OpenDreamKit Work Package 6
Data/Knowledge/Software-Bases

Michael Kohlhase
FAU Erlangen-Nürnberg
http://kwarc.info/kohlhase

OpenDreamKit Final Review, Luxembourg, October 30. 2019
1 Introduction
Background: WP6 (Data/Knowledge/Software-Bases)

From the Proposal:

- **Knowledge**
  - Scientific journals
  - wikipedia.org
  - Teaching resource repositories
  - mathoverflow.net
  - Collaborative help centers
  - Mailing lists

- **Researcher Communities**
  - Preprint servers
    - arxiv.org
  - oeis.org
  - lmfdb.org
  - findstat.org
  - User interfaces
    - (Jupyter notebook, ...)
  - OpenDreamKit
  - Collaborative workspaces
  - Continuous integration tools
  - Code repositories
  - Issue trackers
  - Software Forges

- **Data**
  - Databases
  - Online databases
    - oeis.org
    - lmfdb.org
    - findstat.org
  - Preprint servers
    - arxiv.org

- **Software**
  - User interfaces
    - (Jupyter notebook, ...)
  - OpenDreamKit
  - Collaborative workspaces
  - Continuous integration tools
  - Code repositories
  - Issue trackers
  - Software Forges

- **Computational resources**
  - (local, super computer, cloud)

- **Storage resources**
  - (local, shared folder, cloud)
From the Proposal:

Proposed Focus: Supply this data to VRE components in an integrated fashion programmatically
Results of the WP6 Workshops: Semantic Interoperability

The WP6 group had a series of workshops

Kickoff in Paris (Sep ’15): strategies for joint knowledge representation

- WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
- WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
- WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
- WS in Oslo (Aug ’17): Active, Structured Documents
- WS in Cernay (May ’18): Interfacing Systems
- WS at FLoC Oxford (July ’18): Aspects of Modular Knowledge
- WS in Cernay (August ’19): Math Data Workshop

Kohlhase: ODK WP6 3 ODK Final Review, Oct. ’19
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug ’17): Active, Structured Documents
  - WS in Cernay (May ’18): Interfacing Systems
  - WS at FLoC Oxford (July ’18): Aspects of Modular Knowledge
  - WS in Cernay (August ’19): Math Data Workshop

Kohlhase: ODK WP6 3 ODK Final Review, Oct. ’19
Results of the WP6 Workshops: Semantic Interoperability

The WP6 group had a series of workshops

- Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
- WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.

- WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
- WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
- WS in Oslo (Aug ’17): Active, Structured Documents
- WS in Cernay (May ’18): Interfacing Systems
- WS at FLoC Oxford (July ’18): Aspects of Modular Knowledge
- WS in Cernay (August ’19): Math Data Workshop

Kohlhase: ODK WP6 3 ODK Final Review, Oct. ’19
The WP6 group had a series of workshops:

- Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
- WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
- WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
- WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
- WS in Oslo (Aug ’17): Active, Structured Documents
- WS in Cernay (May ’18): Interfacing Systems
- WS at FLoC Oxford (July ’18): Aspects of Modular Knowledge
- WS in Cernay (August ’19): Math Data Workshop
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug ’17): Active, Structured Documents
  - WS in Cernay (May ’18): Interfacing Systems
  - WS at FLoC Oxford (July ’18): Aspects of Modular Knowledge
  - WS in Cernay (August ’19): Math Data Workshop
  - Kohlhase: ODK WP6 3 ODK Final Review, Oct. ’19
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug ’17): Active, Structured Documents

Mass-energy equivalence

The energy \( E \) is the quantity of energy transferred to an object in order to alter the object. The mass \( m \) is both a measure of the object's resistance to acceleration (state of motion) when a net force acts on the object.

The speed of light in vacuum, \( c \), is a universal, physical constant important to our everyday life. Its exact value is 299,792,458 m/s (approximately 300,000 km/s or 18,628 miles/s).

Combining these quantities with the equation \( E = mc^2 \).

In [1]: theory MassEnergyEquivalence

theory MassEnergyEquivalence

In [2]: include ?MEC

include http://cds.cern.ch/record/120591

In [3]: active computation \( m, c, mc^2 \)

\((mc^2)\)

m

\(c\)

Simplify

Click simplify to start
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep '15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb '16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June '16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb '17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug '17): Active, Structured Documents
  - WS in Cernay (May '18): Interfacing Systems

Sage    MMT    GAP

LMFDB
Results of the WP6 Workshops: Semantic Interoperability

The WP6 group had a series of workshops:

- **Kickoff in Paris (Sep ’15):** strategies for joint knowledge representation
- **WS in St. Andrews (Feb ’16):** Math in the Middle Arch. for System Interop.
- **WS in Bremen (June ’16):** GAP/SageMath API Content Dictionaries (CDs)
- **WS in Berlin (Feb ’17):** Math-in-the-Middle Ontology
- **WS in Oslo (Aug ’17):** Active, Structured Documents
- **WS in Cernay (May ’18):** Interfacing Systems

1: \( Q@S \)

\[ \text{Sage} \quad \text{MMT} \quad \text{GAP} \]

\[ \text{LMFDB} \]
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug ’17): Active, Structured Documents
  - WS in Cernay (May ’18): Interfacing Systems

\[
\begin{align*}
1: & \quad Q@S \\
\text{Sage} & \quad \rightarrow & \quad \text{MMT} \\
2: & \quad Q’@L \\
\text{LMFDB} & \quad \rightarrow & \quad \text{GAP}
\end{align*}
\]
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug ’17): Active, Structured Documents
  - WS in Cernay (May ’18): Interfacing Systems

![Diagram]

1: $Q@S$

Sage $\rightarrow$ MMT $\rightarrow$ GAP

2: $Q’@L$

3: $34G@L$

LMFDB
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep ‘15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug ’17): Active, Structured Documents
  - WS in Cernay (May ’18): Interfacing Systems

- WS at FLoC Oxford (July ’18): Aspects of Modular Knowledge
- WS in Cernay (August ’19): Math Data Workshop

Kohlhase: ODK WP6 3 ODK Final Review, Oct. ’19
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep '15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb '16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June '16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb '17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug '17): Active, Structured Documents
  - WS in Cernay (May '18): Interfacing Systems

\[
\begin{align*}
1: & \quad Q@S \\
2: & \quad Q'@L \\
3: & \quad 34G@L \\
4: & \quad 34G@G \\
5: & \quad [t,t,\ldots]
\end{align*}
\]
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops
  - Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug ’17): Active, Structured Documents
  - WS in Cernay (May ’18): Interfacing Systems

1: $\mathcal{Q}@\mathcal{S}$
4: $34G@G$
6: [t,t,...]
5: [t,t,...]
3: $34G@L$
2: $Q’@L$

LMFDB
The WP6 group had a series of workshops

- Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
- WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
- WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
- WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
- WS in Oslo (Aug ’17): Active, Structured Documents
- WS in Cernay (May ’18): Interfacing Systems
- WS at FLoC Oxford (July ’18): Aspects of Modular Knowledge
Results of the WP6 Workshops: Semantic Interoperability

- The WP6 group had a series of workshops:
  - Kickoff in Paris (Sep ’15): strategies for joint knowledge representation
  - WS in St. Andrews (Feb ’16): Math in the Middle Arch. for System Interop.
  - WS in Bremen (June ’16): GAP/SageMath API Content Dictionaries (CDs)
  - WS in Berlin (Feb ’17): Math-in-the-Middle Ontology
  - WS in Oslo (Aug ’17): Active, Structured Documents
  - WS in Cernay (May ’18): Interfacing Systems
  - WS at FLoC Oxford (July ’18): Aspects of Modular Knowledge
  - WS in Cernay (August ’19): Math Data Workshop
WP6 Focus in the Final Review Period

- Inference \(\sim\) the Isabelle Library and MitM
- Mathematical Data
  - (Semantic) Interoperability with Mathematical Data
  - Strengthening Organization via stronger Schemata
  - Collecting mathematical Data during computation
- Data and Inference are a central part of “doing mathematics”.

Math Data

MitM

Documents

Inference

Computation

(Math-in-the-Middle)

(LMFDB)

(Persistent Memoization)
2 Extending OpenDreamKit (MitM) to Inference
New Task 6.11: *Isabelle Case Study* (last Amendment)

**Idea:** Math uses a mixture of computation and proving.

**Isabelle:** One of the most mature and widely used proof assistants
- 82 out of Wiedijk’s top 100 math theorems formally proved
- L4 microkernel verification: > $10^5$ loc
- Archive of Formal Proof
  - > 300 authors, > 500 articles, > $10^5$ lemmas, > $10^6$ loc

**Subcontract:** Collaboration with Makarius Wenzel (main Isabelle developer)
Serialize Isabelle libraries in exchange formats (OMDoc/MMT) ($\approx 6 + 4$ PM)

**input**
- > $10^4$ theories/locales
- > $10^6$ definitions and theorems
- 135 MB uncompressed text files

**output (without proofs)**
- 206 MB compressed OMDoc (37.5 GB uncompressed)
- > $10^8$ RDF triples

**run time:** 12 hours with 8 CPU cores, 50 GB memory
3 MathHub Data – your dataset, but FAIR
Definition 3.1. Research data is recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings.

Background: Virtually all scientific funding agencies now require some kind of research data strategy (tendency: getting stricter)

Definition 3.2 (Gold Standard Criteria). Research data has to be FAIR, i.e.

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.

Questions: What does this mean for mathematics, in particular

What is mathematical research data?

How can we make mathematical data FAIR?
Definition 3.1. Research data is recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings.

Background: Virtually all scientific funding agencies now require some kind of research data strategy (tendency: getting stricter)

Definition 3.2 (Gold Standard Criteria). Research data has to be FAIR, i.e.

- 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.
**FAIR Research Data: The Next Big Thing**

- **Definition 3.1.** Research data is recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings.

- **Background:** Virtually all scientific funding agencies now require some kind of research data strategy (tendency: getting stricter).

- **Definition 3.2 (Gold Standard Criteria).** Research data has to be FAIR, i.e.
  - **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.

**Questions:** What does this mean for mathematics, in particular

- What is mathematical research data?
- How can we make mathematical data FAIR?
The Current Reality in Mathematical Practice

▶ 80% of the datasets are not FAIR.

(here are three silos)

- Graphs of order 4 to 300 (18 MB)
- Graphs of order 302 to 500 (66 MB)
- Graphs of order 502 to 600 (69 MB)
- Graphs of order 602 to 700 (84 MB)
- Graphs of order 702 to 800 (114 MB)
- Graphs of order 802 to 900 (147 MB)
- Graphs of order 902 to 1000 (183 MB)
- Graphs of order 1002 to 1050 (164 MB)
- Graphs of order 1052 to 1100 (113 MB)
- Graphs of order 1102 to 1150 (103 MB)
- Graphs of order 1152 to 1200 (234 MB)
- Graphs of order 1202 to 1250 (137 MB)
- Graphs of order 1252 to 1280 (131 MB)

▶ Idea: Provide semantic hosting of all of these.

| [N,i] | V | E | Tr | W? | B? | |AGl | vs | ds | #STO | gi |
|-------|---|---|----|----|----|-------|---|---|------|---|
| C4(5.1) | 5 | 10 | DT | W | NB | 120 | 24 | 6 | 0 | 3 |
| C4(6.1) | 6 | 12 | DT | U | NB | 48 | 8 | 2 | 1 | 3 |
| C4(8.1) | 8 | 16 | DT | U | B| 2^7(3^7) | 144 | 36 | 2 | 4 |
| C4(9.1) | 9 | 18 | DT | W | NB | 72 | 8 | 2 | 1 | 3 |
| C4(10.1) | 10 | 20 | DT | U | NB | 320 | 32 | 8 | 1 | 4 |
| C4(10.2) | 10 | 20 | DT | W | NB | 240 | 24 | 6 | 1 | 4 |
| C4(12.1) | 12 | 24 | DT | U | B| 768 | 64 | 16 | 3 | 4 |
| C4(12.2) | 12 | 24 | DT | W | NB | 48 | 4 | 1 | 2 | 3 |
| C4(13.1) | 13 | 26 | DT | W | NB | 52 | 4 | 1 | 0 | 4 |
| C4(14.1) | 14 | 28 | DT | U | B| 2^8(7^1) | 128 | 32 | 1 | 4 |
| C4(14.2) | 14 | 28 | DT | W | B| 336 | 24 | 6 | 0 | 4 |
| C4(15.1) | 15 | 30 | DT | W | NB | 60 | 4 | 1 | 2 | 4 |
| C4(15.2) | 15 | 30 | DT | W | NB | 120 | 8 | 2 | 0 | 3 |
| C4(16.1) | 16 | 32 | DT | W | B| 2^12(2) | 256 | 64 | 3 | 4 |
| C4(16.2) | 16 | 32 | DT | W | B| 384 | 24 | 6 | 2 | 4 |
| C4(17.1) | 17 | 34 | DT | W | NB | 68 | 4 | 1 | 0 | 4 |
| C4(18.1) | 18 | 36 | DT | U | B| 2^10(3^2) | 512 | 128 | 1 | 4 |
| C4(18.2) | 18 | 36 | DT | W | B| 144 | 8 | 2 | 2 | 4 |
| C4(20.1) | 20 | 40 | DT | U | B| 2^12(5^1) | 2^10 | 256 | 3 | 4 |
| C4(20.2) | 20 | 40 | DT | W | B| 80 | 4 | 1 | 1 | 4 |
| C4(20.3) | 20 | 40 | DT | W | NB | 320 | 16 | 4 | 1 | 4 |
| C4(20.4) | 20 | 40 | SS | U | B| 2^8(3^1)(5^1) | 384 | 96 | 0 | 4 |
| C4(21.1) | 21 | 42 | DT | W | NB | 84 | 4 | 1 | 2 | 4 |
| C4(21.2) | 21 | 42 | DT | W | NB | 336 | 16 | 4 | 2 | 3 |
MathHub Data in a Nutshell

▶ MathHub Data:

- Schema Theory
- Codecs

your data
CSV or JSON

Online Database
https://data.mathhub.info

Community Resource: MitM and Codecs,
Dataset: data in JSON, provenance, and schema theory.
Definition 3.3 (Codec). A codec consists of two functions that translate between semantic types and realized types.

### Codecs

<table>
<thead>
<tr>
<th>codec</th>
<th>type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>codec : type → type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StandardPos : codec ( \mathbb{Z}^+ )</td>
<td></td>
<td>JSON number if small enough, else JSON string of decimal expansion</td>
</tr>
<tr>
<td>StandardNat : codec ( \mathbb{N} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StandardInt : codec ( \mathbb{Z} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IntAsArray : codec ( \mathbb{Z} )</td>
<td></td>
<td>JSON List of Numbers</td>
</tr>
<tr>
<td>IntAsString : codec ( \mathbb{Z} )</td>
<td></td>
<td>JSON String of decimal expansion</td>
</tr>
<tr>
<td>StandardBool : codec ( \mathbb{B} )</td>
<td></td>
<td>JSON Booleans</td>
</tr>
<tr>
<td>BoolAsInt : codec ( \mathbb{B} )</td>
<td></td>
<td>JSON Numbers 0 or 1</td>
</tr>
<tr>
<td>StandardString : codec ( \mathcal{S} )</td>
<td></td>
<td>JSON Strings</td>
</tr>
</tbody>
</table>

StandardInt decodes 1 into the float 1, but \( 2^{54} \) into the string "18014398509481984".
Elliptic Curve Code Operators

```json
{
  "degree": 1,
  "x-coordinates of integral points": "[5,16]",
  "isogeny matrix": [[1,5,25],[5,1,5],[25,5,1]],
  "label": "11a1",
  "_id": "ObjectId('4f71d4304d47869291435e6e')",
  ...
}
```

Matrix in the isogeny_matrix field

```
    1  5  25
   -
    5  1  5
   -
   25 5  1
```

represented as [[1,5,25],[5,1,5],[25,5,1]]
Our approach: Virtual Theories

<table>
<thead>
<tr>
<th>Numbers</th>
<th>Matrices</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{Z}^+$ : type</td>
<td>$\text{matrix} : \text{type} \rightarrow \mathbb{Z}^+ \rightarrow \mathbb{Z}^+ \rightarrow \text{type}$</td>
</tr>
<tr>
<td>$\mathbb{Z}$ : type</td>
<td></td>
</tr>
<tr>
<td>$\mathbb{Z}^+ \subset \mathbb{Z}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codecs</th>
</tr>
</thead>
<tbody>
<tr>
<td>codec : type $\rightarrow$ type</td>
</tr>
<tr>
<td>standardInt : codec $\mathbb{Z}$</td>
</tr>
<tr>
<td>standardMatrix : ${T, n, m}$ codec $T \rightarrow$ codec matrix$(n, m, T)$</td>
</tr>
</tbody>
</table>

**Elliptic Curve Database Theory**

<table>
<thead>
<tr>
<th>Elliptic Curve Database Theory</th>
<th>Elliptic Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>11a1 : ec = ...</td>
<td>ec : type</td>
</tr>
<tr>
<td>11a2 : ec = ...</td>
<td>from_record : record $\rightarrow$ ec</td>
</tr>
<tr>
<td>...</td>
<td>curveDegree : ec $\rightarrow$ $\mathbb{Z}$</td>
</tr>
<tr>
<td></td>
<td>isogenyMatrix : ec $\rightarrow$ matrix$(3, 3, \mathbb{Z})$</td>
</tr>
</tbody>
</table>

**Elliptic Curve Schema Theory**

<table>
<thead>
<tr>
<th>Elliptic Curve Schema Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>degree : ?implements curveDegree</td>
</tr>
<tr>
<td>?codec : StandardInt</td>
</tr>
<tr>
<td>isogeny_matrix : ?implements isogenyMatrix</td>
</tr>
<tr>
<td>?codec : StandardMatrix$(3, 3, \text{StandardInt})$</td>
</tr>
</tbody>
</table>

lmfdb Elliptic Curves lazily loads from degree implements

Kohlhase: ODK WP6 11 ODK Final Review, Oct. '19
First working prototype since August 2019 (https://data.mathhub.info) (more in the pipeline)

Six datasets provided by the community
- Graphs, Maniplexes, Polyhedra, Additive Bases, small Groups...
- together $\sim 13M$ Math Objects, 10 to 20 properties per object

Mathematical variety sufficient to validate the system design.

Wow: The DB researchers are very interested in the DB aspects (complex objects)

Combinatorics community is very interested (Math Data WS $\sim 2020$)

Future: Scaling, Services, Community Building
- Dataset submission process (metadata, descriptions, provenance, ...)
- Working towards a “Journal of Mathematical Data” based on MHD
- Semantic internal references via views.
Come to the MathHub Data Demo

A census of small connected cubic vertex-transitive graphs
All connected cubic vertex-transitive graphs of order at most 1280.

This dataset has 111760 objects.

Matches found: 164

More about this dataset

Display results

Choose columns

<table>
<thead>
<tr>
<th>Order</th>
<th>CVT Index</th>
<th>Graph</th>
<th>Name</th>
<th>Clique Number</th>
<th>Diameter</th>
<th>Girth</th>
<th>Is Arc-Transitive</th>
<th>Is Bipartite</th>
<th>Is Cayley</th>
<th>Is Distance Regular</th>
<th>Is Distance Transitive</th>
<th>Is Edge-Transitive</th>
<th>Is Hamiltonian</th>
<th>Is Partial Cube</th>
</tr>
</thead>
</table>
4 Persistent Memoization
What is memoisation?
- Store program results in a permanent cache when they are computed.
- Retrieve these results from the cache later instead of recomputing.
- Cache can be local or online.

**Example 4.1 (Persistent Memoization in GAP/python).**
Persistent Memoization in Python and GAP

- Calling program
- Memoised function (wrapper)
  - New result
  - Known result
  - Cache
    - Default
    - If URL specified
      - Local file system
      - HTTP interface
        - MongoDB database
Advantages

- Avoids re-running programs that are guaranteed to return the same answer
- Allows us to create an archive of results that can be used for other purposes
- Share results between users, locations, and even programming languages
Persistent Memoization in Python and GAP

Advantages

- Avoids re-running programs that are guaranteed to return the same answer
- Allows us to create an archive of results that can be used for other purposes
- Share results between users, locations, and even programming languages
5 Recommendations, Deliverables, KPIs, Lessons
Recommendations

**Recommendation 7.** *To develop a comic explaining the MitM approach.*

- The comic has been published on: https://github.com/OpenDreamKit/OpenDreamKit.github.io/blob/master/public/images/use-cases/MitM.png.
- It has already been used in the MitM use case description at https://opendreamkit.org/2018/05/16/lfmfd-b-usecase/, in conference presentations and posters.

**Recommendation 8.** *To disseminate the Adoption by Logipedia of the MitM principle of integrating (logical) systems by aligning concepts.*

- We have made a blog post about this, see https://opendreamkit.org/2019/01/24/logipedia/
Deliverables in WP6

All Deliverables were delivered (mostly on time)

<table>
<thead>
<tr>
<th>Task</th>
<th>Name</th>
<th>RP1</th>
<th>RP2</th>
<th>RP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T6.1</td>
<td>Search</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T6.2</td>
<td>Survey</td>
<td></td>
<td>D6.1</td>
<td></td>
</tr>
<tr>
<td>T6.3</td>
<td>DKS-Design</td>
<td></td>
<td>D6.2</td>
<td>D6.3</td>
</tr>
<tr>
<td>T6.4-8</td>
<td>Case Studies</td>
<td></td>
<td>D6.5</td>
<td>D6.8</td>
</tr>
<tr>
<td>T6.9</td>
<td>Memoization</td>
<td></td>
<td></td>
<td>D6.9</td>
</tr>
<tr>
<td>T6.10</td>
<td>Math Search</td>
<td></td>
<td></td>
<td>D6.10</td>
</tr>
<tr>
<td>T6.11</td>
<td>Isabelle Lib</td>
<td></td>
<td></td>
<td>D6.11</td>
</tr>
</tbody>
</table>
KPIs and Deliverables for WP6

- The Math-in-the-Middle Ontology (largely unchanged from last time)
  - MitM-connected Systems: four (GAP, Sage, LMFDB, Singular) (See D6.5)
  - Formal MitM Ontology: 60 files, 3000 LoF, 500 commits (See D6.8)
  - Informal MitM Ontology: 900 theories, 1900 concepts in English, German, (Chinese, Romanian)
  - MitM System API Theories (GAP, Sage, LMFDB, Singular): 1.000+ Theories, 22.000 Concepts.
  - Isabelle Library: > $10^5$ lemmas, > $10^6$ loc
  - Heavy interest by the theorem proving community about MitM Ontology
  - Logipedia (http://logipedia.science) adopts the MitM principle of integrating (logical) systems by aligning concepts.
KPIs and Deliverables for WP6

- **The Math-in-the-Middle Ontology** (largely unchanged from last time)
  - MitM-connected Systems: four (GAP, Sage, LMFDB, Singular) (See D6.5)
  - Formal MitM Ontology: 60 files, 3000 LoF, 500 commits (See D6.8)
  - Informal MitM Ontology: 900 theories, 1900 concepts in English, German, (Chinese, Romanian)
  - MitM System API Theories (GAP, Sage, LMFDB, Singular): 1.000+ Theories, 22.000 Concepts.
  - Isabelle Library: $>10^5$ lemmas, $>10^6$ loc
  - Heavy interest by the theorem proving community about MitM Ontology
  - Logipedia (http://logipedia.science) adopts the MitM principle of integrating (logical) systems by aligning concepts.

- **MathHub** (Front-End re-implemented 2018/19)
  - 327 archives $\sim 61\, GB$; 25 archives in web UI; $\sim 2.5\, GB$
KPIs and Deliverables for WP6

▶ The Math-in-the-Middle Ontology (largely unchanged from last time)
  ▶ MitM-connected Systems: four (GAP, Sage, LMFDB, Singular) (See D6.5)
  ▶ Formal MitM Ontology: 60 files, 3000 LoF, 500 commits (See D6.8)
  ▶ Informal MitM Ontology: 900 theories, 1900 concepts in English, German, (Chinese, Romanian)
  ▶ MitM System API Theories (GAP, Sage, LMFDB, Singular): 1.000+ Theories, 22,000 Concepts.
  ▶ Isabelle Library: $> 10^5$ lemmas, $> 10^6$ loc
  ▶ Heavy interest by the theorem proving community about MitM Ontology
  ▶ Logipedia (http://logipedia.science) adopts the MitM principle of integrating (logical) systems by aligning concepts.

▶ MathHub (Front-End re-implemented 2018/19)
  ▶ 327 archives $\sim 61\text{GB}$; 25 archives in web UI; $\sim 2.5\text{GB}$

▶ MathHub Data (new, since August 2019)
  ▶ $12M$ Math Objects with $\sim 15$ properties, $\sim 80\text{GB}$ in DB.
  ▶ 4/6 data sets provided externally (four groups/researchers).
Come to the MathWebSearch Data (n-Category Lab)

\[ Y: C \to [C^{op}, Set] \]

**totally distributive category in nLab (1)**

**Yoneda embedding in nLab (1)**

**presheaf in nLab (2)**

via the Yoneda embedding \( Y: C \to [C^{op}, Set] \) The Yoneda embedding sends each object

Substitutions: \( C: C \quad Y: Y \)

view in nLab

Yoneda reduction. See also co-Yoneda lemma. More concretely: let \( Y: C \to [C^{op}, Set] \) denote the Yoneda embedding of an object \( c \in C \)
Lessons Learnt: WP6 (Data/Knowledge/Software)-Bases

- **Generally**: OpenDreamKit was a tremendous opportunity to rethink Math Software Infrastructure
  - Freedom to think/conceptualize/prototype/evaluate/scale for 4 years
  - A common non-trivial infrastructure goal to evaluate (VRE toolkit)
Lessons Learnt: WP6 (Data/Knowledge/Software)-Bases

- **Generally**: OpenDreamKit was a tremendous opportunity to rethink Math Software Infrastructure
  - Freedom to think/conceptualize/prototype/evaluate/scale for 4 years
  - A common non-trivial infrastructure goal to evaluate (VRE toolkit)

- **Problems Encountered**: with taking the high road for system integration
  - MitM takes a large initial investment per system (**Framework + mediator exist now**) (but compilation possible)
  - mediator-based translation is relatively slow
  - correct/complete translations are possible by traditional programming (**by trained mathematicians**) (by trained mathematicians)
  - break-even point seems near 4 systems connected (**$n^2 - n$ vs. $2n$ translations**)
Lessons Learnt: WP6 (Data/Knowledge/Software)-Bases

- **Generally:** OpenDreamKit was a tremendous opportunity to rethink Math Software Infrastructure
  - Freedom to think/conceptualize/prototype/evaluateSCALE for 4 years
  - A common non-trivial infrastructure goal to evaluate (VRE toolkit)

- **Problems Encountered:** with taking the high road for system integration
  - MitM takes a large initial investment per system (Framework + mediator exist now) (but compilation possible)
  - mediator-based translation is relatively slow (by trained mathematicians)
  - correct/complete translations are possible by traditional programming
  - break-even point seems near 4 systems connected ($n^2 - n$ vs. $2n$ translations)

- **Fewer Problems encountered:** for semantic mathematical data
  - semantic description of the dataset is a reasonable investment (Schema theories + JSON + Provenance)
  - BUT author gets a turnkey solution for their data sets! (first digitization)
  - AND the dataset is MitM-enabled. (both intra-MDH and with CAS)
OpenDreamKit Follow-Up Proposal: FAIRMat

- **Call**: European Research Infrastructures: Implementing the European Open Science Cloud  
  *(Deadline 29.1.2019)*

- **FAIRMat**: FAIR Mathematical Data for the European Open Science Cloud
  - FAU Erlangen-Nürnberg (coordinator)
  - Univesité Paris Sud
  - Chalmers University of Technology
  - Univerza v Ljubljani
  - CAE Tech Limited
  - FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
  - European Mathematical Society

- **Work Areas:**
  - **WP2**: Standardized data representation framework *(deep FAIR)*
  - **WP3**: Mathematical Services for the EOSC *(e.g. search, programmatic APIs)*
  - **WP4**: Data Sets for EOSC *(Combinatorics, Algebra, Modelling)*
  - **WP5**: Community Building

- **Result**: Cleared eligibility threshold well, not funded *(too disciplinary)*
Conclusion: What are we doing in WP6 in terms of a VRE

- SageMath and WP6 approach (Math-in-the-Middle; MitM) are both attempts at making a VRE Toolkit.
- SageMath is very successful, because integration is lightweight:
  - It makes no assumption on the meaning of math objects exchanged.
  - Restricts itself to master-slave integration of systems into SageMath. But there are safety, extensibility, and flexibility issues!
Conclusion: What are we doing in WP6 in terms of a VRE

- SageMath and WP6 approach (Math-in-the-Middle; MitM) are both attempts at making a VRE Toolkit.
  - SageMath is **very successful**, because integration is lightweight:
    - It makes no assumption on the meaning of math objects exchanged.
    - Restricts itself to master-slave integration of systems into SageMath.
  - But there are safety, extensibility, and flexibility issues!

- MitM tries to **take the high road** (make possible by OpenDreamKit)
  - **Safety**: by semantic (i.e. context-aware) objects passed.
  - **Extensibility**: any open-API system (i.e. with API CDs) can play.
  - **Flexibility**: full peer-to-peer possibilities. (future: service discovery)
  - But we have to develop a whole new framework! (prototyped it)
Conclusion: What are we doing in WP6 in terms of a VRE

- SageMath and WP6 approach (Math-in-the-Middle; MitM) are both attempts at making a VRE Toolkit.
- SageMath is very successful, because integration is lightweight:
  - It makes no assumption on the meaning of math objects exchanged.
  - Restricts itself to master-slave integration of systems into SageMath.
  But there are safety, extensibility, and flexibility issues!
- MitM tries to take the high road (make possible by OpenDreamKit)
  - Safety: by semantic (i.e. context-aware) objects passed.
  - Extensibility: any open-API system (i.e. with API CDs) can play.
  - Flexibility: full peer-to-peer possibilities. (future: service discovery)
  But we have to develop a whole new framework! (prototyped it)

- Review Period3: Inference & Math Data

- integrated Isabelle Library into MitM
- Semanticizing LMFDB
- Persistent Memoization
- MathHub Data \(\leadsto\) FAIR