FAIRMath: Making Mathematical Data FAIR (Findable, Accessible, Interoperable, Reusable)

Joint work with: Florian Rabe, Dennis Müller, Mihnea Iancu, Katja Bercic, Tom Wiesing...

Michael Kohlhase

Computer Science, FAU Erlangen-Nürnberg

May 30. 2018, Big Proof Workshop, ICMS Edinburgh

Take-Home Message (I will probably run out of time)

- We are all interested in "Big Math", not only "Big Proof"

Take-Home Message (I will probably run out of time)

- We are all interested in “Big Math”, not only “Big Proof”
- We propose a tetrapodal model for “doing/supporting” mathematics
- Mathematical Research Data is a next big thing (FAIR principles)
- Math Data wants to be deep FAIR (accessible semantics crucial)

Take-Home Message (I will probably run out of time)

- We are all interested in "Big Math", not only "Big Proof"
- We propose a tetrapodal model for "doing/supporting" mathematics
- Mathematical Research Data is a next big thing (FAIR principles)
- Math Data wants to be deep FAIR (accessible semantics crucial)
- First steps towards deep FAIR infrastructures/hosting (MathDataHub)

Take-Home Message (I will probably run out of time)

- We are all interested in "Big Math", not only "Big Proof"
- We propose a tetrapodal model for "doing/supporting" mathematics
- Mathematical Research Data is a next big thing (FAIR principles)
- Math Data wants to be deep FAIR (accessible semantics crucial)
- First steps towards deep FAIR infrastructures/hosting (MathDataHub)
- Future: (would be happy to collaborate with you all)
  - get funding for deep FAIR math data, (EOSC proposal FAIRMath rejected)
  - : stabilize MathDataHub, collect data sets and services,
  - extend these ideas to other sciences (the STEM disciplines)

1 Big Math and the One Brain Barrier

Background: Towards Big Math; Details in [Car+19]

- ***Observation 1.1.*** In the last half decade mathematics tackles problems that lead to increasingly large developments: proofs, computations, data sets, and document collections.
- Consequence: Intense discussions about the nature of mathematics
  1. Is a proof that can only be verified with the help of a computer still a mathematical proof? (Appel&Haken '76)
  2. Is a mathematical proofscape that exceeds what can be understood in detail by a single expert a legitimate justification of a mathematical result? (CFSG)
  3. Can a collection of mathematics papers — however big — adequately represent a large body of mathematical knowledge? (DML)
- **Definition 1.2.** Let us call such large developments Big Math– and the (uncontroversial) rest Pen Math.

E.g.: The Classification of Finite Simple Groups (CFSG)

- The CFSG is one of the seminal results of 20$^{\text{th}}$ century mathematics.
- Its status is similar to that of the fundamental theorem of arithmetic.
- Its proof was constructed by a large community over ≥ 50 years (last special cases completed in 2004)
- The CFSG proof spans ≥ 100 articles ($\geq 10,000$ pp)
- Goal since 1985: Condense to "Second-Generation Proof" (~ 5000 pp)

E.g.: The Classification of Finite Simple Groups (CFSG)

- The CFSG is one of the seminal results of 20^{th} century mathematics.
- Its status is similar to that of the fundamental theorem of arithmetic.
- Its proof was constructed by a large community over ≥ 50 years (last special cases completed in 2004)
- The CFSG proof spans ≥ 100 articles ($\geq 10,000$ pp)
- Goal since 1985: Condense to "Second-Generation Proof" (~ 5000 pp)
- ***Observation 1.3.*** The traditional way of "doing Math" $\widehat{=}$
  - well-trained, highly creative individuals deriving insights with "pen and paper",
  - report on them in community meetings, and publishing them in academic journals or monographs.

  is reaching its natural limits posed by the amount of mathematical knowledge that can be held in a single human brain $\rightsquigarrow$ the one-brain barrier (OBB)

E.g.: The Classification of Finite Simple Groups (CFSG)

- The CFSG is one of the seminal results of 20^{th} century mathematics.
- Its status is similar to that of the fundamental theorem of arithmetic.
- Its proof was constructed by a large community over ≥ 50 years (last special cases completed in 2004)
- The CFSG proof spans ≥ 100 articles ($\geq 10,000$ pp)
- Goal since 1985: Condense to "Second-Generation Proof" (~ 5000 pp)
- ***Observation 1.3.*** The traditional way of "doing Math" $\widehat{=}$
  - well-trained, highly creative individuals deriving insights with "pen and paper",
  - report on them in community meetings, and publishing them in academic journals or monographs.

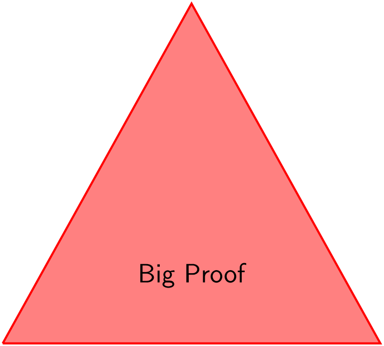
  is reaching its natural limits posed by the amount of mathematical knowledge that can be held in a single human brain $\rightsquigarrow$ the one-brain barrier (OBB)
- OBB can be generalized to small-group-brain barrier (but mind "The Mythical Man Month")

Classifying Math wrt. the One-Brain-Barrier

- A classification of mathematical developments.
  - Pen math: the developments and results that can be obtained by pen, paper, and university library by an individual or small group without too much strain on the process.
  - Big math: all that is beyond small math (i.e. straining the classical math process), but still inside the OBB
  - Trans-OBB math: all that is beyond bigmath but still amenable to the "method of proof".
  - Inaccessible math: all results are are unprovable because of Gödel's incompleteness.
- The agenga of "Big Proofs/Big Math" must be to enabling big/trans-OBB math (leave inaccessiblemath alone)

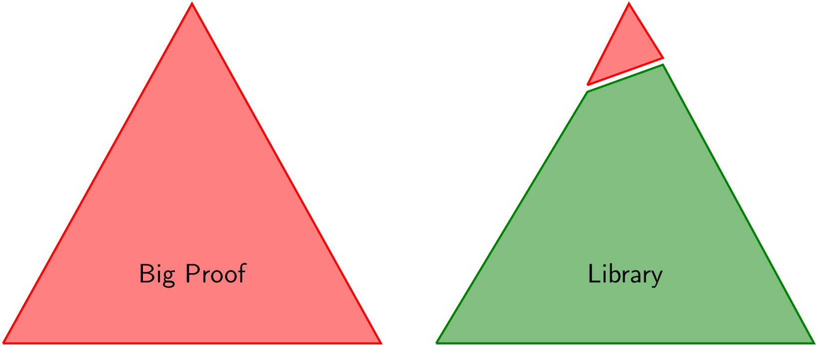
Big Proof $\widehat{=}$ Big Libraries + Little Proofs

- Proofs from first principles are prohibitively BIG.



Big Proof $\widehat{=}$ Big Libraries + Little Proofs

- Proofs from first principles are prohibitively BIG.



Good Practice: Explore theories, prove intermediate results, bild mathematical components/tools. (Digital Mathematical Libraries)

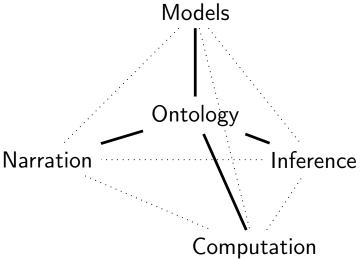
- libraries are a method to achieve big knowledge via little proof!

Knowledge Representation is only Part of "Doing Math"

- One of the key insights is that the mathematics ecosystem involves a body of knowledge described as an ontology and four aspects of it:
  - inference: exploring theories, formulating conjectures, and constructing proofs
  - computation: simplifying mathematical objects, re-contextualizing conjectures. . .
  - models: collecting examples, applying mathematical knowledge to real-world problems and situations.
  - narration: devising both informal and formal languages for expressing mathematical ideas, visualizing mathematical data, presenting mathematical developments, organizing and interconnecting mathematical knowledge

"Doing Math": as a Tetrapod

- We call the endeavour of creating a computer-supported mathematical ecosystem "Project tetrapod" as it needs to stand on four legs.



Collaborators: KWARC@FAU, McMaster University

The Tetrapod in the Big Proofs Workshop

- There are many single/dual-aspect systems, ...
- some are mentioned here at the big proofs workshop
  - ontology: e.g. Paulson – see Ma, what I can inherit
  - inference: everyone of course
  - computation: e.g. Avigad – verification of computation,
  - narration: e.g. Hales/Köpke – CNL for human-readable FAbstracts
  - tabulation: Douglas – DBs in Physics

  Motivation 1 the tabulation (mathematical datasets/databases) aspect is the least represented here.

2 Motivation 2: Mathematical Research Data

Research Data: A general Next Big Thing

- **Definition 2.1.** 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)

Research Data: A general Next Big Thing

- **Definition 2.1.** 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 2.2 (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.

Research Data: A general Next Big Thing

- **Definition 2.1.** 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 2.2 (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 be make mathematical data fair?

German National Research Data Initiative

- NFDI: In November 2018 the federal/state governments agreed to establishment of a national initiative for research data.
- Funding: 900 M€ over 10 years, afterwards institutional funding? (research and competence/service, no hardware)
- Format: ca. 30 Consortia who will form independent organizations.
- Math4NFDI: A consortium for Mathematical Research Data, (Lead: WIAS Berlin)
- Current State of Play: Networking, consortium consolitdation, first NFDI call imminent.

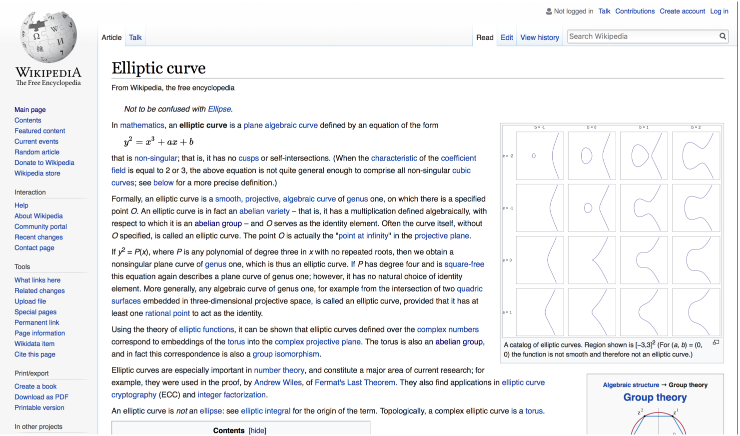
The European Open Science Cloud

- EU Vision: The EOSC will provide 1.7m EU researchers an environment with free, open services for data storage, management, analysis and re-use across disciplines.
- Planned Architecture: Federated meta-archive building on existing infrastrutures: CERN, EMBL, ELIXIR, etc.
- Chicken/Egg Problem: how to get the EOSC off the ground?
  - There is only one mathematical data set on the EOSC (Jukka Kohonen's lattices)
- Current State of Play: some EOSC calls for implementation, data sets, services
- Proposal FAIRMath: from Jan 2019 was unsuccessful, disciplinary proposals apparently not appreciated.
- But: the FAIRMath proposal led to a clarification – for us – of Math Research Data

3 Mathematical Knowledge/Data Bases; State of the Art

Mathematical Knowledge Sources (MKS)

- generic information systems (Wikipedia)


Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

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

Print/export
Create a book
Download as PDF
Printable version

In other projects

Elliptic curve

From Wikipedia, the free encyclopedia

Not to be confused with Ellipse.

In mathematics, an **elliptic curve** is a plane algebraic curve defined by an equation of the form

$$y^2 = x^3 + ax + b$$

that is non-singular; that is, it has no cusps or self-intersections. (When the characteristic of the coefficient field is equal to 2 or 3, the above equation is not quite general enough to comprise all non-singular cubic curves; see below for a more precise definition.)

Formally, an elliptic curve is a smooth, projective, algebraic curve of genus one, on which there is a specified point *O*. An elliptic curve is in fact an abelian variety – that is, it has a multiplication defined algebraically, with respect to which it is an abelian group – and *O* serves as the identity element. Often the curve itself, without *O* specified, is called an elliptic curve. The point *O* is actually the "point at infinity" in the projective plane.

If $y^2 = P(x)$, where *P* is any polynomial of degree three in *x* with no repeated roots, then we obtain a nonsingular plane curve of genus one, which is thus an elliptic curve. If *P* has degree four and is square-free this equation again describes a plane curve of genus one; however, it has no natural choice of identity element. More generally, any algebraic curve of genus one, for example from the intersection of two quadric surfaces embedded in three-dimensional projective space, is called an elliptic curve, provided that it has at least one rational point to act as the identity.

Using the theory of elliptic functions, it can be shown that elliptic curves defined over the complex numbers correspond to embeddings of the torus into the complex projective plane. The torus is also an abelian group, and in fact this correspondence is also a group isomorphism.

Elliptic curves are especially important in number theory, and constitute a major area of current research; for example, they were used in the proof, by Andrew Wiles, of Fermat's Last Theorem. They also find applications in elliptic curve cryptography (ECC) and integer factorization.

An elliptic curve is *not* an ellipse: see elliptic integral for the origin of the term. Topologically, a complex elliptic curve is a torus.

Contents [hide]

b = -1 b = 0 b = 1 b = 2
a = -2 a = -1 a = 0 a = 1

A catalog of elliptic curves. Region shown is $[-3,3]^2$ (For $(a, b) = (0, 0)$ the function is not smooth and therefore not an elliptic curve.)

Algebraic structure → Group theory
Group theory

Mathematical Knowledge Sources (MKS)

- generic information systems (Wikipedia)
- informal mathematical document collections (Cornell preprint arXiv)

arXiv.org > math > arXiv:1711.02170

Search or Article ID | All papers

(Help | Advanced search)

Mathematics > Number Theory

On Elliptic Curves of prime power conductor over imaginary quadratic fields with class number one

John Cremona, Ariel Pacetti

(Submitted on 6 Nov 2017)

The main result of this paper is to generalize from $\Q$ to each of the nine imaginary quadratic fields of class number one a result of Serre and Mestre-Oesterl\'e of 1989, namely that if E is an elliptic curve of prime conductor then either E or a 2-isogenous curve or a 3-isogenous curve has prime discriminant. The proof is conditional in two ways: first that the curves are modular, so are associated to suitable Bianchi newforms; and secondly that a certain level-lowering conjecture holds for Bianchi newforms. We also classify all elliptic curves of prime power conductor and non-trivial torsion over each of the nine fields: in the case of 2-torsion we find that such curves either have CM or with a small (finite) number of exceptions arise from a family analogous to the Setzer-Neumann family of elliptic curves over $\Q$.

Comments: 27 pages
Subjects: Number Theory (math.NT)
MSC classes: 11G05 (Primary), 14H52 (Secondary)
Cite as: arXiv:1711.02170 [math.NT]
(or arXiv:1711.02170v1 [math.NT] for this version)

Download:
- PDF
- PostScript
- Other formats

(license)

Current browse context:
math.NT
< prev | next >
new | recent | 1711

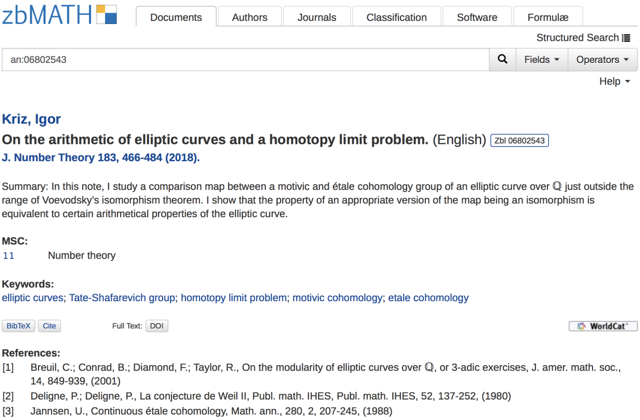
Change to browse by:
math

References & Citations
- NASA ADS

Bookmark (what is this?)

Mathematical Knowledge Sources (MKS)

- generic information systems (Wikipedia)
- informal mathematical document collections (Cornell preprint arXiv)
- literature information systems (zbMATH, MathSciNet)



zbMATH

Documents | Authors | Journals | Classification | Software | Formulæ

Structured Search

an:06802543

Fields | Operators

Help

Kriz, Igor

On the arithmetic of elliptic curves and a homotopy limit problem. (English) Zbl 06802543

J. Number Theory 183, 466-484 (2018).

Summary: In this note, I study a comparison map between a motivic and étale cohomology group of an elliptic curve over $\mathbb{Q}$ just outside the range of Voevodsky's isomorphism theorem. I show that the property of an appropriate version of the map being an isomorphism is equivalent to certain arithmetical properties of the elliptic curve.

MSC:

11 Number theory

Keywords:

elliptic curves; Tate-Shafarevich group; homotopy limit problem; motivic cohomology; etale cohomology

BibTeX | Cite | Full Text: DOI | WorldCat

References:

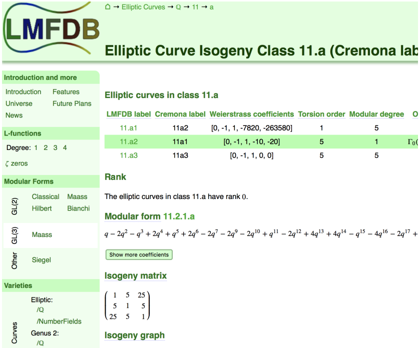
[1] Breuil, C.; Conrad, B.; Diamond, F.; Taylor, R., On the modularity of elliptic curves over $\mathbb{Q}$, or 3-adic exercises, J. amer. math. soc., 14, 849-939, (2001)

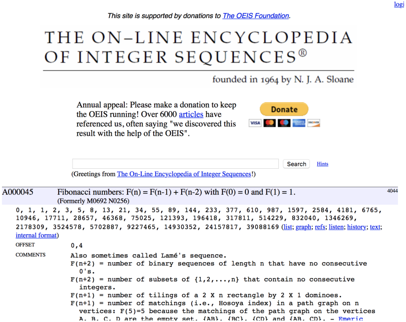
[2] Deligne, P.; Deligne, P., La conjecture de Weil II, Publ. math. IHES, Publ. math. IHES, 52, 137-252, (1980)

[3] Jannsen, U., Continuous étale cohomology, Math. ann., 280, 2, 207-245, (1988)

Mathematical Knowledge Sources (MKS)

- generic information systems (Wikipedia)
- informal mathematical document collections (Cornell preprint arXiv)
- literature information systems (zbMATH, MathSciNet)
- mathematical object databases (GAP libraries, OEIS, LMFDB)





Mathematical Knowledge Sources (MKS)

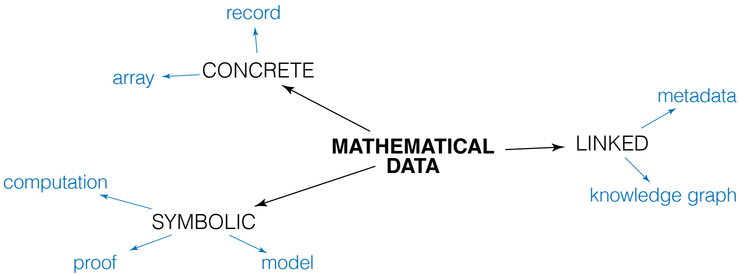
- generic information systems (Wikipedia)
- informal mathematical document collections (Cornell preprint arXiv)
- literature information systems (zbMATH, MathSciNet)
- mathematical object databases (GAP libraries, OEIS, LMFDB)
- formal theorem prover libraries (Mizar, Coq, PVS, HOL)
- We will concentrate on mathematical object databases here.

FAIRness in Mathematics

- Mathematical research is becoming more data-driven (datasets for conjecture induction/testing)
- But: there is no accepted paradigm for producing/working with data sets
- ***Observation 3.1.*** There is a strong open-source/open-data ethos in most of the mathematical community (see e.g. the IMU resolutions and IMKT)
- Consequence: Mathematics is (somewhat) FAIR on the surface (we try to do the right thing)
- But: deep problems remain, e.g. (deep $\widehat{=}$ hard, deep $\widehat{=}$ below surface)
  - accessible: math objects have more and more varied internal structure than e.g. satelite images
  - reusable: no copy/paste from GAP to Sage to Lean (different encodings)
  - interoperable: e.g. dihedral group of order 8 is called D_4 in Sage, but D_8 in GAP.
  - findable: there are attempts at structural math search engines,...
- ***Conjecture 3.2.*** For mathematics, we need deeply FAIR data practices $\leftrightsquigarrow$ math metadata are mathematical objects themselves.

Types of Mathematical Data

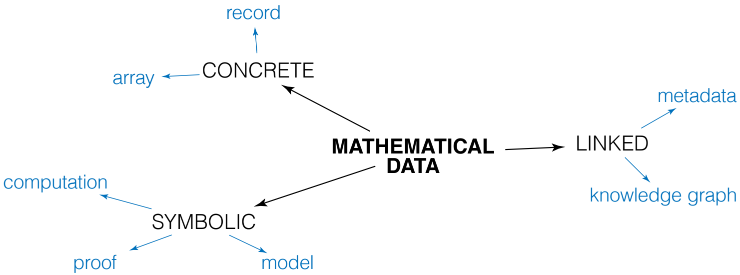
- We see three (or seven) kinds of mathematical data



- Symbolic Data can capture the full semantics of math objects by abstraction principles such as underspecification, quantification, and variable binding.
  $\leadsto$ context-sensitive: moving expressions across environments difficult
  $\leadsto$ F,I,R difficult (mitigate by standardization, e.g. MathML/OMDoc)

Types of Mathematical Data

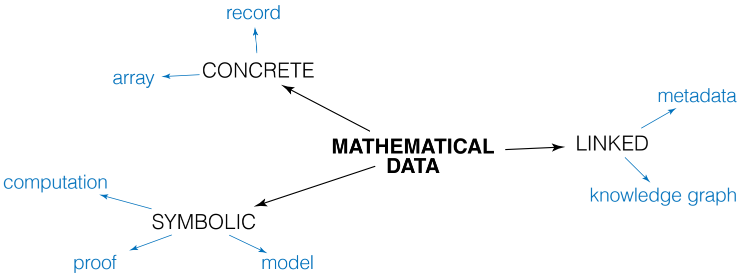
- We see three (or seven) kinds of mathematical data



- Concrete data employs representation theorems that allow encoding math. objects as simple data structures built from numbers, strings, lists, and records.
  $\rightsquigarrow$ Users have to know the repr. theorems to access data (often complex)
  $\rightsquigarrow$ FAIR difficult in practice (mitigate via documentation/Codecs)

Types of Mathematical Data

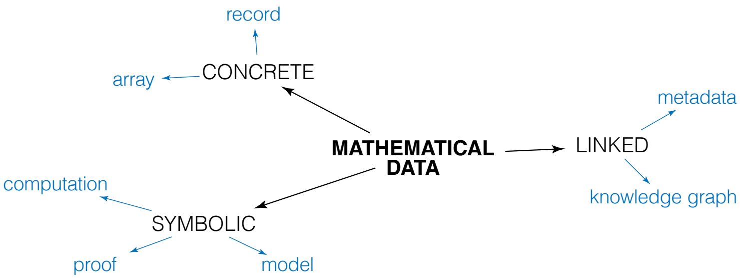
- We see three (or seven) kinds of mathematical data



- Linked data introduces identifiers for objects and then treats them as blackboxes, only representing the identifier and not the original object.
  $\leadsto$ semantics of represented mathematical objects is partial,
  $\leadsto$ F/A limited, I/R subject to misinterpretation (use only as directed)

Types of Mathematical Data

- We see three (or seven) kinds of mathematical data



Kind of data	Symbolic	Concrete	Linked
Allows recovering the represented object	+	+	−
Applicable to all objects	+	−	+
Easy to process	−	+	+

Programatic search in LMFDB

- Actual query:
  `http://www.lmfdb.org/api/transitivegroups/groups/?cyc=1`

- Desired query:

{x declared in 'lmfdb:db/transitivegroups?group | cyclic x ∗=∗ true }

Semantics-aware Open Data and Deep FAIRness

- Idea: We need to keep the semantics near the data (legends in tables)
- Current Practice: add informal labels, e.g. "weight in kg."
- ⚠: This works only where the semantics is very simple $\leadsto$ not in Math!
- **Example 3.3 (Often not even then).** In 2016 [**ZieEreElO:GeneErrors16**], researchers found widespread errors in papers in genomics journals with supplementary MS Excel gene lists. About 20% of them contain erroneous gene name because the software misinterpreted string-encoded genes as months.
- **Remark 3.4.** In engineering, encoding mistakes can quickly become safety-critical, i.e., if a dataset of numbers is shared without their physical units, precision, and measurement type.
- **Example 3.5 (The Mars Orbiter).** NASA specified thruster in SI units, Contractor built thruster using PSI
- **Definition 3.6.** We speak of accessible semantics if data has metadata annotations that allow recovering the exact semantics of the data.
- ***Observation 3.7.*** With accessible semantics, datasets can be validated automatically against their semantic type to avoid such errors

Semantics-aware Open Data and Deep FAIRness

- **Example 3.8.** We can reconstruct the (semantic) type *polynomial with integer coefficients* from its encoding *list of integers* only if its type and encoding function (*coefficients in order of decreasing degree*) are known.
  But: coefficient orders, sparse/dense, or multivariate polynomials.
- **Remark 3.9.** Without accessible semantics mathematical services can only operate on the dataset as a whole, we call them shallow FAIR services.
- **Definition 3.10.** We call a mathematical service deep FAIR, iff it operates on mathematical objects in a semantics-aware manner
- ***Observation 3.11.*** Meaningful mathematical services need to be deep FAIR.

Shallow/Deep Mathematical Services

- ***Observation 3.12.*** In our experience (in Math and elsewhere)
  - General data services are easy to build, iff they are shallow (general IT)
  - deep services are usually system-specific (where we have semantics)
- **Example 3.13.** shallow and deep FAIR services

Service	Shallow	Deep
Identification	DOI for a dataset	DOIs for each entry
Provenance	who created the dataset?	how was each entry computed?
Validation	is this valid XML?	does this XML represent a set of polynomials?
Access	download a dataset	download a specific fragment
Finding	find a dataset	find entries with certain properties
Reuse	impractical without accessible semantics	
Interoperability	impossible without accessible semantics	

Shallow/Deep Mathematical Services

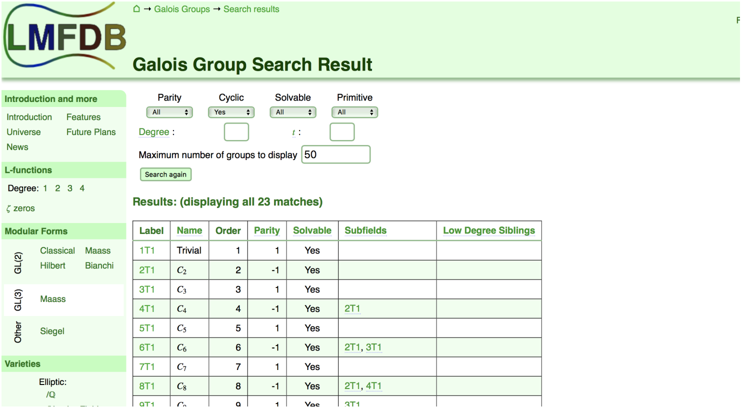
- ***Observation 3.12.*** In our experience (in Math and elsewhere)
  - General data services are easy to build, iff they are shallow (general IT)
  - deep services are usually system-specific (where we have semantics)
- **Remark 3.14 (Deep FAIR readiness of mathematical data).**

Data	Findable	Accessible	Interoperable	Reusable
Symbolic	Hard	Easy	Hard	Hard
Concrete	Impossible without access to the encoding function			
Linked	Easy, but only applicable to the small fragment with exposed semantics			

4 Deep/Shallow FAIR in practice

Searching in in the LMFDB

- Question: Find all cyclic transitive groups



⌂ → Galois Groups → Search results

LMFDB

Galois Group Search Result

Introduction and more
Introduction Features
Universe Future Plans
News

L-functions
Degree: 1 2 3 4
ζ zeros

Modular Forms
GL(2) Classical Maass
Hilbert Bianchi
GL(3) Maass
Other Siegel

Varieties
Elliptic:
/Q

Parity: All
Cyclic: Yes
Solvable: All
Primitive: All
Degree :
t :
Maximum number of groups to display 50
Search again

Results: (displaying all 23 matches)

Label	Name	Order	Parity	Solvable	Subfields	Low Degree Siblings
1T1	Trivial	1	1	Yes		
2T1	C_2	2	-1	Yes		
3T1	C_3	3	1	Yes		
4T1	C_4	4	-1	Yes	2T1	
5T1	C_5	5	1	Yes		
6T1	C_6	6	-1	Yes	2T1, 3T1	
7T1	C_7	7	1	Yes		
8T1	C_8	8	-1	Yes	2T1, 4T1	
9T1	C_9	9	1	Yes	3T1	

- Problem: But what if I want to compute with them?

Searching in OEIS

- **Question:** Find all sequences starting with $0, 1, 1, 2, 3, 5, 8$

0,1,1,2,3,5,8 [Search] Hints

(Greetings from The On-Line Encyclopedia of Integer Sequences!)

Search: **seq:0,1,1,2,3,5,8**

Displaying 1-10 of 124 results found. page 1 2 3 4 5 6 7 8 9 10 ... 13

Sort: relevance | references | number | modified | created Format: long | short | data

A000045 Fibonacci numbers: F(n) = F(n-1) + F(n-2) with F(0) = 0 and F(1) = 1. +20 4044

(Formerly M0692 N0256)

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181, 6765, 10946, 17711, 28657, 46368, 75025, 121393, 196418, 317811, 514229, 832040, 1346269, 2178309, 3524578, 5702887, 9227465, 14930352, 24157817, 39088169 (list; graph; refs; listen; history; text; internal format)

OFFSET 0,4

COMMENTS
```
Also sometimes called Lamé's sequence.
F(n+2) = number of binary sequences of length n that have no consecutive
  0's.
F(n+2) = number of subsets of {1,2,...,n} that contain no consecutive
  integers.
F(n+1) = number of tilings of a 2 X n rectangle by 2 X 1 dominoes.
F(n+1) = number of matchings (i.e., Hosoya index) in a path graph on n
  vertices: F(5)=5 because the matchings of the path graph on the vertices
  A, B, C, D are the empty set, {AB}, {BC}, {CD} and {AB, CD}. - Emeric
  Deutsch, Jun 18 2001
F(n) = number of compositions of n+1 with no part equal to 1. [Cayley,
  Grimaldi]
Positive terms are the solutions to z = 2*x*y^4 + (x^2)*y^3 - 2*(x^3)*y^2 -
  y^5 - (x^4)*y + 2*y for x,y >= 0 (Ribenboim, page 193). When x=F(n), y=F(n
  + 1) and z > 0 then z=F(n + 1).
For Fibonacci search see Knuth, Vol. 3; Horowitz and Sahni; etc.
```

Our Goal here

- Provide a Uniformal Interface to Mathematical Knowledge Bases
  - a mathematical, programatic API
- Idea: Use OMDoc/MMT to represent semantics
  - we can make use of theory graphs
  - we already have the Math-In-The-Middle approach
- use the MMT system
  - MMT terms represent semantic objects
  - has a built-in query language QMT

5 Virtual Theories

LMFDB Data (Database Level)

- **Example 5.1 (A transitive group represented in in LMFDB).**

```
{
  "ab": 1,
  "arith_equiv": 0,
  "auts": 1,
  "cyc": 1,
  "label": "1T1",
  "n": 1,
  ...
}
```

Legend: for understanding them (LMFDB improved documentation)

- the cyc field represents being cyclic (0 is **false**, 1 is **true**)
- the n field represents degree (IEEE Float 1 corresponds to $1 \in \mathbb{N}$)
- ...

Two Problems: that have to be solved for MitM integration

- data base schema is not at the mathematical level (let alone interoperable)
- values are encoded for MongoDB convenience (what do they mean?)

Codecs: Encoding and Decoding Database Values

- **Definition 5.2 (Codec).** A codec consists of two functions that translate between semantic types and realized types.

Codecs

codec : type → type	
StandardPos : codec $\mathbb{Z}^+$ StandardNat :codec $\mathbb{N}$ StandardInt :codec $\mathbb{Z}$	JSON number if small enough, else JSON string of decimal expansion
IntAsArray :codec $\mathbb{Z}$ IntAsString :codec $\mathbb{Z}$	JSON List of Numbers JSON String of decimal expansion
StandardBool :codec $\mathbb{B}$ BoolAsInt :codec $\mathbb{B}$	JSON Booleans JSON Numbers 0 or 1
StandardString :codec $\mathbb{S}$	JSON Strings

- StandardInt decodes 1 into the float 1, but 2^{54} into the string "18014398509481984"

Elliptic Curve Code Operators


```
{
  "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
  - $$\begin{bmatrix} 1 & 5 & 25 \\ 5 & 1 & 5 \\ 25 & 5 & 1 \end{bmatrix}$$
  - represented as [[1,5,25],[5,1,5],[25,5,1]]

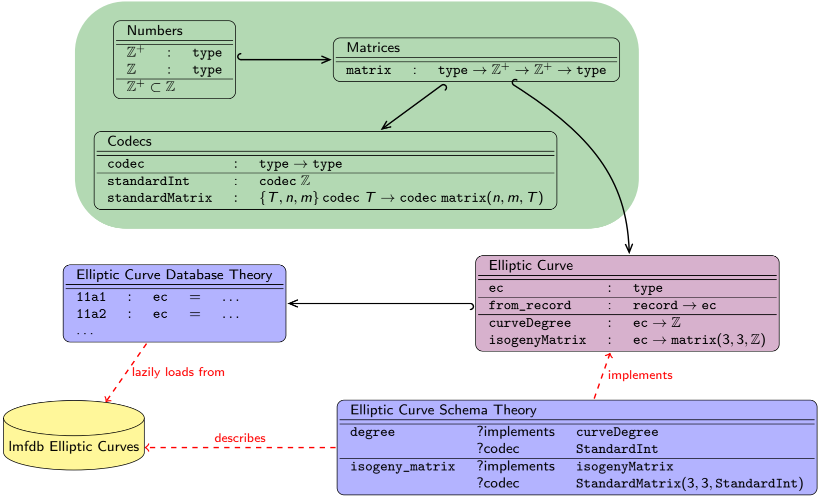
Codec Operator Examples

- **Definition 5.3 (Codec Operator).** A codec operator is a function which takes a codec, a set of parameters, and returns a codec.
- Codecs (continued)

StandardList : $\texttt{codec}\ T \to \texttt{codec}\ \mathrm{List}(T)$	JSON list, recursively coding each element of the list
StandardVector : $\texttt{codec}\ T \to \texttt{codec}\ \mathrm{Vector}(n, T)$	JSON list of fixed length n
StandardMatrix : $\texttt{codec}\ T \to \texttt{codec}\ \mathrm{Matrix}(n, m, T)$	JSON list of n lists of length m

- StandardMatrix(StandardInt, 3, 3) generates the codec we used for the isogeny matrix

Our approach: Virtual Theories



An Example of a Query

- **Example 5.4.** Finding all cyclic transitive groups in LMFDB (recall from above)

```
x in (related to ( literal ‘lmfdb:db/transitivegroups?group ) by (object declares))
| holds x (x cyclic x *=* true)
```

- This example does not rely on the internal structure of LMFDB
- can be translated into an LMFDB query using the just-defined codecs theory
- `http://www.lmfdb.org/api/transitivegroups/groups/?cyc=1`

6 MathDataHub: Hosting Math Datasets FAIRly

MathDataHub: Hosting Math Datasets FAIRly

- Problem: for all math data sets (see `http://mathwb.mathweb.org` for a list)
  MDH1 General data services are easy to build, iff shallow (general IT)
  MDH2 deep services are usually system-specific (where we have semantics)
  In particular, systems/datasets are motivated by **MDH2** and flounder for **MDH1**
- Idea: Supply a deep FAIR infrastructure (**MDH1**) so that authors can concentrate on **MDH2** $\leadsto$ MathDataHub. (share accessible semantics)
- Technically: Extend the virtual theories above to a Math Data Definition Language MDDL and generate MathDataHub infrastructure from that. (details in [**BerKohRab:tumdi19**])

A simple Running Example

- **Example 6.1 (Running Example).**
  - Joe has collected a set of integer matrices together with their trace, eigenvalues, and the Boolean property whether they are orthogonal for his Ph.D. thesis.
  - he develops a MDDL description and submits it to MathDataHub.
  - Jane, a collaborator of Joe's, is interested in characteristic polynomials of integer sequences.
  - she develops a MDDL extension of Joe's.

Joe's dataset				Jane's column
M	TrM)	Orthogonal	σ_M	$\det(\lambda I - M)$
$\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$	2	yes	2, 1	$\lambda^2 - 3\lambda + 2$
$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$	4	no	3, 1	$\lambda^2 - 4\lambda + 3$
$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$	0	yes	1, −1	$\lambda^2 - 1$

A Glimpse of MDDL

- The DbData theory (simplified)

```
theory DbData : ur:?PLF =
  db_tp : type ❙
  db_val : db_tp → type ❘# V 1 prec −5 ❙
  db_null : {a} V a ❙
  db_int, db_bool, db_string, db_uuid : db_tp ❙
  db_array : db_tp → db_tp ❙
  eq  : {a} V a → V a → V db_bool ❘# 1 = 2 ❙...
❚
```

A Glimpse of MDDL

- The DbData theory (simplified)
- An excerpt from the MathData theory: collections

```
vector : type → ℤ → type
  # vector 1 2 prec 10
empty : {a} vector a 0
single : {a} a → vector a 1
matrix : type → ℤ → ℤ → type
  = [a,m,n] vector (vector a m) n
option : type → type
some : {a} a → option a
none : {a} option a
getOrElse : {a} option a → a → a
```

A Glimpse of MDDL

- The DbData theory (simplified)
- An excerpt from the MathData theory: collections
- Joe's Schema Theory (simplified)

```
theory MatrixS : ?MDDL =
    mat: matrix ℤ 2 2 | meta ?Codecs?codec MatrixAsArray IntIdent |
        tag ?MDDL?opaque ‖
    trace : ℤ | meta ?Codecs?codec IntIdent ‖
    orthogonal: bool | meta ?Codecs?codec BoolIdent ‖
    eigenvalues : list ℤ | meta ?Codecs?codec ListAsArray IntIdent |
        tag ?MDDL?opaque ‖
|||
```

A Glimpse of MDDL

- The DbData theory (simplified)
- An excerpt from the MathData theory: collections
- Joe's Schema Theory (simplified)
- Joe runs MBGen on tyhis schema theory

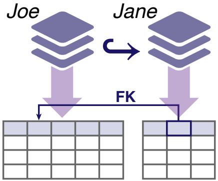
```
    Column       |    Type
---------------+----------
ID               | uuid
MAT              | integer[]
TRACE            | integer
ORTHOGONAL       | boolean
EIGENVALUES      | integer[]
Indexes: "MatrixS_pkey"
PRIMARY KEY, btree ("ID")
```

```
       ID          |     mat      | trace | orthogonal | eigenvalues
------------------+------------+-------+------------+-------------
e278b5e8-4404-... | {2,0,0,1}  |     2 | t          | {2,1}
05a30ff0-4405-... | {2,1,1,2}  |     4 | f          | {3,1}
1be3f022-4405-... | {-1,0,0,1} |     0 | t          | {1,-1}
```

A Glimpse of MDDL

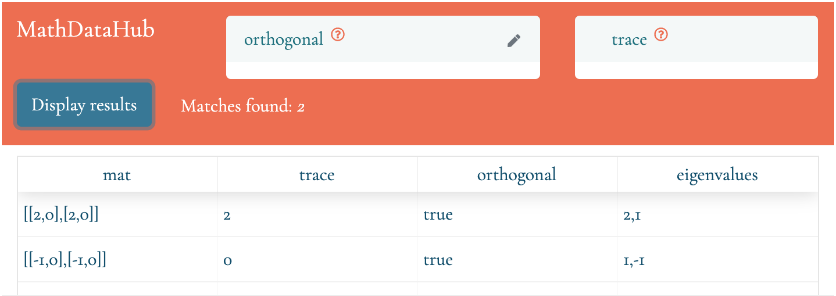
- The DbData theory (simplified)
- An excerpt from the MathData theory: collections
- Joe's Schema Theory (simplified)
- Joe runs MBGen on tyhis schema theory
- Jane's Extensions are compiled into a table referencing Joe's

```
theory MatrixWithCharacteristicS : ?SchemaLang =
  include ?MatrixS ❙
  matrixID: int | meta ?SchemaLang?foreignKey ?MatrixS ❙
  characteristic : Polynomial IntegerRing |
    meta ?Codecs?codec PolynomialAsSparseArray IntIdent ❙
❚
```



MathDataHub: Generating User Interfaces

- The MDDL specifications have annotations that allow to generate modulear UI code – here React.JS that interprets MMT-generated JSON.
- **Example 6.2 (Finishing the Running Example).**

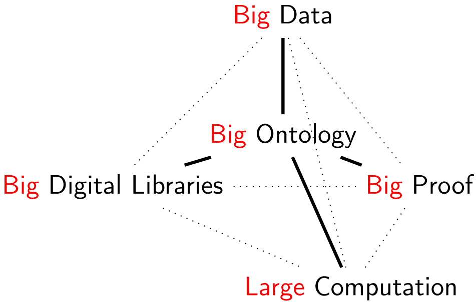


Conclusions & Future Work (if I managed to get here)

- We are all interested in “Big Math”, not only “Big Proof”

Conclusions & Future Work (if I managed to get here)

- We are all interested in "Big Math", not only "Big Proof"
- We propose a tetrapodal model for "doing/supporting" mathematics



- Mathematical Research Data is a next big thing (FAIR principles)
- Math Data wants to be deep FAIR (accessible semantics crucial)

Conclusions & Future Work (if I managed to get here)

- We are all interested in "Big Math", not only "Big Proof"
- We propose a tetrapodal model for "doing/supporting" mathematics
- Mathematical Research Data is a next big thing (FAIR principles)
- Math Data wants to be deep FAIR (accessible semantics crucial)
- First steps towards deep FAIR infrastructures/hosting (MathDataHub)

Conclusions & Future Work (if I managed to get here)

- We are all interested in "Big Math", not only "Big Proof"
- We propose a tetrapodal model for "doing/supporting" mathematics
- Mathematical Research Data is a next big thing (FAIR principles)
- Math Data wants to be deep FAIR (accessible semantics crucial)
- First steps towards deep FAIR infrastructures/hosting (MathDataHub)
- Future: (would be happy to collaborate with you all)
  - get funding for deep FAIR math data, (EOSC proposal FAIRMath rejected)
  - : stabilize MathDataHub, collect data sets and services,
  - extend these ideas to other sciences (the STEM disciplines)