Sage and online databases: from L-functions to combinatorics

Paul-Olivier Dehaye and Nicolas M. Thiéry

November 1, 2011

In the past decade, data mining in huge (open source) databases has become an essential tool in many areas of experimental sciences. This trend is now coming to mathematics. An example is given by the LMFDB project, which studies L-functions [2]. The design of such databases is challenging: besides the usual issues of *scale* and *heterogeneity* and *licensing* for which lessons need to be learned from experimental sciences, one faces a specific issue of *modeling complex mathematical objects*.

We propose a workshop aiming at exploring those challenges, with the specific goal of refining the design and implementation of the LMFDB project. It will bring together number theory experts, mathematicians with strong expertise in object oriented modeling, as well as invited speakers from database projects in experimental sciences and computer science experts in object oriented databases and web services. To widen the discussions, developers of other mathematical databases, in particular in combinatorics, will be invited too; this will ensure that the lessons learned and tools developed within the Sage system during that workshop will be of immediate wide use.

1 Context

1.1 Mathematics

At present there are several online resources which are used for research mathematics, such as: preprint servers (the arXiv), MathSciNet, MathOverflow [3], Sage [30], and websites with encyclopedic content (with various quality levels: Wikipedia, PlanetMath [5], Wolfram MathWorld [8]). Their aims are either to share mathematical results or mathematical software.

In addition, there are websites with mathematical data, specialized to one area of mathematics. Interaction with the data then tends to be very limited: it is most often restricted to a full download of the database, cut-and-paste, browsing via a fixed webpage, or querying via a limited interface.

Beyond the pioneer example of the Atlas of Finite Group Representations [6], we cite examples in a few areas, and emphasize their particularities.

In combinatorics, the most famous is certainly the Online Encyclopedia of Integer Sequences [29]. Another website has recently sprung up, that offers the option to recover information about a combinatorial statistic from a few of its values: www.findstat.org [10]. Several databases of graphs also exist (*e.g.* [19, 25]), and even a database of graph properties [15]. One should also mention the work of a the Algorithms Project at INRIA, which is building an Encyclopedia of Combinatorial Structures [22] (interestingly, they are reusing a framework already built for their Dynamic Dictionary of Mathematical Functions [21]).

In knot theory, historically many databases of knots and their numerous invariants have been compiled. At some point, the associated algorithms have been reimplemented in Mathematica, into the package KnotTheory‘. Data was then recomputed centrally, which led to unified databases [12, 9] (albeit with coverage over less knots).

In number theory, possibly even more databases exist: number fields [23], algebraic curves of various degree and genus (*e.g.* [13, 32]), modular forms [31], Artin representations [16], values of L-functions or location of their zeroes (*e.g.* [26, 28, 27, 24, 11]). While the computed data cited here is vastly heterogeneous, the Langlands program conjectures that all those objects should fit into one unified pattern.

1.2 Integration of heterogeneous databases in experimental sciences

A database becomes most useful when it integrates data from many different sources and lets users access that data transparently.

Google Earth [17] and Google Maps [18] (and their open source alternative like NASA World Wind [20] and the OpenStreetMap [4]) are very popular examples, giving access to a huge variety of GIS sources. Their implementation even allows exterior developers to easily add new interfaces and innovative ways to exploit that data.

Throughout the experimental sciences, a similar process has been repeated, with great success. For example, the final output of the Human Genome project [1] integrates data processed by thousands of scientists (the final output is only 1.5 Gb, however, allowing each of those scientists to hold a copy of the whole final dataset).

For the Sloan Digital Sky Survey [33], data from different telescopes was compiled, leading to “Virtual Observatories” [7]: requests can be made spanning time (decades), wavelengths, or bearings, and these will return homogeneous data, masking to the user that this data was possibly obtained by very different telescopes.

These projects overcame significant issues of scale, heterogeneity and licensing of the data. This complexification of databases is a trend that is also coming to mathematics. A clear example is given by the LMFDB project.

2 The LMFDB project

LMFDB stands for

> "The database of modular forms, L-functions, and related objects."

Indeed, the objects currently included are L-functions, modular forms (classical, Siegel, and Hilbert), elliptic curves, number fields (both local and global), and Dirichlet characters. Before the website reaches its official release, it will also contain Artin representations, Hecke characters, and hyperelliptic curves. This project is hosted at the website http://www.l-functions.org/. It is implemented in the Sage system.

This AIM proposal originated from an LMFDB workshop in September 2011. Proposer Dehaye began collaborating with Tim Dokchitser to put Artin L-functions into the LMFDB. The plan was to mimic the mechanisms by which other objects had been added to the website. They indeed found that this would be possible, but they had to make a choice of which model to follow because there was a lack of uniformity to how the different objects were handled. This led to a realization that the long-term viability of the LMFDB project, and its ability to serve as a model for mathematicians in other areas, was not guaranteed.

3 Goals

In this workshop, we want to address those issues and make the LMFDB a viable long-term model, which we can then start to export to combinatorics.

Leaning on the experience gained during the first implementation, we propose to revamp the backend of the website. We want to break down the process of integrating data 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 suggested technique to achieve this, that will be refined and evaluated during the workshop, 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. Some deeper mathematical knowledge 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 would be 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 collaboration ready to go through the trouble of formalizing their mathematical objects in the same system.

In practice, all this could be achieved using principles of object-oriented programming to their fullest. From the lightweight implementation of the idealized mathematical objects (for the formalization), all the actual objects would be implemented via inheritance. Data abstraction would separate the data from the idealized versions that the displaying, browsing or searching components work on. Polymorphism would help solve the more subtle issues tied to the different origins for the data.

Proposer Thiery has been leading since 2000 the `Sage-Combinat` project (formerly `MuPAD-Combinat`), whose goal is to improve `Sage` as a platform for computer exploration in algebraic combinatorics. The specific needs in this area led him to gain strong experience in modeling complex mathematical objects in computer algebra platforms, and to become the main architect and implementer of the object oriented foundations of `Sage` (the so-called category framework). This framework is now used pervasively through the `Sage` library, with very much the same structuring aims as what is needed for the LMFBD project.

In addition, he will make sure that the workshop reaches its goal of exporting these tools to the combinatorics community.

4 Organization of the workshop

This workshop will involve participants with the variety of skills needed to match its goals. It will involve research mathematicians with extended programming experience, drawn from the number theory and combinatorics communities. In general, the number theory data tends to be more heterogeneous, while the combinatorics people tend to have more experience with an object-oriented approach.

The circumstances of an AIM workshop will be ideal for the fruitful exchange of needs and ideas between these two groups. We intend to formalize the needs for a database as described above, implement a new version of the LMFDB database according to these wishes, and package the parts of this work that are generic in such a way that it can serve for other areas of mathematics, and foremost for graphs, combinatorial statistics, and congruence groups [14]. Talks will mostly serve to expose the issues we face, discuss solutions and structure our progress.

References

[1] *Human Genome Project.* `http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml`.

[2] *The L-functions and Modular Forms DataBase.* http://www.l-functions.org.

[3] *MathOverflow.* http://www.mathoverflow.net.

[4] *OpenStreet Map.* http://www.openstreetmap.org/.

[5] *PlanetMath.* http://www.planetmath.org.

[6] Atlas *of Finite Group Representations.* http://brauer.maths.qmul.ac.uk/Atlas/v3/.

[7] *Virtual Observatory, wikipedia definition.* http://en.wikipedia.org/wiki/Virtual_Observatory.

[8] *Wolfram MathWorld (formerly Eric's Treasure Trove of Mathematics).* http://http://mathworld.wolfram.com.

[9] D. Bar-Natan, S. Morrison, et al., *The Knot Atlas.* http://katlas.org.

[10] C. Berg, F. Saliola, and C. Stump, *Combinatorial statistic finder.* http://http://www.findstat.org/.

[11] A. Booker, A. Strmbergsson, and A. Venkatesh, *Effective computations with Maass forms; data files.* http://www2.math.uu.se/~astrombe/emaass/emaass.html.

[12] J. C. Cha and C. Livingston, *KnotInfo: Table of Knot Invariants.* http://www.indiana.edu/~knotinfo.

[13] J. Cremona, *The elliptic curve database for conductors to 130000*, in Algorithmic number theory, vol. 4076 of Lecture Notes in Comput. Sci., Springer, Berlin, 2006, pp. 11–29.

[14] C. Cummins and S. Pauli, *Congruence subgroups of PSL(2,Z).* http://page.math.tu-berlin.de/~pauli/congruence/.

[15] H. de Ridder et al., *Information System on Graph Classes and their Inclusions.* http://www.graphclasses.org/.

[16] T. Dokchitser, *Database of Artin representations.* http://www.maths.bris.ac.uk/~matyd/, included in Magma.

[17] Google, *Google Earth.* http://earth.google.com.

[18] ——, *Google Maps.* http://maps.google.com.

[19] J. Grout, *The Graph Database (graphs of fewer than 8 vertices).* http://artsci.drake.edu/grout/graphs/, fully included in sage.

[20] P. Hogan, *NASA World Wind: infrastructure for spatial data.*, in COM.Geo, L. Liao, ed., ACM International Conference Proceeding Series, ACM, 2011, p. 2.

[21] INRIA Algorithms Project, *Dynamic Dictionary of Mathematical Functions*. http://ddmf.msr-inria.inria.fr/1.6.1/ddmf.

[22] ——, *Encyclopedia of Combinatorial Structures*. http://algo.inria.fr/encyclopedia/.

[23] J. Jones, *Number fields*. http://hobbes.la.asu.edu/NFDB/.

[24] S. Lemurell, *Computational data on Maass forms*. http://www.math.chalmers.se/~sj/.

[25] M. Meringer, *Regular graphs*. http://www.mathe2.uni-bayreuth.de/markus/reggraphs.html.

[26] A. Odlyzko, *Tables of zeros of the Riemann zeta function*. http://www.dtc.umn.edu/~odlyzko/zeta_tables/index.html.

[27] M. Rubinstein, *Values of L-functions of degree less or equal to 2*. http://oto.math.uwaterloo.ca/~mrubinst/L_function_public/VALUES/.

[28] ——, *Zeroes of elliptic curve L-functions*. http://oto.math.uwaterloo.ca/~mrubinst/L_function_public/ZEROS/.

[29] N. Sloane and al., *The On-Line Encyclopedia of Integer Sequences*. http://oeis.org.

[30] W. Stein et al., *Sage Mathematics Software (Version 4.7.1)*, The Sage Development Team, 2011. http://www.sagemath.org.

[31] W. A. Stein, *The Modular Forms Database*. http://modular.ucsd.edu/Tables, 2004.

[32] W. A. Stein and M. Watkins, *A database of elliptic curves—first report*, in Algorithmic number theory (Sydney, 2002), vol. 2369 of Lecture Notes in Comput. Sci., Springer, Berlin, 2002, pp. 267–275.

[33] 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.