MathDataHub - your dataset, but FAIR

Katja Berčič, Michael Kohlhase, Florian Rabe, Tom Wiesing
Computer Science, FAU Erlangen-Nürnberg

May 22, 2020
Seminar for Mathematical Data

Motivation: Mathematical Data

- There are a lot of different kinds of mathematical data
  - concrete data (**record** or **array** data)
  - symbolic data (**computation**, **decuction**, **modelling**)
  - linked data (**metadata**, **knowledge graph**s)
  - narrative data (**notations**, **documents**, **visualisations**, **verbalisations**)
- we heard about some of this in more detail last time
  - I will try to keep this talk self-contained
  - But: I will try to avoid going into too much details if we already knew them

Motivation: FAIR Data

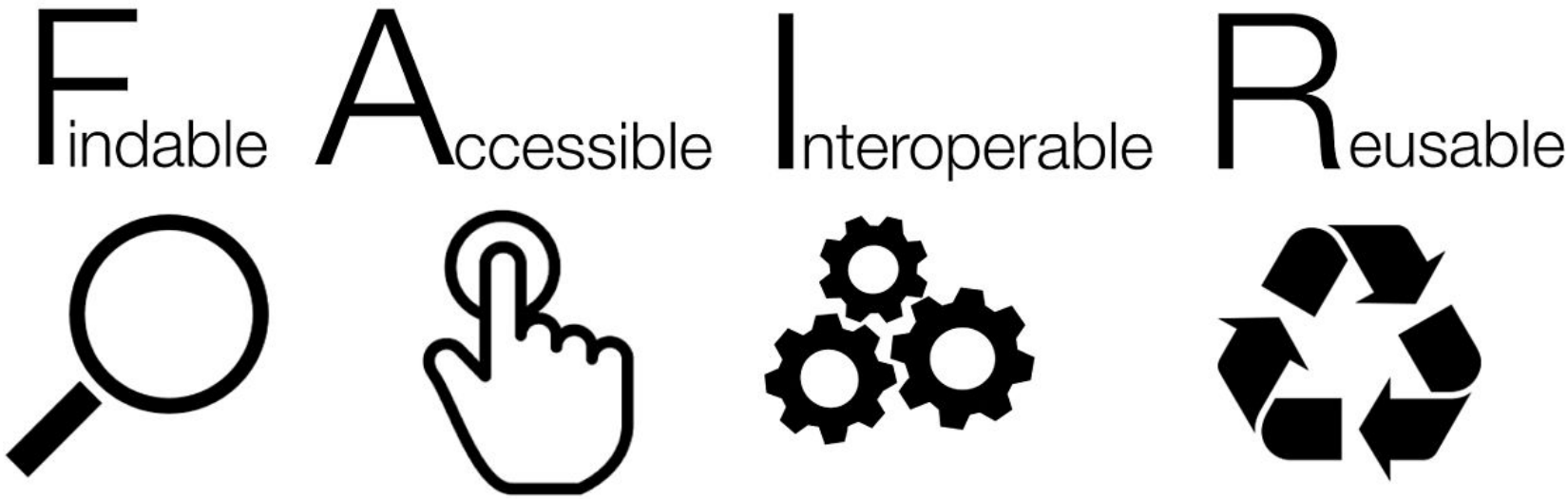


Image Source: Wikipedia, licensed under CC BY-SA 4.0.

Goals of MathDataHub

- Problem: Typical Math Datasets are not FAIR
  - hard to achieve, especially if it is not in focus
- Solution: Provide a **generic** infrastructure
  - make it easy for mathematicans
- MathDataHub aims to be such an infrastructure

What MathDataHub Can Do

A census of small connected cubic vertex-transitive graphs

All connected cubic vertex-transitive graphs of order at most **1280**.

This dataset has 111360 objects.

Matches found: *111355*

More about this dataset

Display results

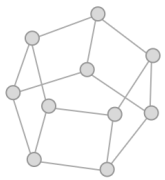
Available conditions

- Order
- CVT Index
- Graph
- Name
- Clique Number

Active conditions

- Order>=10

Choose columns

	Order	CVT Index	Graph	Name	Clique Number	Diameter	Girth	Is Arc-Transitive	Is Bipartite	Is Cayley	Is Hamiltonian	Is Prism	Is Split Praeger-Xu
	10	1		5-Prism	2	3	4	false	false	true	true	true	false
	10	2		5-Möbius	2	3	4	false	true	true	true	false	false

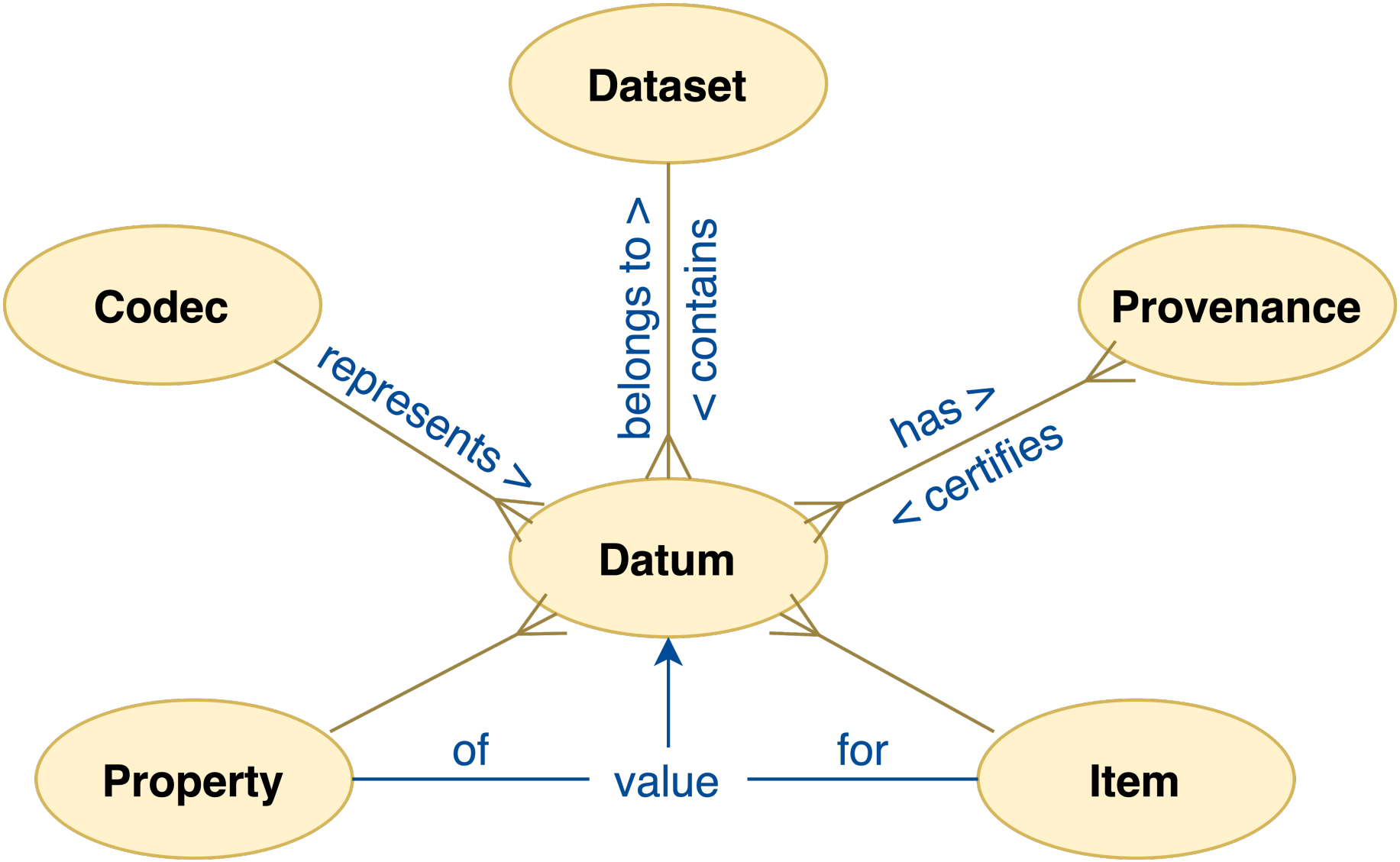
MathDataHub – Architecture Overview

- stores and represents mathematical data in a generic data model
  - (more about this on the next slide)
- all data is stored in a **PostgreSQL** database
  - Pros: this can handle a lot of data efficiently
  - Cons: Requires some optimization (e.g. using "materialized database views")
- Backend written in Python using a web-framework called **Django**
  - Pros: We do not have to manually create (and update) SQL table structures
  - Cons: We had to write a lot of custom code to make **importing** datasets faster
- Frontend written in **TypeScript** and **React**
  - TypeScript is a typed version of JavaScript
  - React is an MVC framework originally developed by Facebook
- developed as a part of MathHub

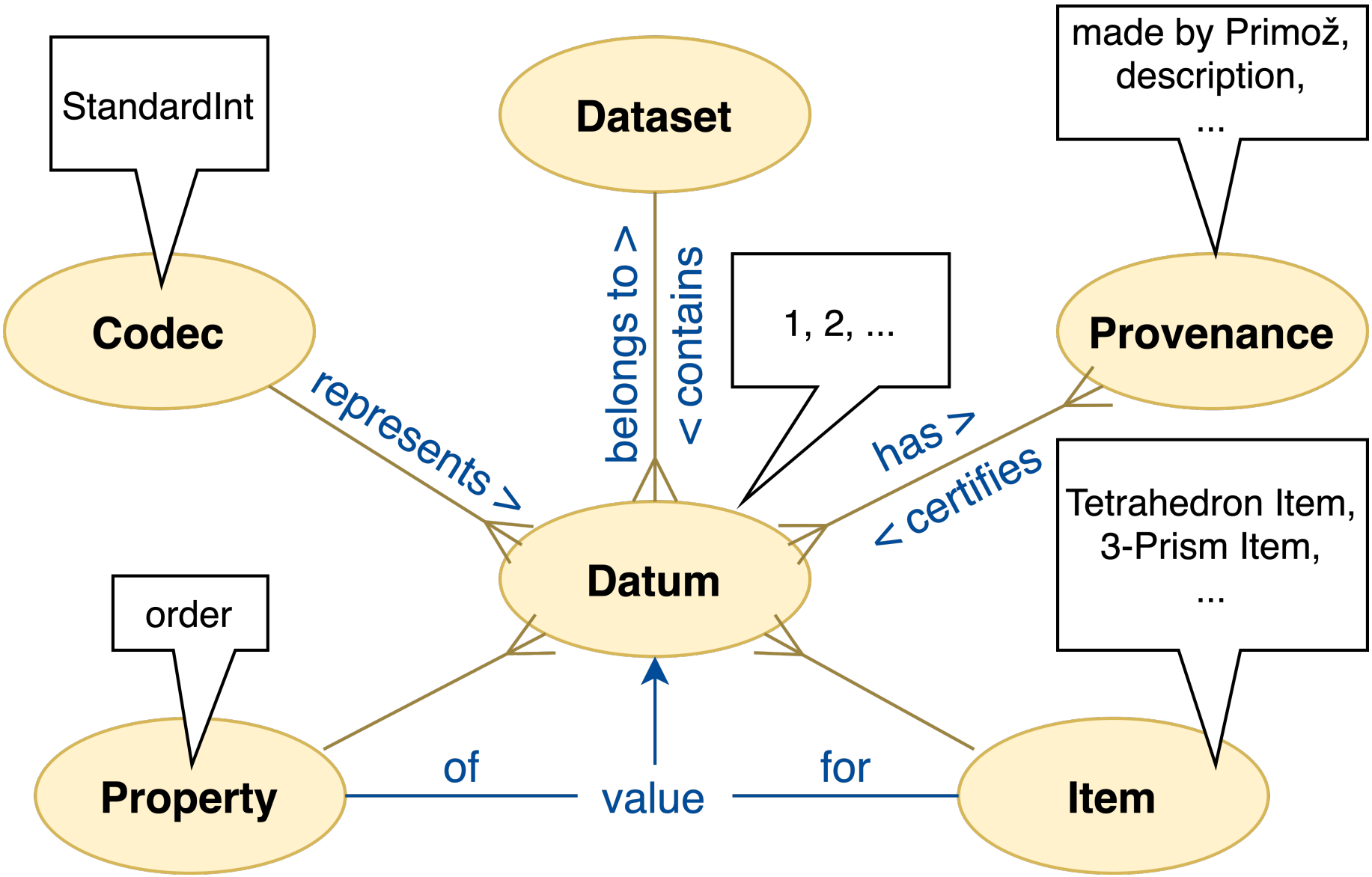
A concrete example

- Example: "A census of small connected cubic vertex-transitive graphs"
  - all connected cubic vertex-transitive graphs of order at most 1280
  - **cvt** for short
  - contributed and authored **Primož Potočnik** et al.
  - now available at `https://data.mathhub.info/collection/cvt`
- collection has several properties
  - 22 properties e.g. **order**, **name**, **graph**, **girth**, ...
  - 111360 items
- we will investigate the **order** property
  - an **integer** value
  - represents the **number of vertices in the graph**
  - stored using **database integers**

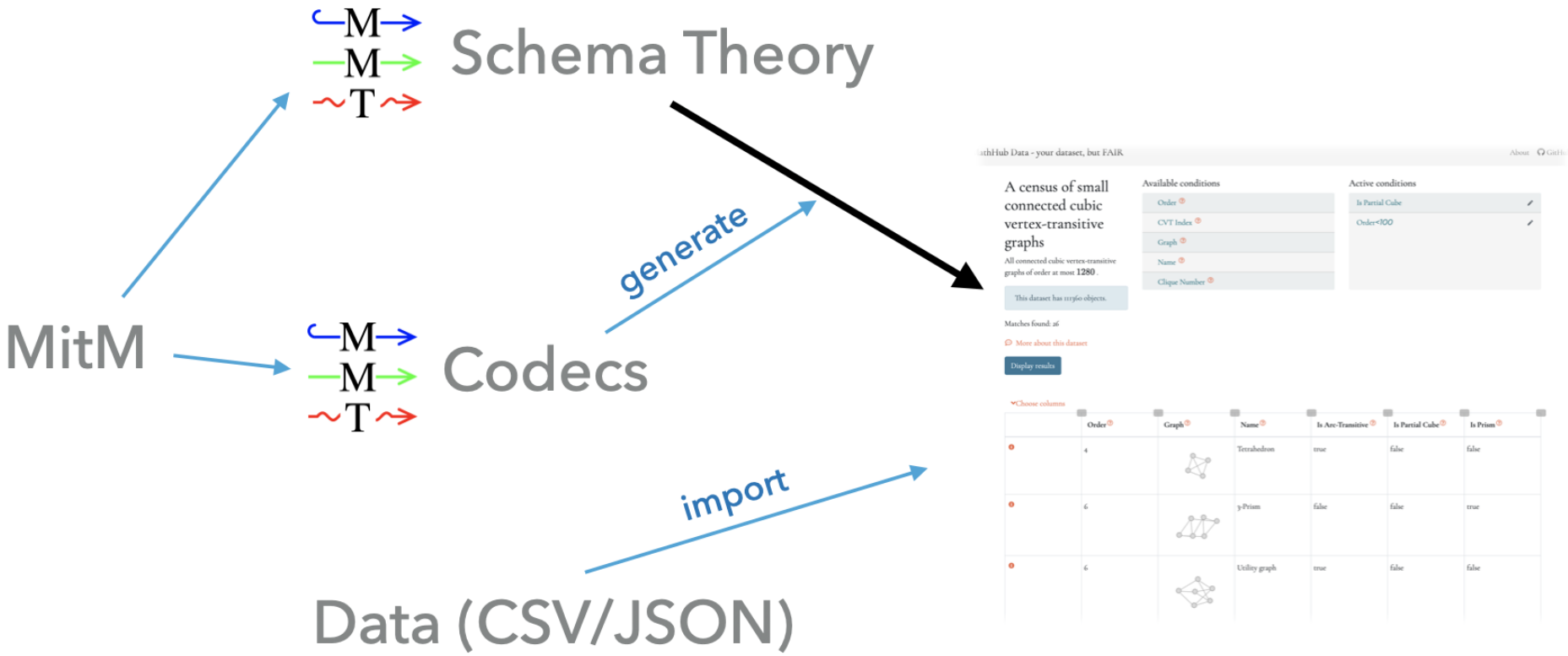
Under the Hood – Data Model



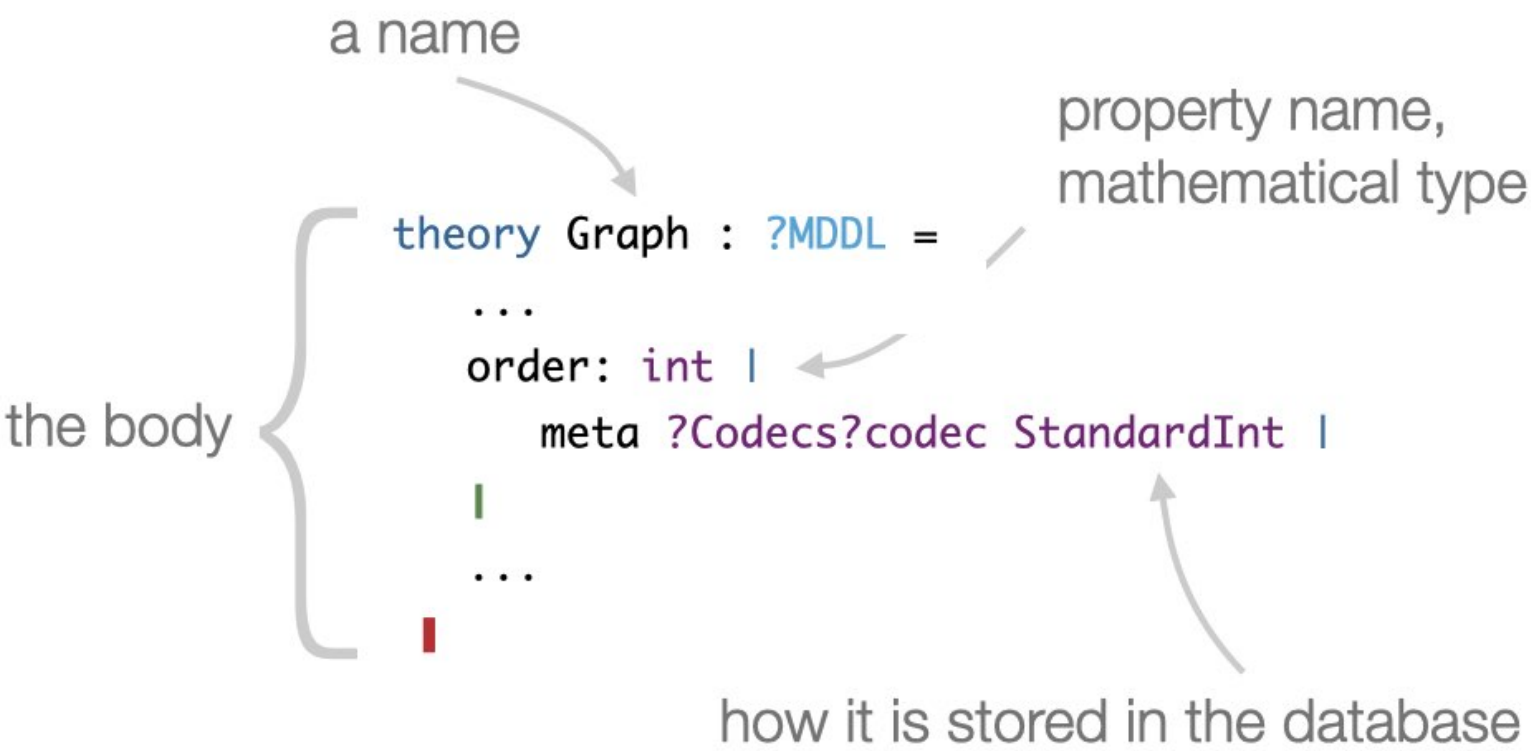
Under the Hood – Data Model



How To Import Your Dataset



How To Import Your Dataset – Schema Theory



How To Import Your Dataset – Schema JSON

```
{
 "slug": "cvt",
 "displayName": "A census of small connected cubic vertex-
     transitive graphs",
 "description": "connected cubic vertex-transitive graphs",
 // ... some properties omitted ...

 "metadata": {
  "schemaTheoryURL": "gl.mathhub.info/ODK/mbgen/cvt_schema.mmt",
  // ... other metadata omitted ...
 },

 "properties": [
  {
   "slug": "order",
   "displayName": "Order",
   "codec": "StandardInt",
   "description": "Number of vertices in the graph."
  },

  // ... more properties ...
 ]
}
```

Summary

- Summary
  - there is a lot of **mathematical datasets** out there
  - it is desirable to make them **FAIR**
  - **MathDataHub** is a generic system that allows you doing so
  - **Codecs** tell the system how a certain object is represented
  - an **MDDL** schema is required to import a new dataset
  - the system will then **generate** the userinterface automatically
  - check out `https://data.mathhub.info`
- Questions, Comments, Concerns?
- Thank You For Listening!