩≀\⟩\⊔]∇\]⊔↙

Solution: XML markup with “meaningful” Tags

- **Example 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>

What can we do with this?

- **Example 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!

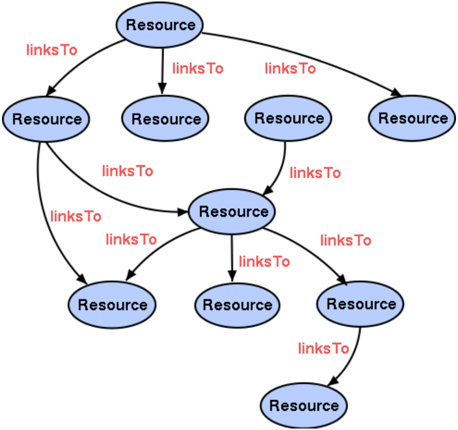
What the machine sees of the XML

- **Example 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∇⟩[⇔⊔⟨]\]§⊔}]\]∇⊣⊔⟩≀\⟩\⊔]∇↖
\]⊔<∫√]⊣∥]∇>
</√∇≀}∇⊣↕>

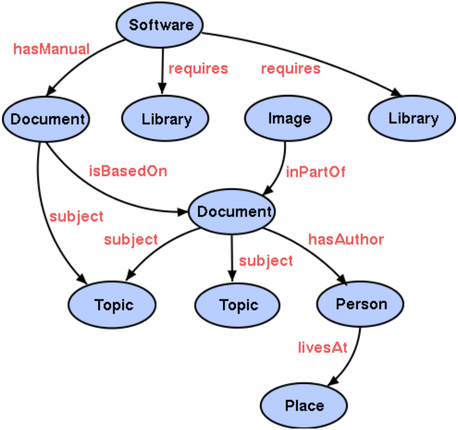
The Current Web

- **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.



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.



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

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

Towards a "Machine-Actionable Web"

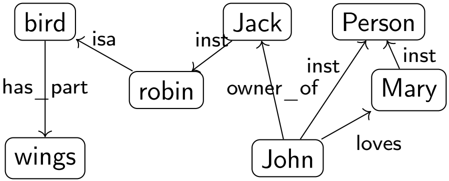
- **Recall:** We need external agreement on meaning of annotation tags.
- **Idea:** standardize them in a community process (e.g. DIN or ISO)
- **Problem:** Inflexible, Limited number of things can be expressed
- **Better:** Use ontologies to specify meaning of annotations
  - Ontologies provide a vocabulary 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+13] 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

Semantic Networks [CQ69]

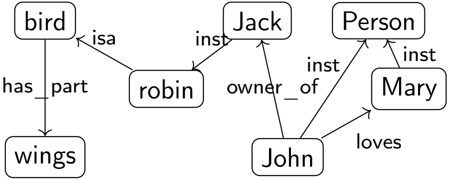
- **Definition 4.1.** A semantic network is a directed graph for representing knowledge:
  - nodes represent objects and concepts (classes of objects)
    (e.g. John (object) and bird (concept))
  - edges (called links) represent relations between these (isa, father_of, belongs_to)
- **Example 4.2.** 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.

Deriving Knowledge Implicit in Semantic Networks

- **Observation 4.3.** *There is more knowledge in a semantic network than is explicitly written down.*
- **Example 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 4.5.** We call links labeled by
  - "isa" an inclusion or isa link (inclusion of concepts)
  - "inst" instance or inst link (concept membership)

Deriving Knowledge Semantic Networks

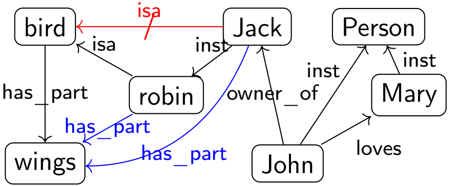
- **Definition 4.6 (Inference in Semantic Networks).** We call all link labels except "inst" and "isa" in a semantic network relations.

  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.

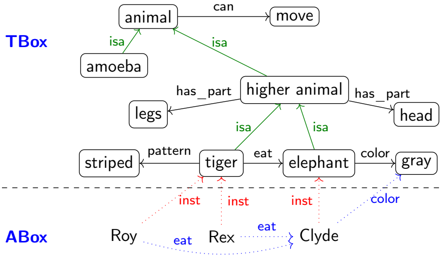
- **Example 4.7.** Derived relations in 4.4



- **Slogan:** Get out more knowledge from a semantic networks than you put in.

Terminologies and Assertions

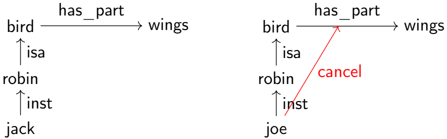
- *Remark 4.8.* We should distinguish concepts from objects.
- **Definition 4.9.** We call the subgraph of a semantic network N spanned by the isa links and relations between concepts the terminology (or TBox, or the famous Isa Hierarchy) and the subgraph spanned by the inst links and relations between objects, the assertions (or ABox) of N.
- **Example 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*).

Limitations of Semantic Networks

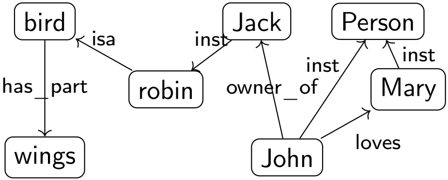
- What is the meaning of a link?
  - link labels are very suggestive (misleading for humans)
  - meaning of link types defined in the process model (no denotational semantics)
- **Problem:** No distinction of optional and defining traits!
- **Example 4.11.** Consider a robin that has lost its wings in an accident:



"Cancel-links" have been proposed, but their status and process model are debatable.

Another Notation for Semantic Networks

- **Definition 4.12.** Function/argument notation for semantic networks
  - interprets nodes as arguments (reification to individuals)
  - interprets links as functions (predicates actually)
- **Example 4.13.**

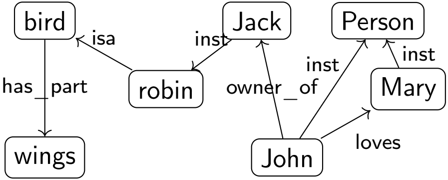


isa(robin,bird)
haspart(bird,wings)
inst(Jack,robin)
owner_of(John, robin)
loves(John,Mary)

- **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)

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)

What is an Ontology

- **Definition 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 4.15.** A vocabulary often includes names for classes and relationship (also called concepts, and properties).
- *Remark 4.16.* If the description logic has a reasoner, we can automatically
  - detect inconsistent axiom systems
  - compute class membership and taxonomies.

Semantic Web Technology in a Nutshell

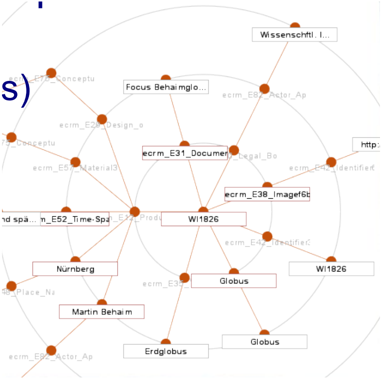
- Ontologies have become one of the standard devices for representing information about the Web and the world.
- **Definition 4.17.** This is facilitated and standardized by the :
  - URIs for representing objects,
  - RDF triples for representing facts,
  - RDFa for annotating RDF triples in XML documents,
  - 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,
  - Protg for authoring and maintaining ontologies,
- Details .

12.5 CIDOC CRM: An Ontology for Cultural Heritage

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



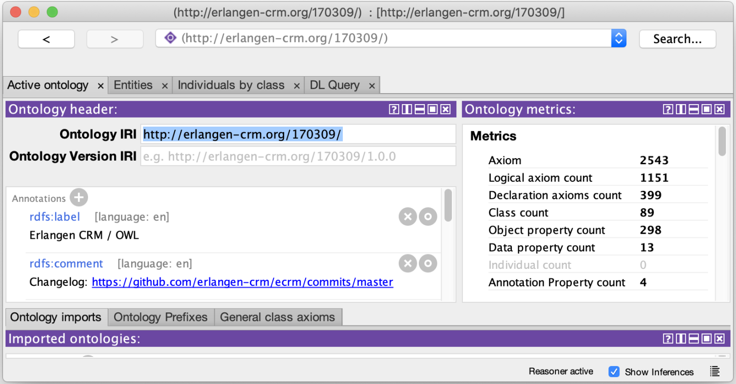
- Reference ontologies for interoperability:
  - SUMO (Suggested Upper Model Ontology) [SUMO] for common knowledge,
  - FOAF (Friend-of-a-Friend) [FOAF14] for persons and relations,
  - CIDOC CRM for documentation of cultural heritage. (up next)

CIDOC CRM (Conceptual Reference Model)

- **Definition 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 5.2.** OWL implements CIDOC CRM in OWL
- Details about CIDOC CRM can be found at [CC] and about OWL at [ECRMb; ECRMa].

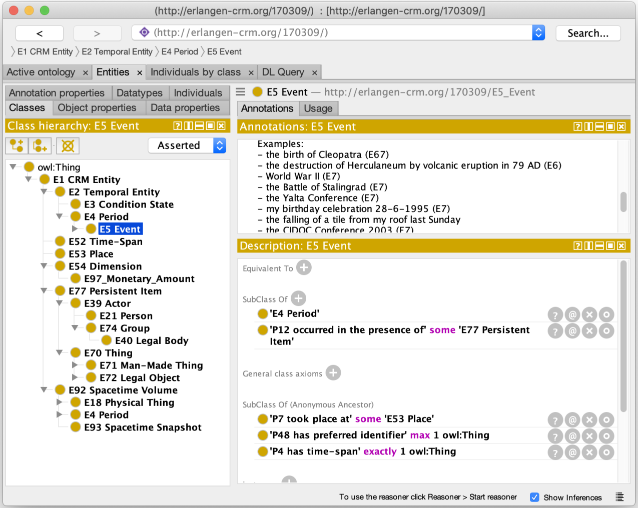
Protege, an IDE for Ontology Development

- **Definition 5.3.** Protg [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 5.4 (CIDOCCRM in Protege).**



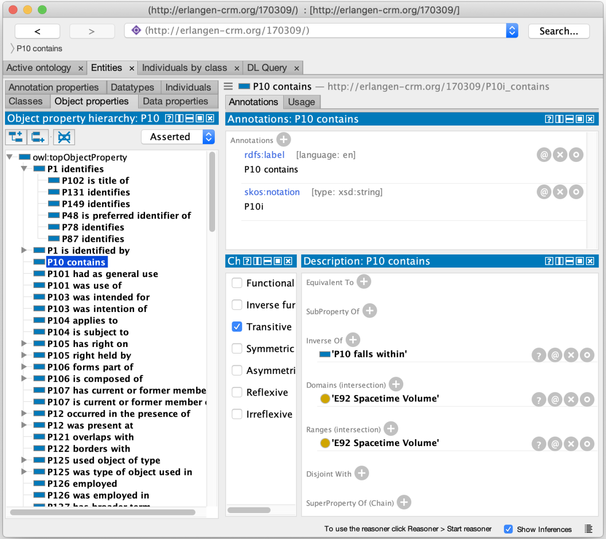
CIDOC CRM Explored (Classes)

- **Idea:** Use semantic web technology to explore OWL.
- **CIDOC CRM Classes:** concept $\widehat{=}$ OWL "Class" (shown in Protege)

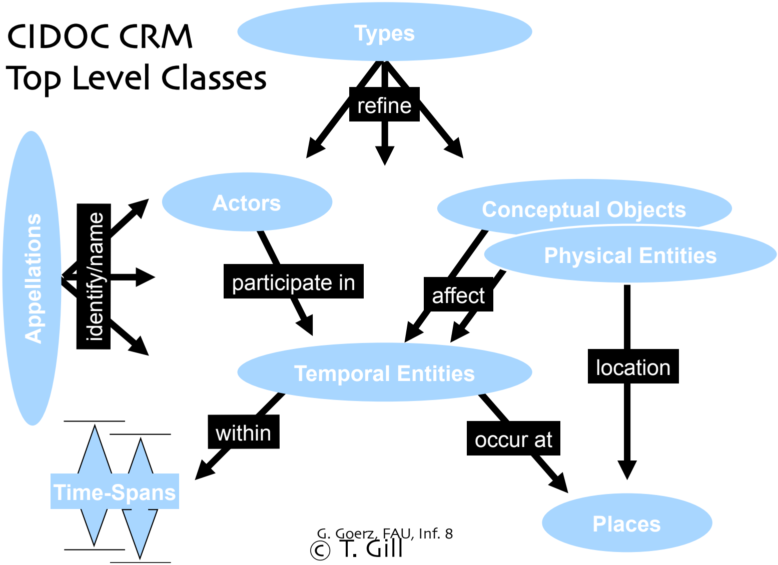


CIDOC CRM Explored (Relations)

- **CIDOC CRM Relations:** relation $\widehat{=}$ OWL "Object Property" (shown in Protege)



CIDOC CRM Structure (Overview)

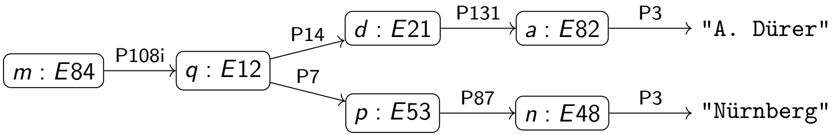


CIDOC-CRM Modeling

- **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 5.5.** *Albrecht Dürer painted Melencolia 1 in Nürnberg*
  We have two units of information here:
  1. Albrecht Dürer painted Melencolia 1
  2. this happened in the city of Nürnberg
- CIDOC CRM modeling decisions; we start with 1. *AD painted M 1*
  1. A painting m is an "Information Carrier" (E84)
  2. It was created in an "Production Event" q (E12)
  3. m is related to q via the "was produced by" relation (P108i)
  4. q was "carried out by" a "person" d (P14 E21)
  5. d "is identified by" an "actor appellation" a (P131 E82)
  6. a "has note" the string "Albrecht Dürer". (P3)
- CIDOC CRM modeling decisions; continuing with 2. *this happened in N*
  1. A painting m is an "Information Carrier" (E84)
  2. It was created in an "Production Event" q (E12)
  3. m is related to q via the " produced by" relation (P108i)
  4. q "took place at" a "place" p (P7 E53)
  5. p "is identified by" a "place name" n (P48 E3)
  6. n "has note" the string "Nürnberg". (P3)

CIDOC CRM Modelling (Ontology Paths)

- Modeling *Albrecht Dürer painted Melencolia 1 in Nürnberg* in CIDOC CRM



  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 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. ($\leadsto$ WissKI)

Event-Oriented Modeling in CIDOC CRM

- **Observation 5.7.** *Ontologies make it easy to model facts with transitive verbs, e.g. Albrecht Dürer created Melencolia 1 (binary relation)*
- **Problem:** What about more complex situations with more arguments? E.g.
  1. *Albrecht Dürer created Melencolia 1 with an etching needle* (ternary)
  2. *Albrecht Dürer created Melencolia 1 with an etching needle in Nürnberg* (four arguments)
  3. *Albrecht Dürer created Melencolia 1 with an etching needle in Nürnberg out of boredom* (five)
- **Standard Solution:** Introduce "events" tied to the verb and describe those
- **Example 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.

12.6 The Semantic Web Technology Stack

Resource Description Framework

- **Definition 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 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

Resources and URIs

- RDF describes resources with properties and property values.
- RDF uses Web identifiers (URIs) to identify resources.
- **Definition 6.3.** A resource is anything that can have a URI, such as `http://www.fau.de`.
- **Definition 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 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 6.6.** Statements: *[This slide]^{subj} has been [author]^{pred}ed by [Michael Kohlhase]^{obj}*

XML Syntax for RDF

- RDF is a concrete XML vocabulary for writing statements
- **Example 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.

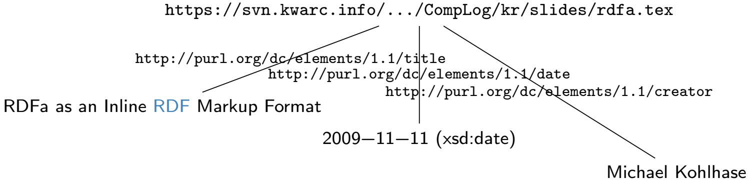
RDFa as an Inline RDF Markup Format

- **Problem:** RDF is a standoff markup format (annotate by URIs pointing into other files)

  Definition 6.8. RDFa (RDF annotations) is a markup scheme for inline annotation (as XML attributes) of RDF triples.

- **Example 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>
```



RDF as an ABox Language for the Semantic Web

- **Idea:** RDF triples are ABox entries h R s or $h{:}\varphi$.
- **Example 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 $\mathcal{ALC}$)

OWL as an Ontology Language for the Semantic Web

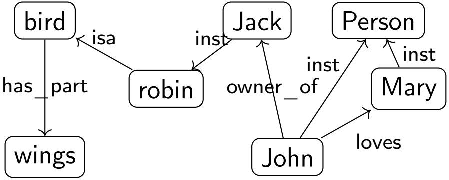
- **Task:** Complement RDF (ABox) with a TBox language.
- **Idea:** Make use of resources that are values in rdf:type. (called Classes)
- **Definition 6.11.** OWL (the ontology web language) is a language for encoding TBox information about RDF classes.
- **Example 6.12 (A concept definition for "Mother").**
  Mother$=$Woman $\sqcap$ 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`
	`  )`
	`)`

Extended OWL Example in Functional Syntax

- **Example 6.13.** The semantic network from 4.4 can be expressed in OWL (in functional syntax)



  - 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*.

Extended OWL Example in Functional Syntax

- **Example 6.14.** The semantic network from 4.4 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*.

SPARQL an RDF Query language

- **Definition 6.15.** 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 6.16.** A system is called a SPARQL endpoint, iff it answers SPARQL queries.
- **Example 6.17.** 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 Applications: DBPedia

- **Typical Application:** DBPedia screen-scrapes Wikipedia fact boxes for RDF triples and uses SPARQL for querying the induced triplestore.
- **Example 6.18 (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
```

Results: Browse Go! Reset

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

Triple Stores: the Semantic Web Databases

- **Definition 6.19.** 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 6.20.** 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

Example: Hogwarts Ontology

- **Example 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 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
```

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:

- Closed world assumption (CWA)
  - Missing information treated as false
- Unique name assumption (UNA)
  - Each individual has a single, unique name
- Schema behaves as constraints on structure of data
  - Define legal database states.

Ontology:

- Open world assumption (OWA)
  - Missing information treated as unknown
- No UNA
  - Individuals may have more than one name
- Ontology axioms behave like implications (inference rules)
  - Entail implicit information

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 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)

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?
- **Counting Query:** How many friends does Harry Potter have?

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?
- **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)

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?
- **Counting Query:** How many friends does Harry Potter have?
- **How about:** if we add

```
DifferentIndividuals: RonWeasley HermioneGranger
```

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?
- **Counting Query:** How many friends does Harry Potter have?
- **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)

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?
- **Counting Query:** How many friends does Harry Potter have?
- **How about:** if we add

```
DifferentIndividuals: RonWeasley HermioneGranger
```

- **And:** if we also add

```
Individual: HarryPotter
  Types: hasFriend only RonWeasley or HermioneGranger
```

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?
- **Counting Query:** How many friends does Harry Potter have?
- **How about:** if we add

```
DifferentIndividuals: RonWeasley HermioneGranger
```

- **And:** if we also add

```
Individual: HarryPotter
  Types: hasFriend only RonWeasley or HermioneGranger
```

  - DB: 2
  - Ontology: 2

DB vs. Ontology by Example (Insertion)

- **Given:** the ontology from 7.1 and 7.2 insert

```
Individual: Dumbledore
Individual: Fawkes
  Types: Phoenix
  Facts: isPetOf Dumbledore
```

- **System Response:**

DB vs. Ontology by Example (Insertion)

- **Given:** the ontology from 7.1 and 7.2 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)

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 7.3.** We call a query language semantic, iff query answering involves derived axioms and facts.
- **Observation 7.4.** *Ontology queries are semantic, while database queries are not.*

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.

Chapter 13
The WissKI System: A Virtual Research Environment for Cultural Heritage

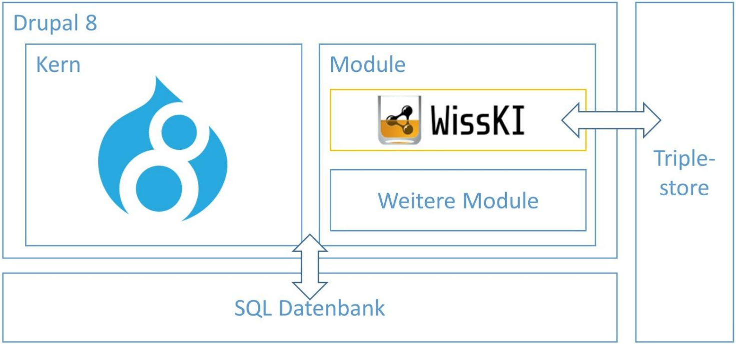
WissKI: a Virtual Research Env. for Cultural Heritage

- **Definition 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

13.1 WissKI extends Drupal

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.



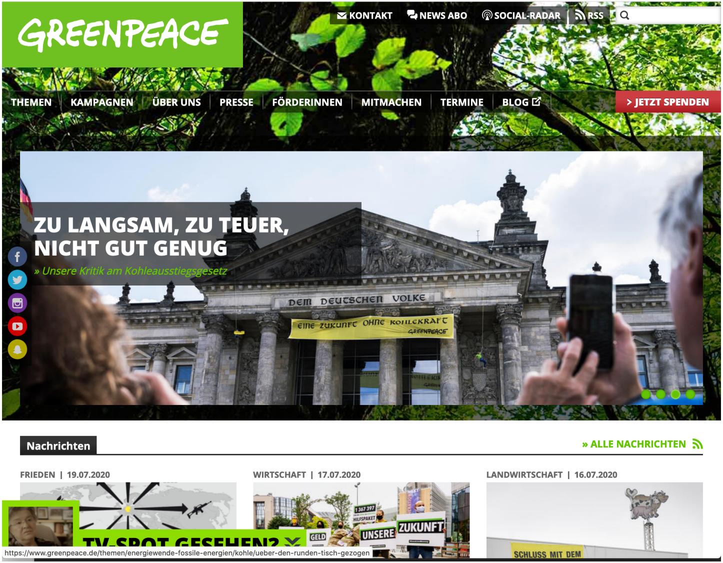
Drupal: A Web Content Managemt Framework

- **Definition 1.1.** Drupal is an open source web content management application. It combines CMS functionality with knowledge management via RDF.
- **Definition 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.

Assembling a Web Site via Drupal Blocks (Example)

- **Example 1.3 (Greenpeace via Drupal).** Can you find the blocks?



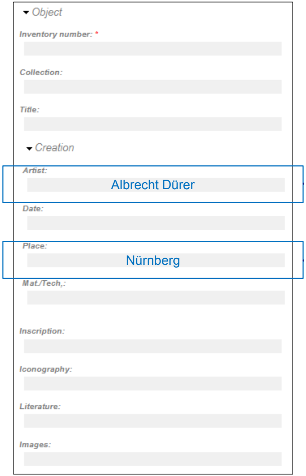
Drupal Modules and Themes

- **Idea:** Drupal is designed to be modular and extensible (so it can adapt to the ever-changing web)
- **Definition 1.4 (Modular Design).** Drupal functionality is structured into
  - drupal core – the basic CMS functionality
  - modules which contribute e.g. new block types ($\sim$ 45.000)
  - themes which contribute new UI layouts ($\sim$ 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

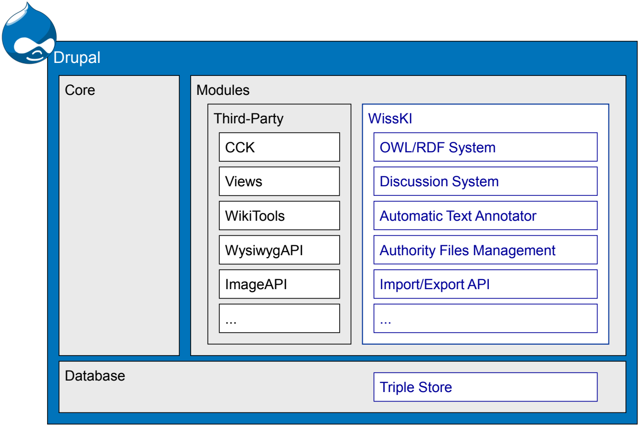
Bundles and Fields in Drupal (Data Entry)

- **Definition 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 $\rightsquigarrow$ 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.



WissKI System Architecture (Recap)

- WissKI = drupal + CIDOC CRM + triplestore + WissKI modules

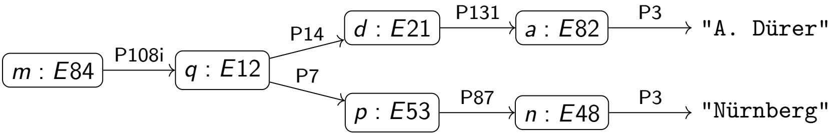


- **Note:** Much of WissKI functionality is configurable via the drupal config bar.

13.2 Dealing with Ontology Paths: The WissKI Pathbuilder

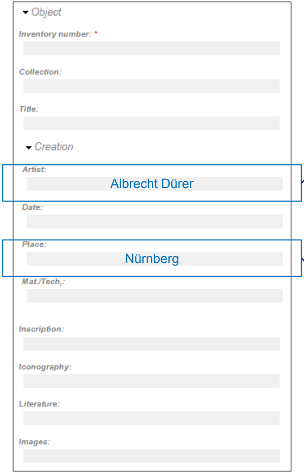
The WissKI Path Builder (Idea)

- **Recall:** *Albrecht Dürer painted Melencolia 1 in Nürnberg*



- **Idea:** Hide the complexity induced by the ontology from the user
  - Form-based interaction with categories and fields (as in a RDBMS UI)
- **Definition 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)

- **Example 2.2 (A WissKI Group).**

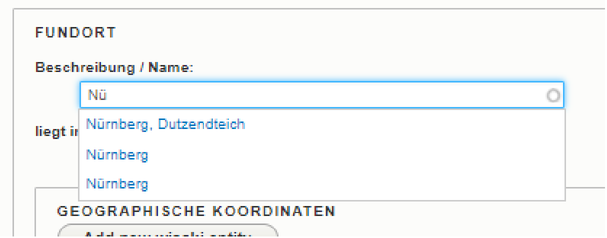


E84 Information Carrier
→ P108i was produced by →
E12 Production
→ P14 carried out by →
E21 Person
→ P131 is identified by →
E82 Actor Appellation
→ P3 has note →
„Albrecht Dürer“

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 “

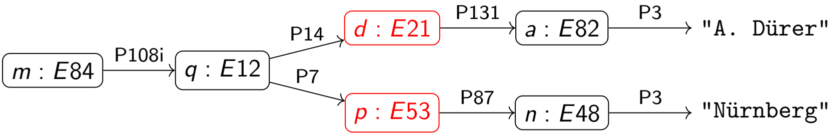
Sharing and Disambiguation in Path Builders

- **Observation 2.3.** *Sometimes we want to refer to existing entities in WissKI.*
- **Example 2.4 (Referring to Nürnberg).** (We love tab completion)



Sharing and Disambiguation in Path Builders

- **Observation 2.8.** *Sometimes we want to refer to existing entities in WissKI.*
- **Example 2.9 (Referring to Nürnberg).** (We love tab completion)
- **Example 2.10 (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 2.11.** A disambiguation point in a path marks an entity that can be re used in data acquisition.
- **Example 2.12.** Disambiguation points are highlighted in red on paths.



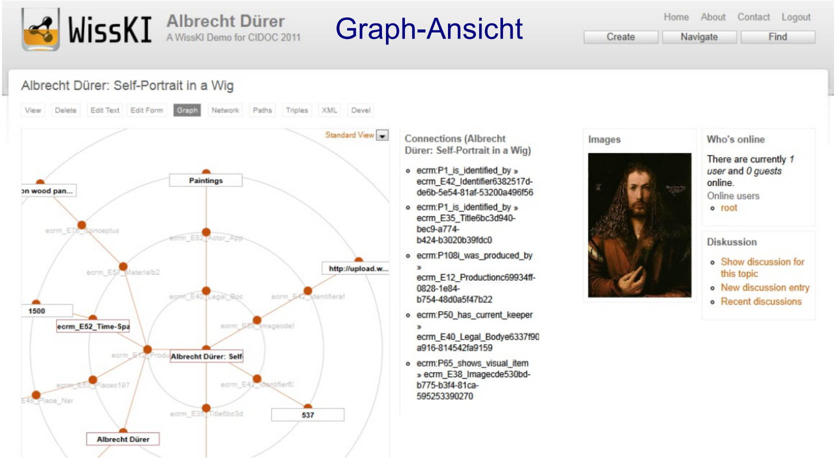
Specifying/Maintaining WissKI Path Builders

- **Recall:** A WissKI path builder maps ontology groups and ontology paths to drupal bundles and fields.
- **Example 2.13 (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

WissKI Path Builders as Graphs

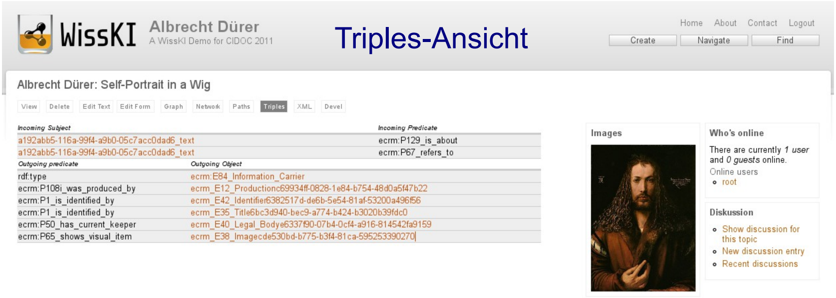
- **Example 2.14 (A WissKI Path Construtor as a Graph).**



- Very nice and helpful, but does not work currently!

WissKI Path Builders as Triples

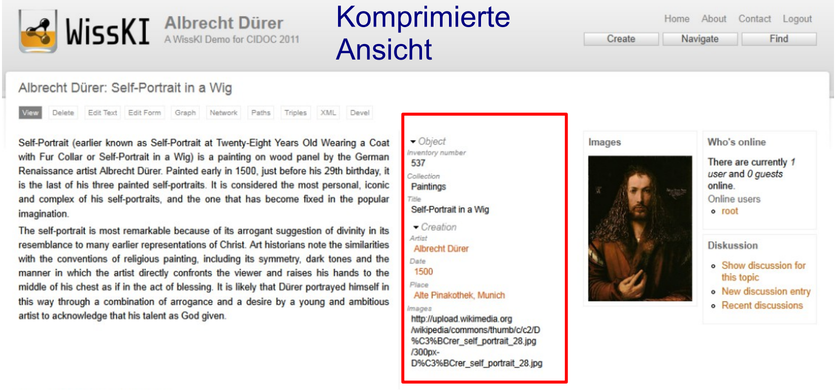
- **Of course** we can view path builders as sets of triples.
- **Example 2.15 (A WissKI Path Construtor as Triples).**



- Such an export also allows standardized communication.

Data Presentation using Path Builders in WissKI

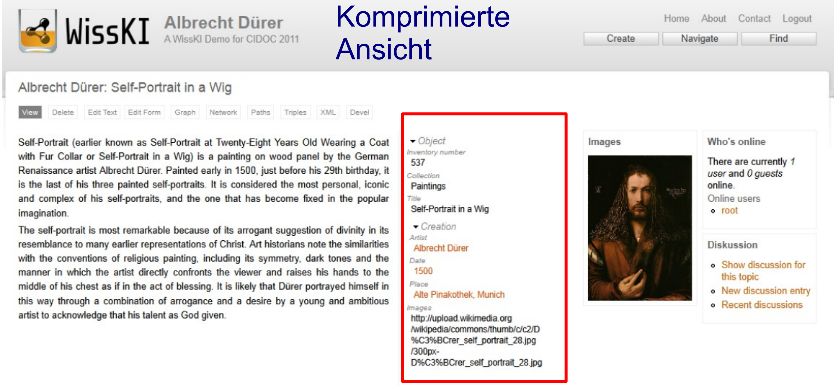
- Path builders can be used as drupal blocks for data presentation.
  - For every object o, aggregate the values of the paths starting in o.
- **Example 2.16 (Compressed View).**



13.3 The WissKI Link Block

The WissKI Link Block (Idea)

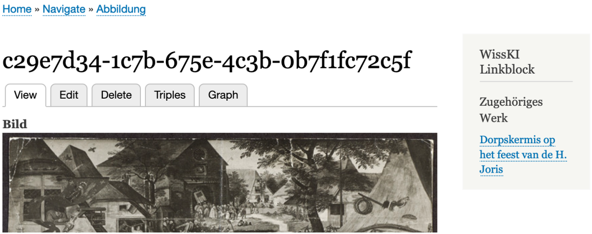
- **Observation 3.1.** *For an entity in a RDF graph, both the outgoing and the incoming relations are important for understanding.*
- **Example 3.2.** This view only shows the outgoing edges!



- **Idea:** Add a block with "incoming links" to the page, use the path builder.

Link Blocks (Definition)

- **Definition 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 3.4 (A link block for Images).**



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 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

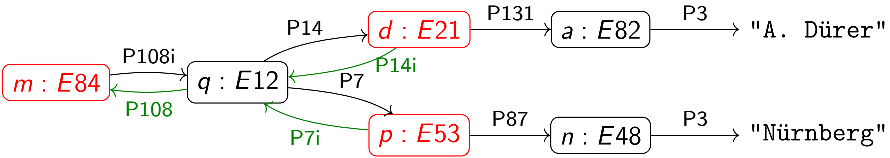
Zugehöriges Werk

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)



- **Note:** With this setup, you never have to fill out the link block paths!

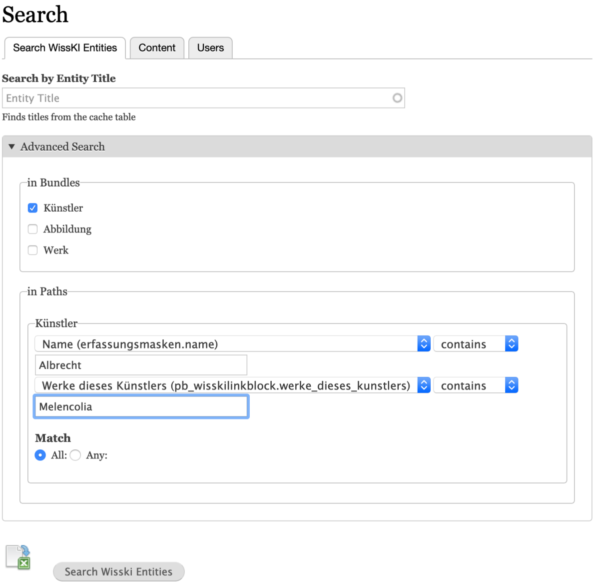
13.4 Cultural Heritage Research: Querying WissKI Resources

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.

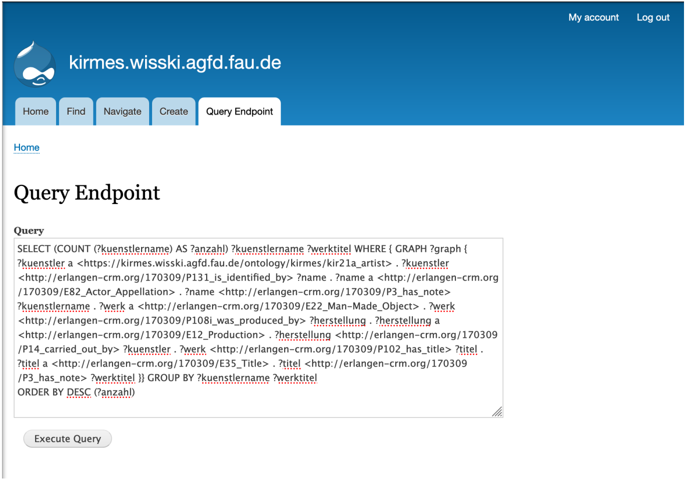
Drupal Search in WissKI

- **Example 4.1.**



SPARQL Endpoint in WissKI

- **Example 4.2.** Find kirmes paintings and their painters and count them



SPARQL Endpoint in WissKI

- **Example 4.3.** Find kirmes paintings and their painters and count them

kirmes.wisski.agfd.fau.de

Home | Find | Navigate | Create | Query Endpoint

Home

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"

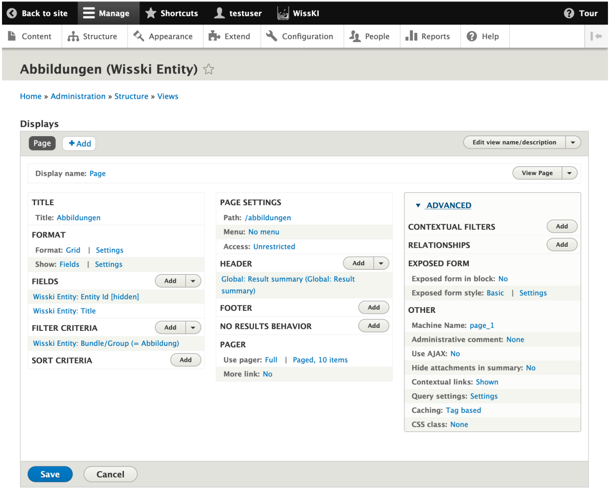
Query

```
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
```

Execute Query

Data Presentation via Views in WissKI

- **Example 4.4 (Configuring a View).** This makes a drupal block.



Drupal generates a SPARQL query, aggregates results into a block.

This Research is WissKI-instance-local

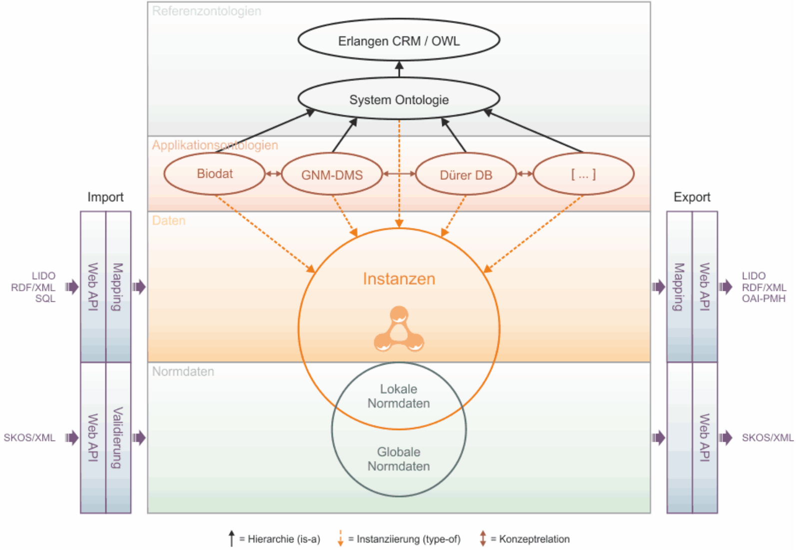
- **Observation 4.5.** *All these research queries only work in the current WissKI instance.*
- **Observation 4.6.** *There is probably much more about the entities you are interested in outside your particular WissKI instance.*
- **Problem:** How to make use of this?
- **Solution:** We need to do two 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

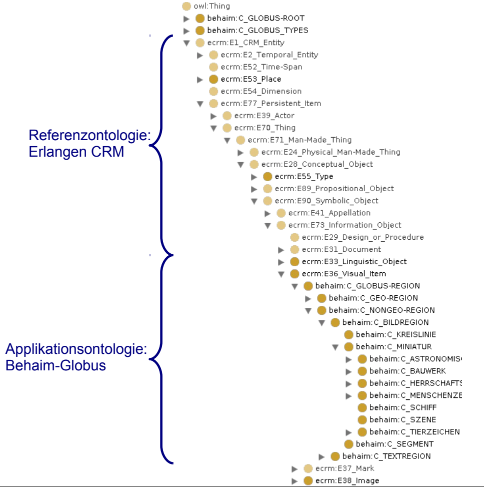
WissKI Information Architecture (Ontologies)

- Ontologies, instances, and export formats



Application Ontologies extend CIDOC CRM

- **Observation 5.1.** *Sometimes we need more than CIDOC CRM.*
- **Definition 5.2.** A WissKI application ontology is one that extends CIDOC CRM, without changing it.
- **Example 5.3 (Behaim Application Ontology).**



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 Protg for that.

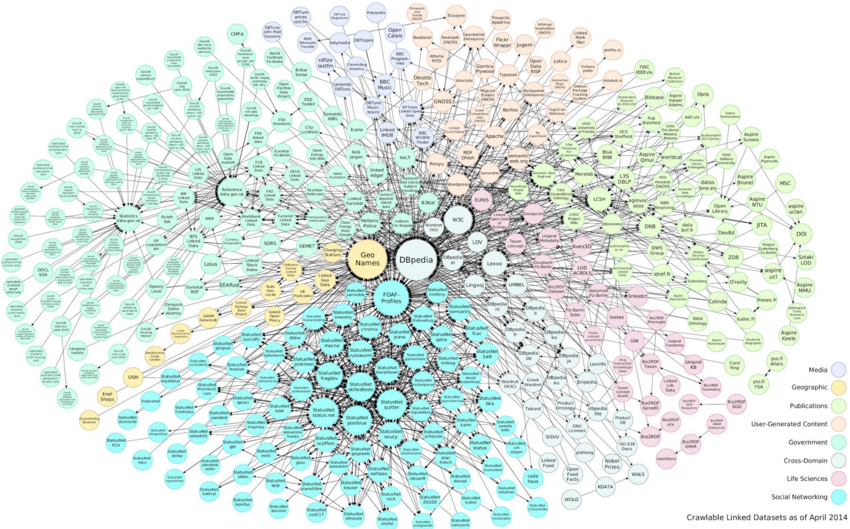
13.6 The Linked Open Data Cloud

Linked Open Data

- **Definition 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 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 6.3.** Given the semantic web technology stack, we can create interoperable ontologies and interlinked data sets, we call their totality the .
- **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.

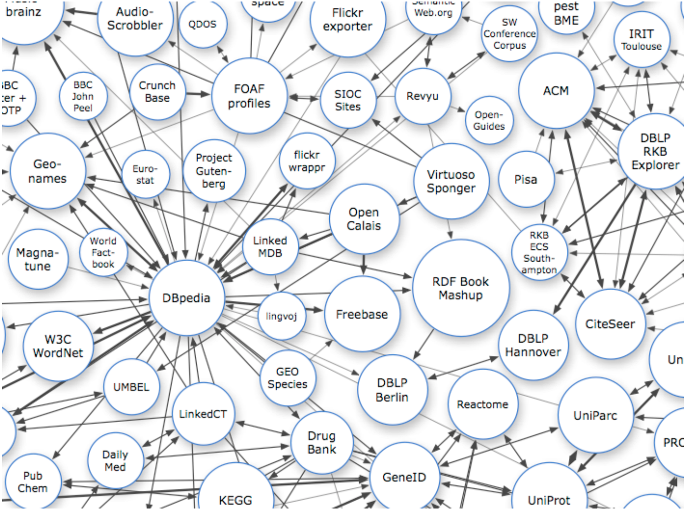
The Linked Open Data Cloud

- The linked open data cloud in 2014 (today much bigger, but unreadable)



The Linked Open Data Cloud

- ▶ Zooming in (data sets and their – interlinked – ontologies)



Using the LOD-Cloud in WissKI

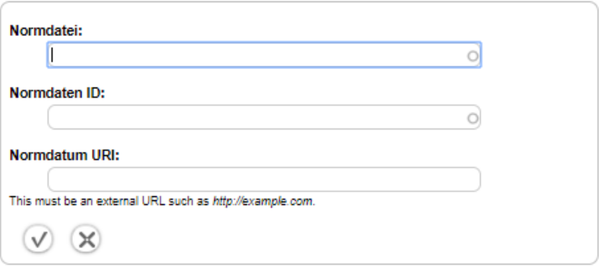
- **Idea:** Do not re-model entities that already exist (in the LOD Cloud)
- **Problem:** Most of the LOD Cloud is about things we do not want.
- But there are some sources that are useful
  - the GND (Gemeinsame Normdatei [GND]), an authority file for personal/corporate names and keywords from literary catalogs,
  - geonames[GN], a geographical database with more than 25M names and locations
  - Wikipedia
- **Observation 6.4.** *All of them provide URIs for real world entities, which is just what we need for objects in RDF triples.*
- **Definition 6.5.** WissKI provides special modules called adapters for GND and geonames.

Using Geonames in WissKI (Example)

1. **Example 6.6.** We want to use the "Meilwald" (Erlangen) in WissKI.

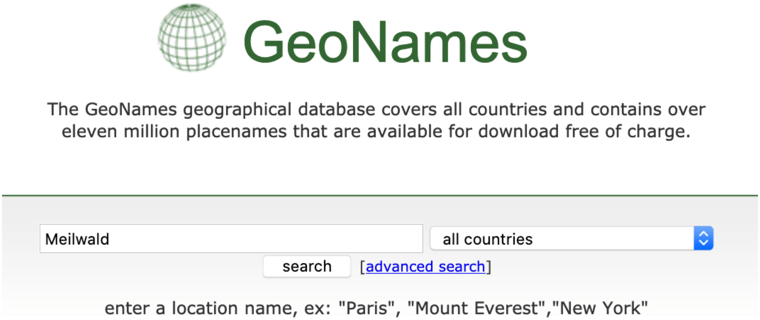
Using Geonames in WissKI (Example)

1. **Example 6.7.** 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:



Using Geonames in WissKI (Example)

1. **Example 6.8.** 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:
4. We enter `https://geodata.org` for "Normdatei" and go there to find out the URI for "Meilwald" which goes into "Normdatum URI".



Using Geonames in WissKI (Example)

1. **Example 6.9.** 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:
4. We enter `https://geodata.org` for "Normdatei" and go there to find out the URI for "Meilwald" which goes into "Normdatum URI".
5. there may be multiple results (here only one)


Meilwald | all countries
search [advanced search]

1 records found for "Meilwald"

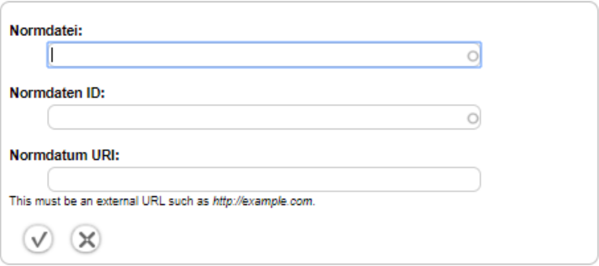
	Name	Country	Feature class	Latitude	Longitude
1	Erlanger Meilwald Erlanger Meil-Wald,Erlanger Meilwald,Meilwald	Germany, Bavaria	forest(s)	N 49° 36' 30''	E 11° 1' 39''

Using Geonames in WissKI (Example)

1. **Example 6.10.** 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:
4. We enter `https://geodata.org` for "Normdatei" and go there to find out the URI for "Meilwald" which goes into "Normdatum URI".
5. there may be multiple results (here only one)
6. Select/click the intended one, check the details

Using Geonames in WissKI (Example)

1. **Example 6.11.** 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:
4. We enter `https://geodata.org` for "Normdatei" and go there to find out the URI for "Meilwald" which goes into "Normdatum URI".
5. there may be multiple results (here only one)
6. Select/click the intended one, check the details
7. Enter the URL from the URL bar into "Normdatum URI".



Towards a WissKI Commons in the LOD Cloud

- **Recap:** We can directly refer to (URIs of) external objects in WissKI.
- **Observation 6.12.** *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 6.13.** We call the part of this resource that can be created by aggregating WissKI exports the WissKI commons.
- **Observation 6.14.** *WissKI exports meet the FAIR principles quite nicely already.*
- We will be working on a FAU WissKI commons in the next years. (help wanted)

Chapter 14
Legal Foundations of Information Technology

14.1 Intellectual Property

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 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.

Background: Property and Ownership in General

- **Definition 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 1.3.** Ownership involves multiple rights (the property rights), which may be separated and held by different parties.
- **Definition 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 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 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.

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.

Legal Traditions

- The various legal systems of the world can be grouped into “traditions”.
- **Definition 1.7.** Legal systems in the common law tradition are usually based on case law, they are often derived from the British system.
- **Definition 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.

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 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 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.

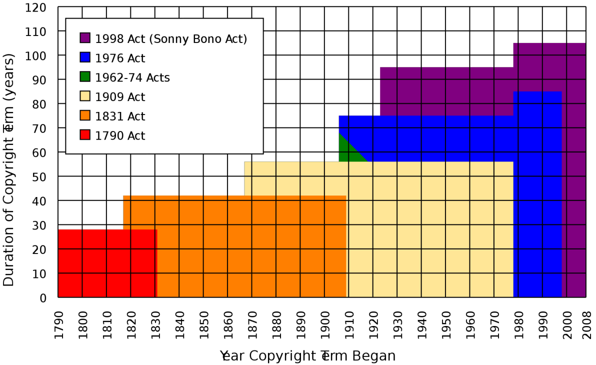
14.2 Copyright

Copyrightable Works

- **Definition 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)

Limitations of Copyrightabilitiy: The Public Domain

- **Definition 2.2.** A work is said to be in the public domain, if no copyright applies, otherwise it is called copyrighted.
- **Example 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 2.4 (US Copyright Terms).** Some people claim that US copyright terms are extended, whenever Disney's Mickey Mouse would become public domain.



Rights under Copyright Law

- **Definition 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 2.6.* Formally, it is not the copyrightable work that can be owned itself, but the copyright.
- **Definition 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.

Copyright Holder

- **Definition 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 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 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.*

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 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

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 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.

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 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.

GPL/Copyleft: Creating a FLOSS Public Domain?

- **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 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 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, ...)

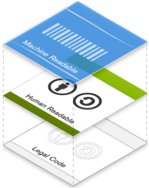
Open Content/Data via Open Licenses

- **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!

Creative Commons a System of Open Content Licenses

Definition 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 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:

+/- attribuition	(must reference the author)
+/- commercial use	(can be restricted)
+/- derivative works	(can allow modification)
+/- share alike (copyleft)	(modifications must be donated back)

14.4 Information Privacy

Information/Data Privacy

- **Definition 4.1.** The principle of information 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 General Data Protection Regulation (GDPR)

- **Definition 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 4.3.* As an EU regulation, the GDPR is directly effective in all EU member countries. (enforced since 2018)
- **Axiom 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 4.5.** Physical access control: Unauthorized persons may not be granted physical access to data processing equipment that process personal data. ($\rightsquigarrow$ locks, access control systems)
- **Definition 4.6.** System access control: Unauthorized users may not use systems that process personal data. ($\rightsquigarrow$ passwords, firewalls, ...)
- **Definition 4.7.** Information access control: Users may only access those data they are authorized to access. ($\rightsquigarrow$ access control lists, safe boxes for storage media, encryption)
- **Definition 4.8.** Data transfer control: Personal data may not be copied during transmission between systems. ($\rightsquigarrow$ encryption)
- **Definition 4.9.** Input control: It must be possible to review retroactively who entered, changed, or deleted personal data. ($\rightsquigarrow$ authentication, journaling)
- **Definition 4.10.** Availability control: Personal data have to be protected against loss and accidental destruction. ($\rightsquigarrow$ physical/building safety, backups)
- **Definition 4.11.** Obligation of separation: Personal data that was acquired for separate purposes has to be processed separately.

Personally Data (GDPR)

- **Definition 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 4.13.** We collectively call direct identifiers and identifying collections of quasi-identifiers personally identifying information (PII).
- **Example 4.14.** Quasi-identifiers include name, date of birth, race, location, ...
- **Definition 4.15.** Personal data (also called personal information) is any information relating to an identified or identifiable person.
- **Example 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 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 $\leadsto$ 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)

Chapter 15
Collaboration and Project Management

15.1 Revision Control Systems

15.1.1 Dealing with Large/Distributed Projects and Document Collections

Web Development Scenario

▶ **Example 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.

Web Development Scenario

- **Example 1.2.**
  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.

Web Development Scenario

- **Example 1.3.**
  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 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.

Web Development Scenario

- **Example 1.4.**
  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 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.

Web Development Scenario

- **Example 1.5.**
  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 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.

Web Development Scenario

- **Example 1.6.**
  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 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).

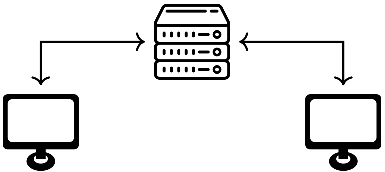
Architecture of Revision Control Systems

- **Observation:** We distinguish three large classes of RCS.
- In local RCS, a working copy uses a repository on the same machine.

- We will go through these in explaining the respective features as we go along.

Architecture of Revision Control Systems

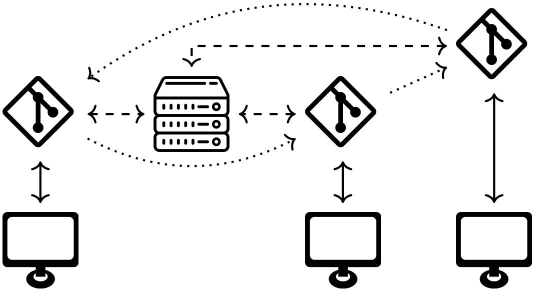
- **Observation:** We distinguish three large classes of RCS.
- In local RCS, a working copy uses a repository on the same machine.
- In a centralized RCS, the repository is on a central repository server.



- We will go through these in explaining the respective features as we go along.

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

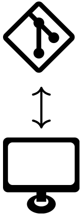
- GIT is a powerful distributed revision control system.
- GIT is the current dominant RCS, exceeding 90% adoption in open source projects and high utilization in industry.
- GIT features a well-designed set of primitive revision control actions, from which complex behaviours can be composed.
- **In 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

Revision Control Systems

- **Definition 1.7.** 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 1.8.** In a RCS, users do not directly work on the repository, but on a working copy that is synchronized with the repository.
- **Definition 1.9.** 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 1.10.** *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.

Computing and Managing Differences with `diff` & `patch`

- **Definition 1.11.** `diff` is a file comparison utility that computes differences between two strings or text files: the source f_1 and the target f_2. Differences are output linewise in a diff $\delta(f_1, f_2)$.
- **Definition 1.12.** `patch` is a sister utility that applies a diff $\delta{:=}\delta(f_1, f_2)$ to f_1 – resulting in f_2; we say it patches f_1 with δ.
- **Example 1.13.** We compare two simple text files:

The quick brown fox jumps over the lazy dog	The quack brown fox jumps over the loozy dog	1c1,2 < The quick brown --- > The quack brown > 3c4 < the lazy dog --- > the loozy dog

- **Definition 1.14.** A diff 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.

15.1.3 GIT as a local Revision Control System

Working with GIT

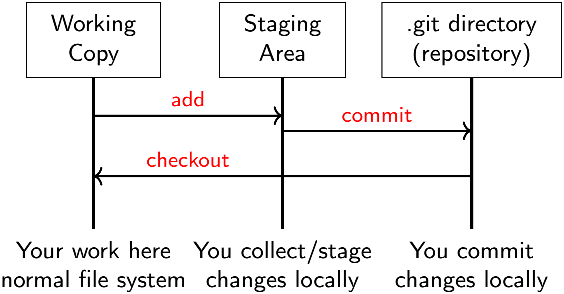
- **Observation:** GIT can be used in many situations.
- **On your Laptop:** for software development
  - Download GIT from `https://git-scm.com/downloads`, install (you want to use it on your local machine)
  - We will use GIT from the shell on your system (`MacOSX` or `linux`) or `GitBash`, a shell that comes with your GIT download (`Windows`). (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.

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 (`MacOSX` or `linux`) or `GitBash` that comes with your GIT download (`Windows`). (graphical front ends exist but hinder understanding)
- Test whether your installation works: git version (should be $\geq$ 2.30)
- **Definition 1.15.** 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 $\langle\langle \mathrm{name} \rangle\rangle$ makes working copy + local repository in the $\langle\langle \mathrm{name} \rangle\rangle$ subdirectory.

Working with GIT (Staging and Committing)

- **Overview:** GIT local workflow: staging files for commit using git add



commits acts only on staged files $\leadsto$ git add foo.tex (GIT must know about them)

Working with GIT (Staging and Committting)

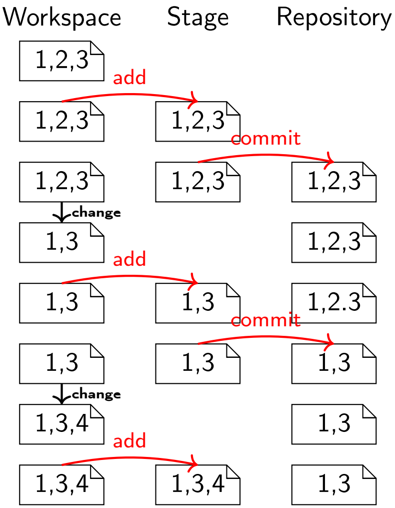
- Basic GIT commands: (many variants and options ⇝ study them)

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.

An Example Git Workflow

- **Example 1.16.** A typical, elementary workflow in GIT in a shell.

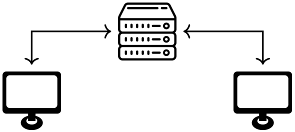
```
> git init
Initialized empty Git repository in /tmp
> echo "1,2,3" > test.txt
> git add test.txt
> git commit −m'initializing'

> 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
```



15.1.4 Centralized Revision Control: Collaboration

Collaboration via Centralized RCS

- **Definition 1.17.** 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.

Merging Differences

- There are basically two ways of merging the differences of files into one.
- **Definition 1.18.** In two way merge, an automated procedure tries to combine two different files by copying over differences by guessing or asking the user.
- **Definition 1.19.** 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 $\mu_i^3(f_1, f_2, p)$, 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

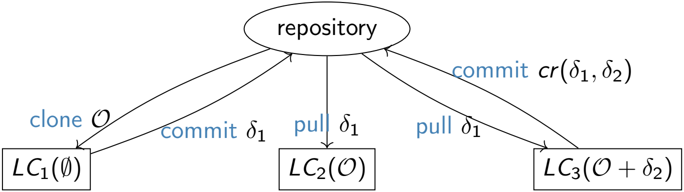
- **Definition 1.20.** The `merge3` tool computes a three way merge.
- **Example 1.21.** We compare two simple text files with a parent:

<table>
<tr><td>mine.txt</td><td>your.txt</td><td>parent.txt</td><td>conflict marker</td></tr>
<tr><td>This is the file.
Hello</td><td>This is the file.
hello</td><td>This is the file.
hi</td><td>This is the file.
<<<<<<< mine.txt
Hello
||||||| parent.txt
hi
=======
hello
>>>>>>> your.txt</td></tr>
</table>

- **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.

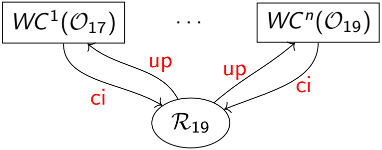
Collaboration via Centralized RCS (Example)

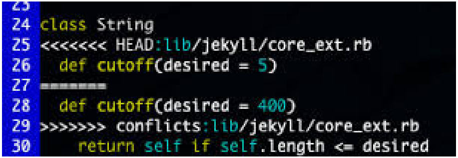
▶ **Example 1.22 (A Workflow with three Working Copies).**



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:



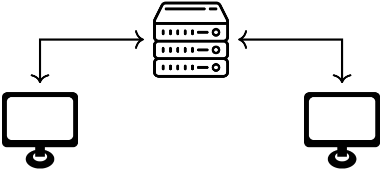


  - you can only commit, if your revision is the head (otherwise update)
  - update merges the changes into your working copy.
  - If there are changes on the same line, you have a conflict, which must be resolved.

15.1.5 GIT as a centralized RCS

Recap: Centralized RCS

- **Idea:** In a centralized RCS, the repository resides on a repository server.



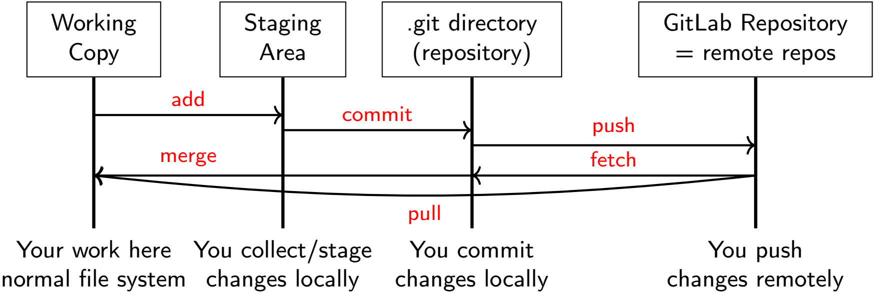
- **Problem:** We need some generalizations over local RCS:
  - Identifying the repository server.
  - Pushing and fetching over the network.

Working with 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 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':
...
```

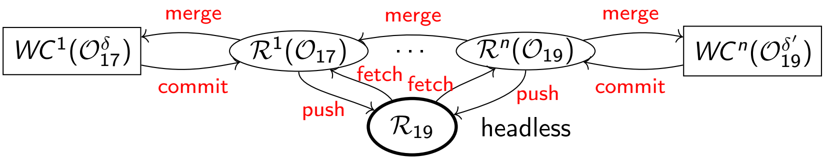
15.1.6 Distributed Revision Control

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)

Distributed Version Control

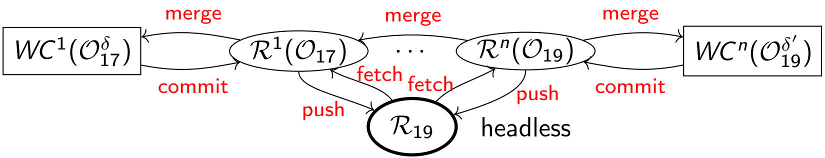
- **Problems with Centralized Revision Control:**
  1. We can only commit when online! (but we work on the train)
  2. All collaboration goes via one, central repository. (prescribes workflow)
- **Idea:** Distribute the repositories and move patches between them.



  1. local commits to local repositories
  2. all repositories created equal (flexible organization)

Distributed Version Control

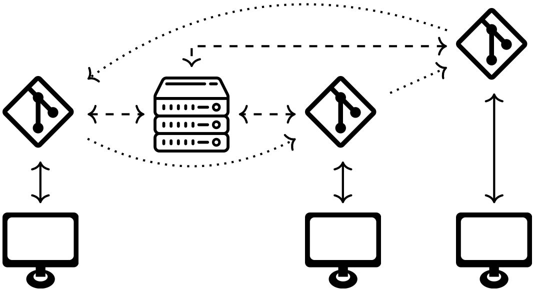
- **Problems with Centralized Revision Control:**
  1. We can only commit when online! (but we work on the train)
  2. All collaboration goes via one, central repository. (prescribes workflow)
- **Idea:** Distribute the repositories and move patches between them.



  1. local commits to local repositories
  2. all repositories created equal (flexible organization)
- **Definition 1.27.** We call a revision control system distributed, iff it allows multiple repositories that can exchanged patches.
- **Definition 1.28.** We call a repository headless (or bare), if used without a working copy.

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.



- **Definition 1.29.** We call a revision control system distributed, iff it allows multiple repositories that can exchanged patches.
- **Definition 1.30.** 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.

Distributed Version Control with GIT

- **Definition 1.31.** 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

GIT Branches and Forks

- GIT special commands for making, switching, and merging branches.

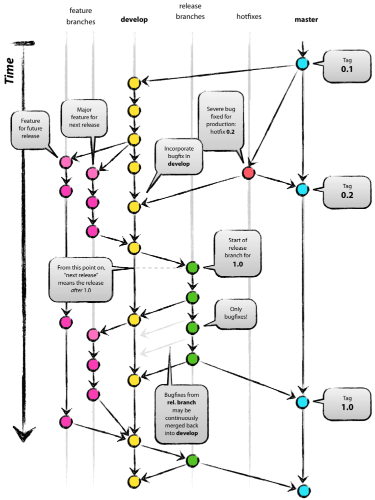
git branch ⟪branch⟫	makes a branch with name ⟪name⟫
git checkout ⟪branch⟫	switches a working copy to branch ⟪branch⟫
git branch −v	shows all branches
git branch −d ⟪branch⟫	deletes branch ⟪branch⟫

- **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.

GitFlow: An Elaborate Development Model based on GIT

- **Definition 1.32 (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.



15.2 Working with GIT and GitLab/GitHub

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 2.1.** A repository management system is an web application that supports the administration of a repository server and manages authentication and authorization.
- **Example 2.2.** GitLab is an open source repository management system and repository hosting service at `http://GitLab.com`. (free public/private repositories)
- **Definition 2.3.** A repository hosting service is a web based repository management system that also offers storage space for repositories.
- **Example 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).

Working with GitLab/GitHub (continued)

- **Definition 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. 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!* ⚠

Making Repositories on GitLab

- Make a new project with , play with it (you can always delete it)
- **Definition 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

Authorization in GitLab: Managing Access Permissions

- **Definition 2.7.** Authorization refers to a set of rules that determine who is allowed to do what.
- **Definition 2.8.** Authorization is often operationalized by assigning permission levels and binding the authorization to execute particular interactions to permission levels.
- **Definition 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

Authentication

- **Definition 3.1.** Authentication is the process of ascertaining that somebody really is who they claim to be.
- **Definition 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 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.

Authentication by Cryptographic Public Keys

- **Definition 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 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 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 $t=t'' \leadsto$ OK, iff $t=t''$.
- **Advantage:** Passwords canot be leaked, need not be transmitted, retyped.

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.

15.4 Bug/Issue Tracking Systems

Bug/Issue Tracking Systems

- **Definition 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 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.

The Anatomy of an Issue

- **Definition 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 4.4.** A feature request is an issue that only specifies the expected behavior and proposes ways of implementing that.

Markdown a simple Markup Format Generating HTML

.

- **Idea:** We can translate between markup formats.
- **Definition 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 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`.

Markdown a simple Markup Language Generating HTML

▶ **Example 4.7.** We show the most important Markdown commands.

<table>
<tr><th>Markdown syntax</th><th>Generated HTML</th></tr>
<tr><td>

```
# 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).
```

</td><td>

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, <code>monospace</code>.
Bullet list:
• apples
• oranges
• pears
Numbered list:
1. apples
2. oranges
3. pears
A link.

</td></tr>
</table>

Markdown a simple Markup Language Generating HTML

- **Example 4.8.** We show the most important Markdown commands.

Markdown syntax	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](http://example.com).
	<**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**>

GitHub flavored markdown: Tracker Specific Extensions

- *Remark 4.9.* Source code hosting systems offer special extensions for referencing their components.
- **Definition 4.10.** GitHub flavored markdown (GFM) is a markdown dialect extended for the use in GIT-based issue tracking systems; see [Gfm] for the specification.
- **Example 4.11.** 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 4.12.** *Very useful for project planning and reporting in GitLab and GitHub.*

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 4.13 (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 4.14 (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.

Bugtracker Workflow

- **Definition 4.15 (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 4.16.** *An issue tracker can serve as a full blown project planning system, if used accordingly.*
- **Definition 4.17.** For timing work plans, most issue trackers provide milestones that issues can be targeted to.

Administrative Metadata for Issues

- To make the issue based workflows work we need data.
- **Definition 4.18 (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?

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 strutures like numbers and strings, character encodings, unicode, and regular expressions
- Digital 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 servers
  - serverside computing: bottle routing and
  - client-side 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

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

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

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

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)

References I

[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).

[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.

[CSS] *CSS Specificity*. url: https://en.wikipedia.org/wiki/Cascading_Style_Sheets#Specificity (visited on 12/03/2018).

References II

[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).

[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).

[FAIR18] European Commission Expert Group on FAIR Data. *Turning FAIR into reality*. 2018. doi: 10.2777/1524.

References III

[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).

[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+13] 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/.

References IV

[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/.

[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.

References V

[LM] *LabelMe: the open annotation tool*. url: http://labelme.csail.mit.edu (visited on 08/28/2020).

[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).

[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).

References VI

[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).

[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/.

References VII

[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.

[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.

[UL] *urllib – URL handling modules*. url: https://docs.python.org/3/library/urllib.html (visited on 04/15/2021).

References VIII

[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.