Making data useful for mathematical collaboratories and the example of the LMFDB project

The online sharing of data and computational resources has so far revolutionized all sciences but mathematics. While other fields have to contend with issues concerning reproducibility, copyright, ethics, etc a repository for mathematical data has to face its own set of problems to be fully useful and directly assist in research. It is remarkable that many of these hurdles have fallen down recently (for instance with the arrival of sage as an open-source computer algebra system), but specific issues remain relating to the integration of disparate mathematical databases. These issues have crept up in the LMFDB project, which has a very vertical approach, focused on one relatively narrow area of mathematics but facing several obstacles at each stage of the process. The perspective here is more horizontal, targeted at the specific issue of integration, expansion and maintenance of existing databases, but with a wider intended audience, essentially the whole pure mathematics community. I am party to the sage and LMFDB projects.

I. A broad view of collaboratories

In 2002–2007, the U.S. National Science Foundation funded a research program called the *Science of Collaboratories* [17] to survey large-scale academic research collaborations in scientific disciplines or the humanities. This program used the following definition:

> "A collaboratory[1] is an organizational entity that spans distance, supports rich and recurring human interaction oriented to a common research area, and fosters contact between researchers who are both known and unknown to each other, and provides access to data sources, artifacts, and tools required to accomplish research tasks."

(In a mathematical context, I will take *artifacts* and *tools* to mean computational resources, either hardware or software, since they are essential for an experimental mathematician).

This research program compiled a list of over 200 collaboratories that includes some very large scale projects (such as the Sloan Digital Sky Survey [24], the Atlas experiment at CERN [1] or the Human Genome Project [6]) and some much smaller projects (such as the e-Mouse Atlas Project [4], papyri.info [13] or the FaceBase Hub [5]). They also developed a taxonomy for these collaboratories, and a checklist for their success.

While the larger projects are clearly exceptional, many of the smaller ones are also actively transforming their respectively fields. It is clear that some projects are missing in this survey, such as the ATLAS of Finite Group Representations [19] and the Atlas of Lie Groups and Representations [2]. The only pure[2] mathematics projects in the full list are the Mathematics Genealogy Project [8] and the Polymath project [15], and neither is concerned with mathematical data.

Clearly, while other scientists and humanities researchers are actively sharing open data and interacting with it in complex ways *across databases*, mathematicians have very rarely broken the model of each putting their own data up on their own repository/homepage, for others to download.

[1] A *portmanteau* of *collaboration* and *laboratory*.

[2] My count excludes financial/economics projects as well as projects covering all the sciences such as the Open Science Grid project [10], or the DOE Science Grid [3] (as far as I know, these have little involvement from pure mathematicians at the moment).

II. Mathematical near-collaboratories

There are other projects that are nearly collaboratories, and they are interesting to consider as well. Some courageous *individuals* or very select groups, through hard work, gather all the data in their own mathematical area of interest (hence also data produced by others). For instance, one can find (precomputed) data on knots and their invariants at KnotInfo [22] or at the KnotAtlas [21]. In combinatorics, there is also the Online Encyclopedia of Integer Sequences [9], which is now more collaborative but offers very limited browsing capabilities. In neither of these two cases do we have a true collaboratory :

- If the user wants to perform more complex searches, the only option[3] is to download the whole data and do these searches locally. This is only possible because these data sets are modest in size.
- The process of inclusion of new data relies on very few individuals to do the actual work, which prevents the community to effectively contribute its expertise.

In order words, the ability for the end-user to enrich the data or process it differently is still fairly limited. Despite these shortcomings, these examples are still important since they have truly changed the way many mathematicians do their work in those areas.

Since 2007 and that NSF-sponsored survey, some of the environment in mathematics has changed. The sage project (that has enjoyed exponential growth since its start around 2005) is helping many mathematical subcommunities to coalesce around one common open-source framework. It provides a unifying[4] and inclusive[5] setting for mathematicians to collaborate, at the level of algorithms. As its intended audience includes the whole mathematical community, I will qualify this project as *horizontal*. As highlighted in [23], the existence of sage is essential as it provides a common base to the collaboration. GAP and custom software served the same purpose for the very successful Atlases.

Data sharing in the sage project is currently minimal and limited to the download of various heterogeneous databases (each database is linked to sage itself, but there is no deeper interaction between the various databases, which causes inefficient queries across databases).

III. The example of the LMFDB

I have recently participated in a workshop on an exciting and ongoing project (the LMFDB project [7]). This would be among the largest mathematical projects so far to fully qualify for the 2007 *Collaboratories* survey.

The LMFDB project concerns L-functions and associated objects such as modular forms, Dirichlet characters, elliptic curves, abelian varieties, Artin representations, etc. Its main purpose is first to aggregate data already computed by dozens of researchers in number theory into one coherent framework (either served on the web[6] or more directly accessed by tapping into the database from standard computer algebra software). In a

[3] The OEIS also has a superseeker functionality, which is based on email and limited to one query an hour. It tries to find matches for a predefined set of transforms of the query sequence.

[4] See [18] for a partial list of mathematical areas of coverage.

[5] sage is written in Python [16], a language particularly easy to interface with other languages and programs. As such, it is ideal to recycle previous work and seamlessly includes nearly all the previous open-source mathematical work. It also has interfaces with Mathematica, Maple, Magma, etc.

[6] A prototype exists, but is still in beta version. At the moment, the prototype website *displays* data relatively uniformly (> 1 Tb for a dozen different main types of objects) but does not allow for complex queries. Also, it manages to explain the data in a non-obstrusive way, using an innovative system (interesting in its own right for educational applications).

second stage, the project will include more computational features, such as the ability to generate more data, to check for consistency or to perform complex queries.

Since the LMFDB project concerns a relatively narrow slice of mathematics, but intends to provide many features, I will qualify it as *vertical*.

The unified name of *L-function* can easily mislead on the main difficulty and goal of this project at the mathematical level: it is a consequence of the far-reaching yet conjectural Langlands programme that these functions should have similar properties and fit a unified pattern. In practice however, these objects come from different sources and often originate from the work of different researchers, with different knowledge, assumptions and points of view. This is what will make this project a true mathematical collaboratory: some of the main progress on the understanding of the data and the underlying mathematics is expected to come specifically from data sharing among dozens of mathematicians.

Since L-functions, modular forms and zeta functions are so prevalent, it is expected that many number theorists should find a part of this project actually relevant to them (this might also include some physicists, for instance those studying string theory or quantum chaos).

Before discussing issues in the project, I should insist that the website is already useful in its current, unfinished, state. However, as a torch-bearer in mathematics, this project has to face many new obstacles, and I will now focus on one particular issue.

IV. An issue likely to arise again

The main weakness of the LMFDB system as currently implemented is that there is a bottleneck in introducing new types of objects, or even different data fields to already known objects. The data can be uploaded easily, but ends up forming one more collection in a very heterogeneous bunch of data. Some tedious matching has to be done with all the collections already included, which leads to serious issues of scalability. Without a coherent database of objects, more complicated yet realistic and interesting queries are impossible (study of the smallest non-real zero ever computed for an L-function of analytic rank 1 of *any* type, for instance).

I would expect this difficulty in integrating new data to be the main cause of the lack of mathematical collaboratories. As such, it should be solved as a horizontal rather than vertical problem. From now on, I only discuss this problem, *i.e.* the creation of a flexible and coherent mathematical database starting from heterogenous data. Additional layers such as data browsing, data searching, data checking or data improvement will be easier to implement if the data structure is judiciously chosen.

V. A proposed solution

My suggested approach to this issue of integrating mathematical data is to break the process down into many concurrent steps, with each dependent on very few skills or little knowledge. The clear intention of structuring the workflow in this way is to make a larger set of individuals capable of performing any single step.

The best method to achieve this is to separate mathematical from programming skills, and data from theory. Mathematicians should feel like they are writing mathematics in a very lightweight language, and contribute to the formalization of the mathematics studied *regardless of the data currently known*[7]. Some deeper mathematical knowledge

[7] This is the approach that has been taken in chemistry, for instance, with the OpenBabel project [12]. One of the main steps in that project was to establish some standards on how to refer to the chemical objects in the abstract, not as the results of specific experiments (via the definition of C++ classes such as OBMol, OBAtom [11], etc). This separates theory and experimental results. The approach would be similar here, but has to be more dynamic, acknowledging the constant "risk" in

in the LMFDB resides only with people with limited programming experience, but they still need to be able to contribute to this first step of standardization. Mathematicians who have produced the data should specify how their data ties up with that formalism (or simply pass on the raw data for someone else to do that step). Finally, it is up to people with more programming experience (some of them not necessarily mathematicians) to actually implement this interface, or even better to automate that step (this is best done gradually, with more and more components being automated). Once automated, this last step would be of great benefit to the mathematical community, as it could then be reused for any other mathematical collaboratory ready to go through the trouble of formalizing their mathematical objects in the same system. This automatic interfacing is actually a fairly frequent problem in computer science (called *Object* or *Document Relational Mapping*). In Python, that type of problem is usually solved using *metaclasses*. Conveniently, metaclasses have already been successfully used by mathematicians in implementing the *categories* framework in the sage subproject sage-combinat [20], so there is some support available there.

To summarize, I have in fact described two projects in this annex.

The LMFDB project as a whole will bring a more complete understanding of L-functions and as such will be tremendously interesting. I will extensively contribute to this project, probably more in the analysis tools than in providing the data itself.

I have also described my more personal project, which is concerned with the integration of mathematical data for collaboratories. While I do not consider this project a mathematical project, I think it could be of great value to research in mathematics in general. I intend on pursuing it on the side as a service to the community, just as paper refereeing, outreach, CAS development, etc. I am only mentioning it here because of its (presumed) originality and wide audience.

References

[1] *The Atlas experiment at CERN*. http://atlas.web.cern.ch/Atlas/Collaboration/.
[2] *Atlas of Lie Groups and Representations*. http://www.liegroups.org/.
[3] *DOE Science Grid*. http://doesciencegrid.org/.
[4] *The e-Mouse Atlas Project*. http://www.emouseatlas.org/emap/home.html.
[5] *The FaceBase Hub*. http://www.facebase.org/.
[6] *Human Genome Project*. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.
[7] *The L-functions and Modular Forms DataBase*. Website currently in private beta.
[8] *The Mathematics Genealogy Project*. http://genealogy.math.ndsu.nodak.edu/.
[9] *The On-Line Encyclopedia of Integer Sequences*. http://oeis.org.
[10] *Open Science Grid*. http://www.opensciencegrid.org/.
[11] *OpenBabel: OBMol class reference*. http://openbabel.org/dev-api/classOpenBabel_1_1OBMol.shtml.
[12] *OpenBabel: the open source chemistry toolbox*. http://openbabel.org.
[13] *The papyri.info project*. http://papyri.info/.
[14] *Persistency [of data] in ATLAS*. http://twiki.cern.ch/twiki/bin/view/Persistency/WebHome.
[15] *The Polymath project*. http://polymathprojects.org/.
[16] *The Python Programming Language*. http://www.python.org/.
[17] *Science of Collaboratories*. http://soc.ics.uci.edu/.
[18] *Some components integrated in sage*. http://www.sagemath.org/links-components.html.
[19] Atlas *of Finite Group Representations*. http://brauer.maths.qmul.ac.uk/Atlas/v3/.
[20] *Sage-Combinat: enhancing Sage as a toolbox for computer exploration in algebraic combinatorics*, 2011. http://combinat.sagemath.org.
[21] D. Bar-Natan, S. Morrison, and et al., *The Knot Atlas*. http://katlas.org.
[22] J. C. Cha and C. Livingston, *KnotInfo: Table of Knot Invariants*. http://www.indiana.edu/~knotinfo.

mathematics that someone would want to consider your objects from a more general standpoint. See also the CORAL and POOL subprojects for Atlas [14], for instance.

[23] J. S. Olson, E. C. Hofer, N. Bos, A. Zimmerman, G. M. Olson, D. Cooney, and I. Faniel, *A Theory of Remote Scientific Collaboration*, 2008, pp. 73–+.

[24] A. S. Szalay, J. Gray, A. R. Thakar, P. Z. Kunszt, T. Malik, J. Raddick, C. Stoughton, and J. vandenBerg, *The SDSS skyserver: public access to the Sloan digital sky server data*, in Proceedings of the 2002 ACM SIGMOD international conference on Management of data, SIGMOD '02, New York, NY, USA, 2002, ACM, pp. 570–581.