

Collections of mathematical objects

Katja Berčič

University of Ljubljana and IMFM

Computers in Scientific Discovery 11

Kranjska Gora, May 4-8, 2026

It is a truth almost universally
acknowledged, that a discipline
in possession of elegant
theorems must be in want of
nothing so vulgar as data.



1. An overview

* The rest of this talk (left)

2. What is data (in math)?

Terminology, history and the present

3. Where is it & what is it like?

Towards a survey

4. What can we do better?

Data management, a TL;DR

5. New(ish) uses?

Validation and correctness

Data Management Plan

No data management plan is necessary, since the research outlined in this proposal is in the realm of Mathematics and by nature theoretical. The PI will make his research freely and publicly accessible through articles, graphics and programs on the web page <INSERT WEBSITE>. The PI will also submit articles and papers to appropriate peer-reviewed journals for publication. Finally this work will be disseminated through academic research seminars and conferences.

www.math.harvard.edu/media/DataManagement.pdf
(link curtesy of Boege et al)

2. Research data

Collections of examples and beyond

- "Research data are all digital and analog objects generated or handled in the process of doing research" (*MaRDI*)
- *right*
The listing on the Cornell Mathematics Library page on "Math Databases".
- Today: collections of examples.

- MathSciNet
- Zentralblatt fur Mathematik
- Google Scholar
- Wikipedia
- MacTutor History of Mathematics
- Scopus
- The Web of Science
- Mathworld
- Jahrbuch-Project Electronic Research Archive for Mathematics (mathematics literature 1868 - 1943)
- arXiv
- ERIC (index in the field of Education, including Education in Mathematics)
- Wolfram|Alpha ("this search engine allows you to enter a query and returns an answer from structured data")

"All digital and analog objects" includes: paper publications, proofs, computational results (and more).

Data outlasting the computation

Persistence before computers

- *Earliest datable*: length measurements and corresponding areas. (Sumer, c. 2600 BCE)
- *picture*:
Pythagorean triples (Babylon, c 1800 BCE).
- c. 417 - c. 369 BCE: Platonic solids (*Theaetetus of Athens*).



Bill Casselman

Eleanor Robson, "Tables and tabular formatting in Sumer, Babylonia, and Assyria, 2500 BCE-50" *The History of Mathematical Tables from Sumer to Spreadsheets* [2003]
Campbell-Kelly et al [eds].

Data outlasting the computation

Persistence before computers

- Trigonometric, logarithmic, and exponential functions become subjects of tables.
- picture:
Napier, *Mirifici logarithmorum*,
trig and log trig data for 34°

Gr. 34		+ -			
34	Sinus	Logarithmi	Diferentie	Logarithmi	Sinus
34	5664062	5684436	3750122	1934314	8241262
35	5664459	5680105	3743801	1936314	8239614
36	5668856	5675976	3737561	1938315	8237965
37	5671252	5671750	3731413	1940317	8236316
38	5673648	5667527	3725206	1942321	8234666
39	5676043	5663305	3718980	1944326	8233015
40	5678438	5659082	3712750	1946332	8231363
41	5680832	5654872	3706532	1948340	8229711
42	5683226	5650659	3700310	1950349	8228058
43	5685619	5646449	3694090	1952359	8226405
44	5688011	5642241	3687871	1954370	8224751
45	5690404	5638036	3681653	1956383	8223096
46	5692796	5633834	3675437	1958397	8221440
47	5695187	5629635	3669222	1960412	8219784
48	5697578	5625438	3663010	1962428	8218127
49	5699968	5621244	3656799	1964445	8216469
50	5702358	5617052	3650588	1966464	8214810
51	5704747	5612862	3644379	1968484	8213151
52	5707135	5608676	3638171	1970505	8211491
53	5709524	5604492	3631965	1972527	8209831
54	5711912	5600311	3625761	1974550	8208170
55	5714300	5596132	3619557	1976575	8206508
56	5716686	5591956	3613354	1978601	8204846
57	5719072	5587781	3607154	1980628	8203183
58	5721458	5583611	3600956	1982657	8201519
59	5723844	5579441	3594756	1984687	8200854
60	5726229	5575277	3588559	1986718	8199188
61	5728613	5571114	3582364	1988750	8197522
62	5730997	5566953	3576169	1990784	8195855
63	5733381	5562793	3569976	1992819	8194188
64	5735764	5558639	3563784	1994855	8192520

1111.

grd

55

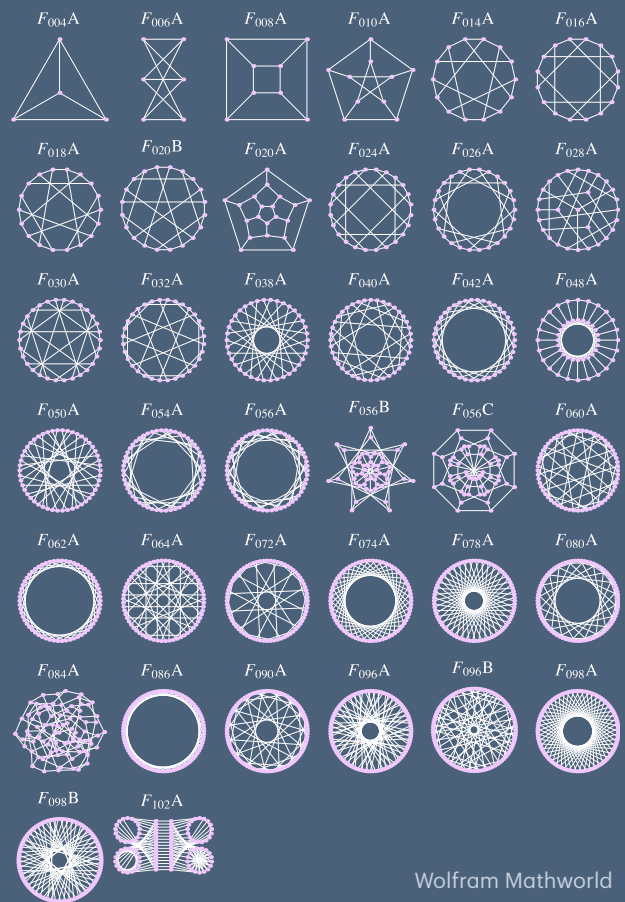
55'

John Napier, 1614

Data outlasting the computation

Persistence before computers

- 1938 - 1946: *Math Tables Project* (Works Progress Administration)
Human computers (450) constructed tables for hand computation.
- picture:
The Foster census (1930, published 1988).
Cubic symmetric graphs with up to 512 vertices. There are 207 such, Foster only missed a handful.



Wolfram Mathworld

Modern mathematical databases

Four well-established examples, just in case

The OEIS is supported by [the many generous donors to the OEIS Foundation](#).

**THE ON-LINE ENCYCLOPEDIA
OF INTEGER SEQUENCES®**

founded in 1964 by N. J. A. Sloane

The On-Line Encyclopedia of Integer Sequences (OEIS)

Enter a sequence, word, or sequence number:

[Search](#) [Hints](#) [Welcome](#) [Video](#)

For more information about the Encyclopedia, see the [Welcome](#) page.

The L-functions and modular forms database (LMFDB)

Introduction

Overview Random
Universe Knowledge

Functions

Rational All

Modular forms

Classical Maass
Hilbert Bianchi

Varieties

Elliptic curves over $\mathbb{Q}$
Elliptic curves over $\mathbb{Q}(i)$
Genus 2 curves over $\mathbb{Q}$
Higher genus families
Abelian varieties over $\mathbb{F}_q$

Fields

Number fields
 p -adic fields

Representations

Dirichlet characters
Artin representations

A database

The LMFDB is a database of mathematical objects arising in number theory and arithmetic geometry that illustrates some of the mathematical connections predicted by the Langlands program.

Click a heading on the left to browse, or go to a random page.

Learn more

Information is available regarding the source, reliability, and completeness of the database.

Known: provide explanations when you need them.

Overview LMFDB Data

Announcements

Continuing the LMFDB series of conferences after a successful first LMFDB at ICMSA, we invite you to join us at LMFDB 2025 with a call for papers! Check out the recently updated abstract groups database [beta]. Check out the new modular curves database [beta].

Citations and acknowledgments

- How to cite the LMFDB
- Source code repository
- Editorial board
- Acknowledgments
- Activities

The House of Graphs Search Search History Draw Graph Meta-directory Publications Help Log in Register

20 graphs found that satisfy the following conditions:

20 most recently added graphs [✕](#)

[Edit conditions](#)

Download these 20 graphs in the selected format:

Graph6 [Download](#)

See also: [supported formats](#)

Visible columns

HoG Id [✕](#) Name [✕](#) Number of Vertices [✕](#) Number of Edges [✕](#) Minimum Degree [✕](#) Maximum Degree [✕](#)

HoG Id	Name	Number of Vertices	Number of Edges	Minimum Degree	Maximum Degree
53687	Asymmetric graph (3322110)	7	6	0	3

π -Base Explore Spaces Properties Theorems Questions Advanced Contribute Help

π -Base

a community database of topological counterexamples

- Search spaces by name/description: [compactifications](#)
- Search spaces by properties: [non-metric continua](#)
- Find counterexamples: [connected spaces need not be path connected](#)

Topology is a dense forest of counterexamples. A usable map of the forest is a fine thing.
— Paraphrased from Mary Ellen Rudin's review of *Counterexamples in Topology*

3. MathBases.org

An index of mathematical data

- An analysis of zbMATH references revealed a long tail of data
(Hulek, Müller, Schubotz and Teschke; *Mathematical Research Data*)
- *Call to action*
Community curated index (authors, area, access method, license) and resources for authors.
Contributors welcome!
(GitHub PRs, issues, Zulip, email)

MATHBASES

INDEX OF MATHEMATICAL DATABASES

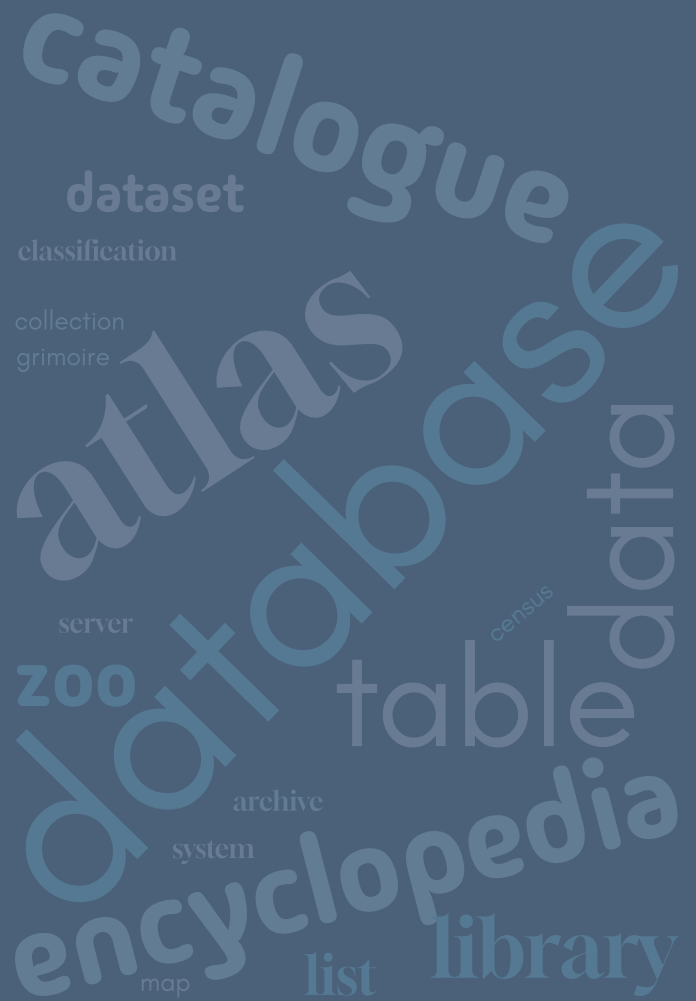
Info	Name
	(S-)Unit Groups of Orders  Gabriele Nebe , Renaud Coulangeon , Oliver Braun , Sebastian Schönnenbeck
	A Catalogue of Lattices  Neil Sloane , Gabriele Nebe
	A Census of Edge-Transitive Tetravalent Graphs  Primož Potočnik , Steve Wilson
	A Database for Number Fields  Jürgen Klüners , Gunter Malle
	A Database of Continued Fractions of Polynomial Type  Henri Cohen
	A Database of Galois polynomials  Bill Allombert , Igor Schein
	A Database of Graphs in Combinatorica Format  Sriram Pemmaraju , Steven Skiena
	Adjectives Project  Jesse Vogel , David Holmes
	An Atlas of Abstract Regular Polytopes for Small Almost Simple Groups  Laurence Vauthier , Dimitri Leemans
	An Atlas of Chiral Polytopes for Small Almost Simple Groups  Dimitri Leemans , Michael Hartley , Isabel Hubard

MathBases: index and showcase mathematical databases. 130+ databases indexed so far, with structured metadata: area, access method, license, authors.
A second function: tutorials and templates for people building new databases.

Purposes

Structure reflects intended usage

- Curating (counter)examples
- Index (fingerprint) theorems via simpler objects
- Knowledge reference
- Concept instantiation
- Benchmarking



Curation of examples: topological spaces, properties and theorems in π -base, graphs and invariants in the House of Graphs.

Index theorems: integer sequences (OEIS), Parameters of Strongly Regular Graphs.

Knowledge reference: definitions and properties of special functions in DLMF

Instantiation: datasets in algebraic geometry (only one object, variety)

Benchmarking: SuitSparse matrix dataset

The diversity of collections of examples (databases?)

Four dimensions

small - large

Atlas of Small Chiral Polytopes (56)

Lists of finite lattices ($17 \cdot 10^9$)

stored - generated on demand

House of Graphs - The Small Groups Library in GAP - nauty

enumerated - curated/collected

The Foster census

House of Graphs

easy to obtain values

nauty

- a value corresponds to a paper

Parameters of Strongly Regular

4. Doing better

FAIR and RDMP, buzzwords of the day

- Findable, Accessible, Interoperable and Reusable.
FAIR $\neq$ open data
- Research Data Management Plan
- *Call to action*
What are relevant metadata for collections of examples?

List of the 17 representatives of $IC(6,3)$, ordered by the RevLex-Index. The numbers above the signs indicate the elements of the corresponding basis.

```
11121121231121231234
22332334442334445555
34445555555666666666
IC(6,3, 1) = ++++++---++
IC(6,3, 2) = ++++++---
IC(6,3, 3) = ++++++---
IC(6,3, 4) = ++++++
IC(6,3, 5) = 0+++++---
IC(6,3, 6) = 0+++++---
IC(6,3, 7) = 0+++++
IC(6,3, 8) = 0+++++0
IC(6,3, 9) = 0+++++0---
IC(6,3,10) = 0+++++0---
IC(6,3,11) = 0+++++0---
IC(6,3,12) = 0+++++0+---
IC(6,3,13) = 0+++++0+0--
IC(6,3,14) = 0+++++0+---
IC(6,3,15) = 0000+++++
IC(6,3,16) = 0000+++++0+++
IC(6,3,17) = 000000000+++++
```

Open in "open access" refers to the removal of financial, legal and technical barriers to data, while accessibility in FAIR refers to the data being retrievable by humans and machines.

The FAIR guidelines

Mostly about metadata

Actionable advice

- **Findable:** globally unique, persistent IDs, rich metadata, indexing. *Find & identify.*
 - **Accessible:** long term storage, well-defined access conditions. *Repositories.*
 - **Interoperable:** FAIR knowledge representation language. *Ready to be combined with other data.*
 - **Reusable:** clear usage license, provenance, domain-relevant standards, comprehensive and relevant attributes.
- More and better metadata and documentation.
 - Snapshots in machine readable formats on Zenodo, GitHub, etc.
 - Record software info (version), attach code.
 - Problem with above: journal publications count, data (mostly) does not.

The FAIR guidelines (2016) are intentionally broad and somewhat vague; they are designed to provide communities with a flexible framework that can be further developed and adapted to specific needs. They focus primarily on metadata — covering aspects such as authorship, provenance, licensing, and descriptions of the dataset's contents.

Accessible: personal webpages are not considered longterm storage.

A taste of an RDMP questionnaire

What we may end up dealing it if the current trends continue

- *Project metadata:*
title, ID, grant reference, PI names, institution, ...
- *Expected questions:*
 - RDMP author,
 - data types, formats, volume and organization,
 - secure storage and backup,
 - documentation,
 - storage type, ...
- *Often disregarded in mathematics:* license.

Possibly does not apply to mathematics:

- Ethics approval, legal issues, IP, culturally sensitive issues.
- Data confidentiality and sensitivity, access restrictions (incl. cost)
- Non-digital data questions
- Data destruction

Licensing is complicated by the legal status of mathematical data.

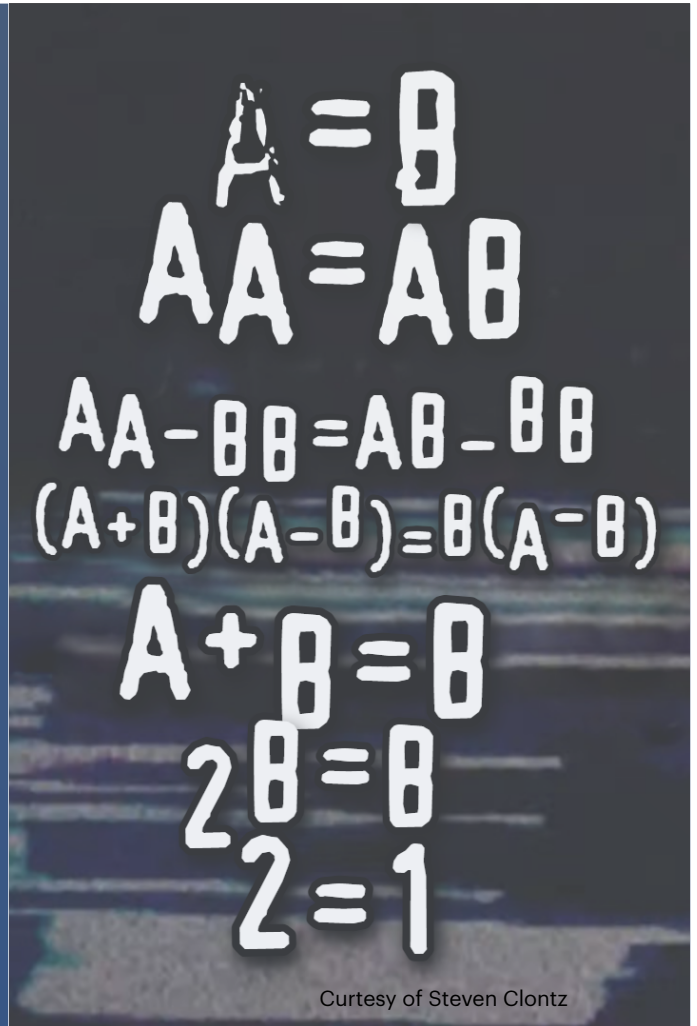
Individual mathematical facts and formulas are generally not copyrightable, but a database as a whole can be protected when the selection or arrangement of its contents constitutes an original intellectual creation; both the European Union's Database Directive and United States copyright law recognize this principle.

A curated collection or a completeness result may therefore carry copyright protection even when the individual entries do not.

5. New(ish) uses

Proof assistants + AI

- Experiments with machine learning on databases of objects.
- Databases as sources of (counter)examples for formalized mathematics.
- Recent developments seem to cause a growing interest in both types of uses.



Standard options

And a few deeper questions that we will not answer, just for fun

- **Standard checks:** format, type, consistency, uniqueness, ...
- **Testing:** software is run on test cases, the results compared to (trusted) reference results.
- **Redundancy:** several versions of software performing the same task are developed and executed independently, results compared.
- Correctness of code or data established by **formal proofs**.

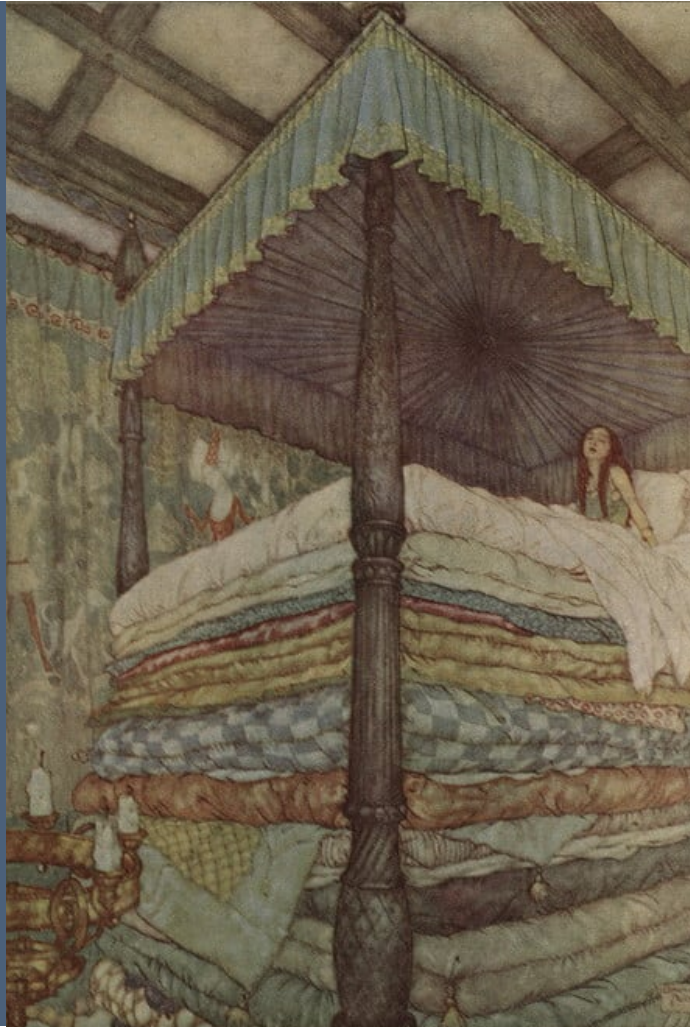
- *How can we trust that a list of examples is complete (if applicable) and correct?*
- *Is the connection between theory and code sound?*
- *Are the results of computations correct?*

No answers to the questions above, just an example: Lean-HoG

Proof assistants

Extremely finicky beasts

- Prove the properties of each example by hand.
- Implement algorithm(s) in the proof assistant.
- Encode as SAT, verify encoding to be correct, use a (trusted) solver, check the solver's certificates.
- Compute properties *and their certificates* with external software, check correctness with a proof assistant.



For few objects and properties, simple.
Few properties, many objects, efficiently computable: can be difficult.
We used a combination of the last two.

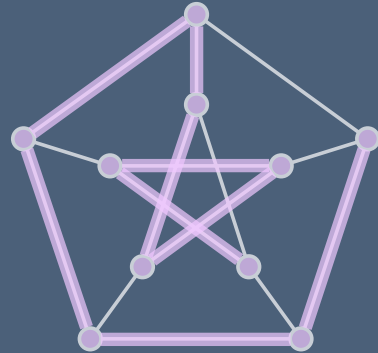
Certificates (a.k.a. witnesses)

Instead of computing values, just check correctness

- Check that 112909084933 is not a prime vs.
 $132241 \cdot 853813 = 112909084933$.

✓ Certificates *explain* why a property holds.

✗ Some properties do not have a certificate.



Easy: find a path in the graph that visits all vertices exactly once.

Hard: prove that we can't find such a cycle in the graph.

Lean-HoG

A Lean 4 library for finite simple graphs incorporating the House of Graphs

- Import efficiently represented graphs into Lean,
- with values and certificates for: the number of connected components, bipartiteness, Hamiltonian paths.
- A tactic to search the database and a tactic to close a goal by finding an example.
- Checking the number of connected components on (almost) all graphs took ~16h.

If you want to try it for yourself:

- Lean may be a sensible proof assistant to start with. A lot depends on Mathlib.
- Checking a database may make you feel like you are doing CS 50 years ago.
- Consider certificates whenever possible.
- Alternative: formal verification of algorithms.

Thank you!

Recap of what we have been up to

- **MathBases**

Adam Towsley, Ben Spitz, David Roe, David Lowry-Duda, Benjamin Hutz, Edgar Costa, KB

- **Lean-HoG**

Jure Taslak, Gauvain Devillez, KB, Andrej Bauer

- **MathDataHub**

Tom Wiesing, KB



Slides (hopefully) and references at katja.not.si, as promised

<https://katja.not.si/do-mathematicians-dream-of-data.html>