Towards a Census of Concrete Data in Mathematics

Katja Berčič[1][0000−0002−6678−8975]

August 21, 2019

Abstract. Research data are becoming ever more important in science as well as in the humanities. This is reflected in the various national and international initiatives that are aimed at developing and supporting data stewardship and the related area of knowledge management. It may come as a surprise to some that research data is experiencing a similar boom in mathematics. However, it could be argued that mathematics is lagging behind other disciplines in using the tools of the trade when it comes to data.

This work-in-progress census aims to shed light on how a large class of mathematical datasets looks like. An increased understanding of data in mathematical research is an important step towards building better infrastructure for these data. The author would like to encourage authors and curators to contribute information about their datasets for future versions of this census.

1 Introduction

Mathematicians have long been computing, collecting, and storing interesting, often hard to obtain facts and used them as reference, source of examples and counter examples, and generally to better understand the structure of objects they study. The early examples were all obtained painstakingly by hand. One such example is the computation of logarithm tables. Another example which (at least partially) predates systematic use of computers, is the Foster census of cubic symmetric graphs. This project was begun in 1930 and remarkably contained nearly all cubic symmetric graphs of up to 512 vertices by the time it was published in book form in 1988[1]. More interesting early examples can be found in a MathOverflow thread started by Gordon Royle [con].

The book form was indeed the norm until the internet revolution. The Atlas of Graphs [RW05] would probably have been published digitally if its creation

---

[1] http://staffhome.ecm.uwa.edu.au/~00013890/remote/foster/

was shifted by only a few years forward. The first two iterations of well known On-Line Encyclopedia of Integer Sequences [OEIS] were published as books in 1973 and 1973.

The circumstances that gave rise to the increasing importance of research data in general have had a corresponding effect on research data in mathematics. With access to computing power, the complexity and size of the datasets grew significantly. Mathematical datasets also found new uses, such as algorithm benchmarking (an example here is the ARG database for benchmarking of graph isomorphism algorithms [FSV01]). This growth sparked a need for tools to manage the data.

In the scientific community, we have seen the formation of FAIR principles [Wil+16], which break down the vague concept of usefulness into properties that form a basis for guidelines. The data should be findable, accessible, interoperable, and reusable. As stressed by the authors of the principles, these still need to be adapted for the needs of specific scientific communities. In a work that predates the FAIR principles (and even an earlier paper discussing similar ideas about accessibility), Billey and Tenner [BT13] outlined a set of desirable properties in a certain class of mathematical databases they call fingerprint databases for theorems. Their properties had some overlap with the FAIR principles; in particular, they require the databases (and their contents) to be citable via unique identifiers. With Vidali, the author outlined some further recommendations for mathematical databases as part of the work on the DiscreteZOO project [BV]. Data in mathematics is intrinsically linked with knowledge and as such, managing if falls into the intersection of data stewardship and mathematical knowledge management.

Data in mathematics and the scope of this census Similar to data in general, mathematical data appear in several forms. In ongoing joint work with Michael Kohlhase and Florian Rabe [BKR], we propose a division of data in mathematics into four categories (Figure 1).
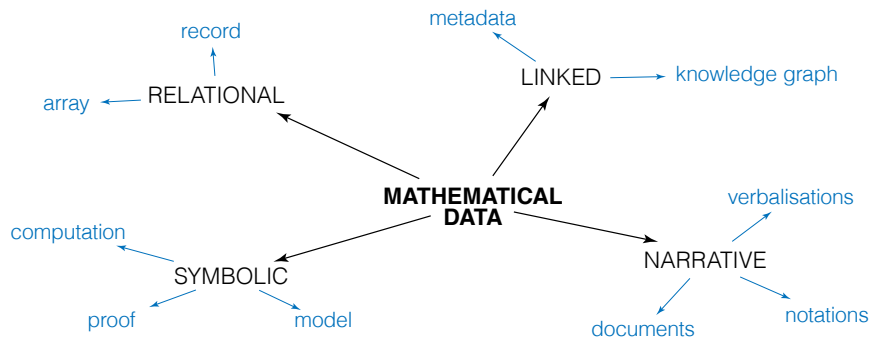

Fig. 1. Kinds of mathematical knowledge and data

Symbolic data (knowledge) are typically found in libraries such as the TPTP Problem Library for Automated Theorem Proving [SS98]. In some sense, these libraries are similar to corpora in linguistics. *Linked data* relates to data in library science and ontologies in information science. A good example of narrative data in mathematics is the repository of electronic preprints arXiv. Finally, *concrete data* (concrete as opposed to abstract) are perhaps closest to what most people think of as data and are what this census focuses on. Examples include lists of mathematical objects that could be (or are) organised into a table, such as censuses of graphs, lists of integer sequences, etc. This classification not strict: for example, the OEIS fits most neatly in the concrete data type, but it includes references to theorems and formulas for closed forms of the generating functions. The census mostly focuses on the datasets themselves, however, we will also briefly mention some of the systems that have been built for them.

Limitations It turns out that finding out what datasets are out there, and what they look like is challenging! With the exception of Billey and Tenner [BT13], there is no literature about concrete math datasets in general. Dataset authors often describe a dataset in a paper. Such papers get lost in a multitude of irrelevant results when searching for keywords such as "database" in databases of mathematical literature like arXiv, MathSciNet and zbMATH. One can find some information on swMATH information service for mathematical software [SWM] by browsing the types Data Collections and Services, Webservices, and the special collection Math.Databases - MathDBS[2]. Another source are computer algebra systems, which integrate some datasets (for example as packages). Unfortunately, most datasets live only on their authors' websites and are not indexed anywhere.

As early work, the census is strongly influenced by the author's area of mathematics. The contents are skewed towards datasets that the author knew about from the start, those discovered through word of mouth, and those datasets mentioned on MathOverflow and swMATH. It is not exhaustive or final in terms of aspects examined, nor examples given. The difficulty in obtaining information is reflected in the uneven coverage of areas of mathematics and in the uneven level of detail about specific datasets. In particular, we have not yet collected much information about datasets that are only available through computer algebra systems. Similarly, we have focused on datasets that do not appear only in commercial systems (such as Magma [BCP97] and Mathematica [Inc]).

Goals More work is needed before a more structured review can be attempted. The work reported here is a necessary first step, as it outlines the use of concrete data in mathematics. Within this larger picture, it aims to set up a foundation for creating community guidelines for FAIR mathematics, and to serve as a reference to anyone who needs to know how data in mathematics look like, such creators of data frameworks for mathematics. Finally, this census aims to increase the visibility of data in mathematics, and contribute towards better recognition of the work that goes into constructing and collecting the datasets.

---

[2] https://swmath.org/browse/types

2 Description of Datasets and Systems

This section aims to illustrate the diversity of the datasets via several aspects. Throughout, we will use the word "dataset" quite loosely to encompass both simple datasets, such as those containing a collection of objects or records with the same structure, as well as organised collections of simple datasets.

Concrete data are best characterised by the utilisation of representation theorems that allow encoding mathematical objects as simple data structures built from numbers, strings, lists and records. Such representations can be quite far from the objects' semantic type. For example, polynomials with integer coefficients can be encoded as lists of integers. Graphs can be represented as adjacency or incidence matrices, or as adjacency or edge lists. These can in turn be represented as arrays or strings (such as `graph6` [McKb]) at the database level. Further more, testing for graph isomorphism (not an uncommon task in a database) is a hard problem in general and results such as canonical forms [BL83] can be used in the encoding to help get around that difficulty. Most concrete data appears as collections of concrete mathematical objects.

2.1 Aspects

Structure Some datasets consist of simple lists of objects (such as Kohonen's giant list of unlabeled lattices on at most 15 nodes [FL]), while others are lists of records, each consisting of an object, together with some of its mathematical invariants. Larger projects can end up organising several simple datasets into a larger one, also storing the interconnection. Arguably the largest such project is the L-functions and Modular Forms Database (LMFDB [LM]) combines more than 30 datasets that have arisen in the context of the Langlands program [Ber03], which explores connections between number theory and geometry.

Content organisation The contents of datasets obtained either by generating all objects that satisfy a given set of parameters, or by collection objects in some other way.

An example of the former are some of the collections of highly symmetric objects (such as graphs). Examples of these are quite rare compared to ones without symmetry and obtaining a complete list of all objects up to a certain size can take months. These two groups of datasets also tend to have correspondingly different authorship. The unsystematic collections tend to become a collaborative effort, while the systematic collections typically have a small number of authors.

On the other hand, unsystematic collections of (rare or interesting) objects and systematic collections of all objects of a specific kind. The On-Line Encyclopedia of Integer Sequences (OEIS [OEIS]) is an example of the former that collects sequences of integers (such as the Fibonacci sequence 0, 1, 1, 2, 3, 5, 8, 13, etc.). Similarly, the stated goal of The House of Graphs [Bri+13] is "to find a workable definition of 'interesting' and provide a searchable database of graphs that conform to this definition". Both of these provide a lot of information about every object, including references to research papers.

Authorship The authorship varies widely.

- The majority of datasets has a single author or group of authors. These datasets are often accompanied by a paper (or a small number of papers) describing the mathematical background, generation, and contents.
- Some datasets have a large number of contributors; these are typically the unsystematic ones, for which a core group of authors contributed a substantive part of the data, together with a large number of authors with smaller contributions. In addition to the OEIS (with thousands of contributors), the LMFDB (with a 100 contributors) and the House of Graphs, Findstat [BSa14] (with 69 contributors) is such an example.
- A somewhat special case are the combinatorial catalogues that can consist of tens of lists of (combinatorial) objects. Some examples of these are catalogues produced by McKay [McKa], Royle [Roy], and Wanless [Wan].

Provenance The provenance of the dataset usually corresponds to its structure and authorship. The datasets with a small number of authors are usually produced via a small number of methods. We did not yet explore the provenance of the larger datasets, especially the unsystematic ones. Important to not here is that datasets can be built on top of other datasets. An example of this is the Census of cubic vertex-transitive graphs [PSV13]. The authors split the graphs into a few cases, each of which required a specialised method. One of these cases was a dataset already generated in previous work.

Infrastructure and Shareability The datasets are usually accessible either through a website or indirectly through a computer algebra system. Exceptions to these are especially older works, such as, the collection of graphs described in the book Atlas of Graphs [RW05]. Many projects with a website also provide an encyclopedic page for every object, and many researchers have commented that this is an important feature.

Especially the larger projects develop some infrastructure for the data, possibly seeding it with the initial contents. The infrastructure then supports contributions of objects (like the OEIS, the House of Graphs and Findstat), or lists of objects. The latter, hosting lists of graphs, is the other stated goal of the House of Graphs. Similarly, the Encyclopedia of Graphs [EG] is a rare example of an online resource developed to help researchers find and use data, without actually producing any of its own datasets. It currently hosts about 30 datasets. Another example is a more recent project DiscreteZOO [BV], which initially aimed to support the community studying symmetric objects.

An important example of a dataset that is only available via CAS is the Small Groups Library [SGL] (in GAP, Magma). It relies on the system to compute a significant part of the information on-the-fly and thus only uses a little under a bit per group. In a bit under 80 MB the library stores enough information to find which of the over 400 million groups a group given by the user is isomorphic to.

A more typical situation is where a dataset is hosted with minimal infrastructure on one of the authors' websites. The website is often browsable with

the browsing interface consisting of HTML tables. An illustrative example here is the Census of edge transitive graphs [EET]. Wilson has the core information about the census stored in a CAS, in which he has also written code that produces the HTML for the website. Another such example is Michael Hartley's atlas of abstract polytopes [AP]. Authors also often provide files with code for the collection as an array in some computer algebra system.

Metrics There is no standard measure for a *size* of a dataset. It is possible to consider the compressed or uncompressed size on disk, the number or size of objects, or the time (itself a problematic metric) it took to generate the dataset, etc. The uncompressed size on disk can range from a few megabytes to over a terabyte, or up to roughly 25 GB with heavy compression. There is a small inverse correlation between the size (on disk or the number of objects) and computational complexity of the process of generation. Kohonen's lattice dataset appears to be a record holder with respect to size, with roughly $17 \cdot 10^9$ objects.

The number of users can be estimated through the number of citations, or the number of citations of the corresponding paper for some projects. The number of downloads would be interesting, but it appears that nearly no-one records it.

FAIR-related aspects For details about the FAIR principles, we refer the reader to the GoFAIR website [GF].

Metadata. Some details about a dataset are typically available on the same website as the dataset, and most datasets have an accompanying paper. Metadata are generally not structured and, with a few possible exceptions, do not specify a license.

Unique IDs and **F**indable. Many datasets have some sort of an unique ID for all the objects Some of the projects also provide some sort of a globally unique ID (in the sense of a URL), but the persistence of it is bounded by the projects' lifespan, as the URLs will expire if the website is decommissioned.

ET [EET]	C4[10, 1]
Findstat	http://www.findstat.org/StatisticsDatabase/St000001/
OEIS	https://oeis.org/A000045

Table 1. Examples of unique IDs

Accessible. While most datasets are in some way available online, the formats are typically ad hoc. Almost all of the others are available through computer algebra systems.

Area of mathematics At least partly due to the author's home area, the information collected so far has mostly been skewed towards combinatorics and geometry. In addition to these, we have found datasets from number theory (LMFDB,

NFDB [NF]), group theory and algebra (the Small Groups Library [SGL], the Graded Ring Database [GRD]), topology (the Knot Atlas [BNMa]), algebraic geometry (the Toric Calabi-Yau Database [CY]), and probability (Distributome [Dst]).

3 Living census

To facilitate the collection of information about data in mathematics, we set up a database with a website frontend [Ber]. While it grew out of the necessity to keep track of the information, it has at least two further goals. First, it aims to make it easy for anyone to see what information has been collected so far. Second, it aims to eventually make it easy to contribute information.

The information about the datasets can be displayed a few different views (with switching implemented through tabs): general information, information about size, information pertaining to the FAIR principles, as well as some other properties.

Fig. 2. The living census website

The FAIR principles in particular are a little unwieldy to get an impression about at a glance, which is why we devised simple diagrams (Figure 3) to aid with that. The design of the diagrams is based off the fact that each of the four principles (findable, accessible, interoperable, and reusable) is composed of 3-4 sub-principles, Findable (**F1**, **F2**, **F3**, and **F4**), Accessible (**A1**, **A1.1**, **A1.2**,

and **A2**), Interoperable (**I1**, **I2**, and **I3**), and Reusable (**R1**, **R1.1**, **R1.2**, and **R1.3**) [Wil+16]. Each of these can be applied to (or not) to one of the three layers of information about the dataset. These layers are not necessarily included in the original FAIR principles, but it seems to be helpful to break information down depending on whether it applies to

- the dataset (**D**) itself (such as whether the dataset has its own globally unique identifier or whether it is registered in a searchable resource),
- the datum (**A**) (each of the objects needs its own globally unique identifier), or
- the metadata (**M**) (such as whether the metadata is accessible even after the data are no longer available).

The colour of each cell in the diagram corresponds to a value for a subprinciple-layer pair: unknown (black), not considered (blank), mostly supported (green), somewhat supported (yellow) and mostly unsupported (red).

For example, let us consider **F1** for FindStat [BSa14], the Combinatorial Statistic Finder. The dataset (but not the metadata) is indexed in zbMATH (Findable **D**, **M**). Each combinatorial statistic in the dataset has a unique identifier, such as `St000081`[3] and can be found through a search interface (Findable **A**).

	D	A	M
F1	🟩	🟩	🟥
F2			⬛
F3			🟨
F4	🟩	🟩	🟥

Fig. 3. An example of a diagram for Findable.

4   Conclusions and Future Work

Perhaps the most important immediate use for this census is as a "market study" for a prototypal unified infrastructure for mathematical data, Math-DataHub [BKR19]. It serves as a source of use cases for the infrastructure, as well as beginnings of a community of researchers that work with mathematical data. Even in this initial stage, the census gives the developers of MathDataHub some idea of the requirements for the system in terms of the ranges of dataset size, complexity, etc.

We will continue to gather information about the concrete datasets in mathematics in the living census website. One way to find more datasets would be to

---

[3] http://www.findstat.org/StatisticsDatabase/St000081/

(in a way that is not yet clear) search for literature in all areas of mathematics (but not computer science) with keywords "database", "atlas", "census" and similar. Such a search currently does not appear to be supported by any of the major databases of mathematical literature. Another large set of datasets that has yet to be added to the census are datasets incorporated into the various computer algebra systems.

Finally, we plan to use the new information as a basis for a more structured census.

References

[AP]   Michael Hartley. *Abstract Polytopes*. URL: http://www.abstract-polytopes.com/atlas/index.html (visited on 01/23/2019).

[BCP97]   Wieb Bosma, John Cannon, and Catherine Playoust. "The Magma algebra system. I. The user language". In: *J. Symbolic Comput.* 24.3-4 (1997). Computational algebra and number theory (London, 1993), pp. 235–265. ISSN: 0747-7171. DOI: 10.1006/jsco.1996.0125.

[Ber]   Katja Berčič. *Math Databases table*. URL: Https://mathdb.mathhub.info/ (visited on 01/15/2019).

[Ber03]   Steve Bernstein Joseph Gelbart, ed. *An Introduction to the Langlands Program*. Birkhäuser, 2003. ISBN: 3-7643-3211-5.

[BKR]   Katja Berčič, Michael Kohlhase, and Florian Rabe. "(Deep) FAIR Mathematics". submitted. URL: https://kwarc.info/kohlhase/submit/it19.pdf.

[BKR19]   Katja Berčič, Michael Kohlhase, and Florian Rabe. "Towards a Unified Mathematical Data Infrastructure: Database and Interface Generation". In: *Intelligent Computer Mathematics (CICM) 2019*. Ed. by Cezary Kaliszyck et al. LNAI. in preparation. Springer, 2019, pp. 28–43. URL: https://kwarc.info/kohlhase/papers/cicm19-MDH.pdf.

[BL83]   László Babai and Eugene M. Luks. "Canonical Labeling of Graphs". In: *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*. STOC '83. New York, NY, USA: ACM, 1983, pp. 171–183. ISBN: 0-89791-099-0. DOI: 10.1145/800061.808746.

[BNMa]   Dror Bar-Natan, Scott Morrison, and et al. *The Knot Atlas*. URL: http://katlas.org.

[Bri+13] Gunnar Brinkmann et al. "House of Graphs: a database of interesting graphs". In: *Discrete Appl. Math.* 161.1-2 (2013), pp. 311–314. ISSN: 0166-218X. DOI: 10.1016/j.dam.2012.07.018.

[BSa14]   C. Berg, C. Stump, and al. *FindStat: The Combinatorial Statistic Finder*. http://www.FindStat.org. [Online; accessed 31 August 2016]. 2014.

[BT13]   Sara C. Billey and Bridget E. Tenner. "Fingerprint databases for theorems". In: *Notices Amer. Math. Soc.* 60.8 (2013), pp. 1034–1039. ISSN: 0002-9920. DOI: 10.1090/noti1029.

[BV]   Katja Berčič and Janoš Vidali. "DiscreteZOO: a Fingerprint Database of Discrete Objects". In: *Mathematics in Computer Science* (). accepted. URL: https://arxiv.org/pdf/1812.05921.pdf.

[con]   MathOverflow contributors. *What are some early examples of creation of lists / catalogues of (particularly) combinatorial objects?* MathOverflow. URL: https://mathoverflow.net/questions/47044/what-are-some-early-examples-of-creation-of-lists-catalogues-of-particularly (visited on 11/20/2018).

[CY]   *Toric Calabi-Yau Database*. URL: http://www.rossealtman.com/index.html (visited on 06/18/2019).

[Dst]   *Distributome*. URL: http://www.distributome.org/ (visited on 06/18/2019).

[EET]   Steve Wilson and Primož Potočnik. *A Census of edge-transitive tetravalent graphs*. URL: https://jan.ucc.nau.edu/~swilson/C4FullSite/index.html (visited on 01/23/2019).

[EG]   *Encyclopedia of Graphs*. URL: http://atlas.gregas.eu (visited on 01/24/2019).

[FL]   Jukka Kohonen. *Lists of finite lattices (modular, semimodular, graded and geometric)*. URL: https://www.shsu.edu/mem037/Lattices.html (visited on 01/25/2019).

[FSV01]   P. Foggia, C. Sansone, and M. Vento. "A Database of Graphs for Isomorphism and Sub-Graph Isomorphism Benchmarking". In: -. Jan. 1, 2001, 176187.

[GF]   *GoFAIR*. URL: https://www.go-fair.org/fair-principles/ (visited on 06/18/2019).

[GRD]   *Graded Ring Database*. URL: http://www.grdb.co.uk/ (visited on 06/18/2019).

[Inc]   Wolfram Research, Inc. *Mathematica, Version 12.0*. Champaign, IL, 2019.

[LM]   *The L-functions and Modular Forms Database*. URL: http://www.lmfdb.org (visited on 02/01/2016).

[McKa]   Brendan McKay. *Combinatorial Data*. URL: http://users.cecs.anu.edu.au/~bdm/data/ (visited on 01/25/2019).

[McKb]   Brendan McKay. *Description of graph6, sparse6 and digraph6 encodings*. URL: http://users.cecs.anu.edu.au/~bdm/data/formats.txt.

[NF]   *Number Fields*. URL: https://hobbes.la.asu.edu/NFDB/ (visited on 06/18/2019).

[OEIS]   *The On-Line Encyclopedia of Integer Sequences.* URL: http://oeis.org (visited on 05/28/2017).

[PSV13]   Primož Potočnik, Pablo Spiga, and Gabriel Verret. "Cubic vertex-transitive graphs on up to 1280 vertices". In: *J. Symbolic Comput.* 50 (2013), pp. 465–477. ISSN: 0747-7171. DOI: 10.1016/j.jsc.2012.09.002.

[Roy]   Gordon Royle. *Combinatorial Catalogues.* URL: http://staffhome.ecm.uwa.edu.au/~00013890/data.html (visited on 01/25/2019).

[RW05]   Ronald C. Read and Robin J. Wilson. *An Atlas of Graphs.* New York, NY, USA: Oxford University Press, Inc., 2005. ISBN: 0198526504.

[SGL]   *The Small Groups Library.* URL: http://www.icm.tu-bs.de/ag_algebra/software/small/small.html (visited on 04/16/2019).

[SS98]   G. Sutcliffe and C. Suttner. "The TPTP Problem Library: CNF Release v1.2.1". In: *Journal of Automated Reasoning* 21.2 (1998), pp. 177–203.

[SWM]   *Mathematical Software – swMATH.* URL: http://swmath.org (visited on 09/07/2017).

[Wan]   Ian Wanless. *Combinatorial Data.* URL: http://users.monash.edu.au/~iwanless/data/ (visited on 01/25/2019).

[Wil+16]   Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3 (2016). URL: https://doi.org/10.1038/sdata.2016.18.