Semantic-aware Fingerprints of Symbolic Research Data

Hans-Gert Gräbe

Univ. Leipzig, Germany
graebe@informatik.uni-leipzig.de,
http://bis.informatik.uni-leipzig.de/HansGertGraebe

Abstract. One of the goals of the SymbolicData Project is to set up a navigational structure on the research data associated with the project. In 2009 we started to refactor the data and metadata along standard semantic web concepts based on the Resource Description Framework (RDF) thus opening the door to the Linked Open Data world.
One of the main metadata concepts used for navigational purposes is that of *semantic-aware fingerprints* as semantically sound invariants of the given data. We applied this principle, first used to navigate within polynomial systems data, to the data sets on polytopes and on transitive groups newly integrated with SymbolicData version 3, and also within the recompiled version of test sets from integer programming.
The RDF based representation of fingerprints allows for a unified navigation and even cross navigation within such data using the SPARQL query mechanism as a generic web service, a clear advantage compared to metadata management traditionally in use within the domain of computer algebra.
In this paper we discuss merely the conceptual background of our *fingerprinting approach* and refer to the SymbolicData wiki for more details and examples how to use that service.

Key words: semantic technology, RDF, computer algebra, metadata management, SPARQL query mechanism

1 Introduction

The section "Information Services for Mathematics" addresses a more complex target compared to the title "Mathematical Software" of this conference at large since mathematical software can be considered as part of a whole infrastructure for mathematical research. Nowadays such an infrastructure goes much beyond the classically hawked "paper and pencil" or "chalk and blackboard" claimed to be sufficient – together with access to the work of colleagues within an, nowadays also not self-evident, information and communication infrastructure – to pursue advanced mathematical research. The themes "software, services, models, and data" point to at least four dimensions to enhance the mathematical research infrastructure in the era of ubiquitous computing and increasingly important digital interconnectedness.

This paper addresses the dimension of *research data* in more detail, in particular aspects of public availability of reliable and well curated *research input data* that is important for the coherence of research questions addressed by communities and thus for the formation of specific research communities themselves.

We discuss relevant questions in the specific context of intra- and intercommunity communication within the specific research domain of *symbolic and algebraic computations* (CA) coarsely defined by the MSC 2010 classification number 68W30. We analyze the situation of public availability of research data in that area on the background of almost 20 years of experience with research data management in that domain within the SymbolicData Project [18]. We address the special challenges to small scientific communities as the CA community compared to larger ones as the whole mathematical community, that nevertheless splits into a number of CA subcommunities. These CA subcommunities are organized around special research topics and in many cases already managed to organize and consolidate their own intracommunity research infrastructures. We discuss lessons to be learned from these activities and hurdles and obstructions to generalize such experience to an intercommunity level within the CA domain.

In section 2 we develop a more detailed view on the interplay between (digital) research data and research infrastructures and discuss the situation of the mathematical digital research infrastructure compared to other sciences.

In section 3 we give a short report about SymbolicData activities in the CA domain during the years. In particular we emphasize the importance of a redesign of the SymbolicData basics during the last years towards standard semantic web concepts and the implementation of an RDF based infrastructure to manage descriptions ("fingerprints") of research data collections of different CA subcommunities and thus to open them for the Linked Open Data world.

Sections 4 and 5 are devoted to a more detailed explanation of the notion of *fingerprints* of research data and our conceptual background of data and metadata management. Further we discuss the advantages of an RDF based approach to metadata management compared to approaches traditionally in use within the CA domain.

2 Research Data and Digital Research Infrastructures

Digital change and the accelerated development of a (seemingly) universally interconnected digital universe lead to an essential reshaping of many areas of life. Also the world of scientific research is affected by these mainly technologically triggered social changes. The public availability and easy accessibility of very detailed descriptions and information about research processes leads to a strong increase of transparency and provides a basis for completely new cooperation forms whose importance for the future hardly can underestimated.

Such a development started to change research methods already in the *computer age* since the 1960th complementing established forms of intermediation of scientific results by journal papers and preprints with computer simulations and

scientific software[1] as an essentially new form of scientific knowledge production. Within the upcoming *networking age* a number of questions of scientific knowledge production have to be addressed anew. Three themes related to simulation are of particular importance: (1) the input data, (2) the simulation procedures (scientific software) and (3) the output data.

Not only with the digital universe the free availability of data for public research from each of these three thematic areas plays an important but scientific-sociologically different role:

(1) The public availability of reliable and well curated *input data* is important for the coherence of research questions addressed by the community and thus for the formation of a specific research community itself around its central research problems.
(2) The public availability of newly developed *simulation methods, procedures and techniques* is relevant for the traceability of the proposed scientific approaches and increasingly accompanies classical forms of description of scientific advancement by academic papers.
(3) The public availability of *output data* is important for the independent reproduction of results and thus of essential importance for the process of academic quality assurance.

It is in the nature of the scientific process that output data is the starting point for new research questions and thus output data mutates to input data. In most of the cases such a mutation happens not immediately but is mediated by a community-internal interpersonal transformation process that transforms the often large output data (or a whole bundle of such data) into (one or several) more compact input data adapted to the new research question(s).

Within the digital change the different scientific communities are faced with the challenge to adapt their research and communication infrastructure to these new socio-technical opportunities. Of central importance – beside a culture of public access – is the allocation of resources for such a mainly non-academic business to restructure this highly technical research infrastructure of the community and keep it running. After many years of community-driven grassroot activities of academic self-organization (e.g., ArXiv) this topic begins to move into the focus of research and political administrations at different levels and is reflected in different calls and rules at German wide (e.g., [3]) or EU level (e.g., [14, 4]).

Other scientific communities (e.g., with programs as TextGrid, DARIAH, CLARIN-PLUS) act very successful to acquire EU funding to upgrade their research data infrastructure mainly at the theme (1) level – in particular to set

[1] *Scientific software* is written to run *computer simulations* – we use this notion in an appropriate broad meaning – and if software is not used in such a way it is of less academic interest. Moreover, computer simulations often require the interplay of several *scientific packages* bundled within an *application*, hence computer simulation is the broader notion and we use it throughout this paper instead of scientific software.

up a sustainable environment for text corpora (e.g., "Deutsches Textarchiv") as the central research data form within Digital Humanities. The mathematical community is much less successful within the EU Research Infrastructures Program [14] (but see the OpenDreamKit Project [11]) and concentrates with projects as swMath [19], sagemath [15] and also this conference on the theme (2) level of sustainably available scientific software. Note that the application of the OpenDreamKit Project was successful also due to the fact that it does *not* address mathematical software as such but *successful cooperate practices* using mathematical software.

Efforts to secure a research infrastructure for mathematical data at the theme (1) or even theme (3) levels are lost in the brushwood of everlasting (for at least a decade) debates about reliable formal but semantically expressive formats as MathML or OpenMath for data resulting from calculi, that are already highly formalized – at least at an informal level – by the internal nature of the research topics themselves. The situation reminds the Tower of Babel Project, since subcommunities are digitally already well established, developed their own formalizations for their own research data at theme (1) level and apply such formalizations very successful within their intracommunity communication processes.

3 The SymbolicData Project

The SymbolicData Project is a small project initiated at the end of the 1990th as an intracommunity project in the area of *Polynomial Systems Solving* to secure a research data infrastructure at the theme (1) level built up within the EU funded PoSSo [13] and FRISCO [5] projects. It grew up from the Special Session on Benchmarking at the 1998 ISSAC conference in a situation where the research infrastructure built up within these projects – the Polynomial Systems Database – was going to break down. After the end of the projects' fundings there was neither a commonly accepted process nor dedicated resources to keep the data in a reliable, concise, sustainably and digitally accessible way. Even within the ISSAC Special Session on Benchmarking the community could not agree upon a further roadmap to advance that matter.

The SymbolicData Project was set up by a small number of volunteers not involved within the EU funded projects, but strongly interested in the public availability of this research data as reference that can be used as input data (1) for certified benchmark activities on specialized mathematical software that was written to run simulations (2) in a special domain of Algebraic Geometry. At those times almost 20 years ago most of the nowadays well established concepts and standards for storage and representation of research data did not yet exist – even the first version of XML as a generic markup standard had to be accepted by the W3C. It was Olaf Bachmann and me who developed during 1999–2002 with strong support by the Singular group concepts, tools and data structures for a structured representation and storage of this data and prepared about 500

instances from *Polynomial Systems Solving* and *Geometry Theorem Proving* to be available within this research infrastructure, see [1].

The main conceptional goal was a nontechnical one – to develop a research infrastructure that is independent of (permanent) project funding but operates based on overheads of its users. This approach was inspired by the rich experience of the Open Culture movement "business models" to run infrastructures. It was an early attempt to emphasize the advantage of an explicitly elaborated concept of a community-based solution to the "tragedy of the commons" [8] within the CA community and to apply such a concept to run a part of its research infrastructure.

Even 15 years later it remains difficult to keep the SymbolicData Project running on such a base, and for many years we concentrate our efforts to secure the sustainable public digital availability of the research input data within our collections and to develop appropriate concepts and tools to manage, search and filter this data. In 2009 we started to refactor the data along standard semantic web concepts based on the Resource Description Framework (RDF). With SymbolicData version 3 released in September 2013 we completed a redesign of the data along RDF based semantic technologies, set up a Virtuoso based RDF triple store and an SPARQL endpoint as Open Data services along Linked Data standards, and started both conceptual and practical work towards a semantic-aware Computer Algebra Social Network [7].

Since then we continued that development. On March 1, 2016, version 3.1 of the SymbolicData tools and data was released. The new release contains

- new resource descriptions ("fingerprints") of remotely available data on transitive groups (*Database for Number Fields* of Gunter Malle and Jürgen Klüners [10]) and polytopes (databases of Andreas Paffenholz [12] within the *polymake* project [6]),
- a recompiled and extended version of test sets from integer programming – work by Tim Römer (*normaliz* group [2]) –,
- an extended version of the *SDEval benchmarking environment* – work by Albert Heinle [9] – and
- a partial integration (SymbolicData People database, databases of upcoming and past conferences) of data from the Computer Algebra Social Network subproject.

Moreover, the github account `https://github.com/symbolicdata` was transformed into an organizational account and the git repo structure was redesigned better to reflect the special life-cycle requirements of the different parts and activities within SymbolicData. We provide the following repos

- *data* – the data repo with a single master branch mainly to backup recent versions of the data,
- *code* – the code directory with master and develop branches,
- *maintenance* – code chunks from different tasks and demos as best practice examples how to work with RDF based data,

- *publications* – a backup store of the LaTeX sources of SymbolicData publications,
- *web* – an extended backup store of the SymbolicData web site that provides useful code to learn how RDF based data can be presented.

The main development is coordinated within the SymbolicData *Core Team* (Hans-Gert Gräbe, Ralf Hemmecke, Albert Heinle) with direct access to the organizational account. We refer to the SymbolicData Wiki [18] for more details about the project's organization and the new release.

4 Research Data and Metadata

From the internal perspective of a research community a special aspect of every research data collection is the design of management, search and filter functionality. For this purpose data is usually enriched with metadata that collect important relevant information of the individual data records in a compact manner. We denote such metadata for an individual data record as its *fingerprint*.

Similar to a hash function a fingerprint function computes a compact metadata record (*resource description* in the RDF terminology) to each individual data record (*resource* in the RDF terminology). As with a hash function one can use the fingerprints to (almost) distinguish different data records within the given collection and to match new records with given ones. But there is an essential difference between (classical) hash functions and well designed fingerprints: fingerprint functions exploit not only the textual representation of the data record as meaningless syntactical character string but convey semantically important information or even compute such information from the string representation. Fingerprints are in this sense *semantic-aware* and can even be designed in such a way that they map ambiguities in the textual representation of records (e.g., polynomial systems given in different polynomial orders and even in different variable sets) to *semantic invariants*.

The design of appropriate fingerprint signatures is an important *intracommunity* activity to structure its own research data collections. Such fingerprint signatures are also very useful for the *intercommunity* usage of research data collections, since they allow to navigate within the (foreign) research data collection without presupposing the full knowledge of the "general nonsense" of the target research domain, i.e., the informal background knowledge required freely to navigate as scientist in that domain. Hence well designed fingerprint signatures are to be considered also as a first class service of a special research community to a wider audience to inspect their research data collections without using the community-internal tools to access the resources themselves.

5 Working with Semantic-aware Fingerprints

Usually the research data collections (resources in the RDF terminology) of a certain community are stored in a specially designed community-internal format,

often as plain text (e.g., the Normaliz Collection [2]), in a special XML notation (e.g., the Polymake Collection [6]) or as SQL database (e.g., the Database for Number Fields [10]). Such formats usually employ special formal semantics agreed within the community as an effective way to store domain specific input and output data and used by commonly developed tools with appropriate parsing functionality.

Usually such formats are extended to store research metadata, i.e., fingerprints or *resource descriptions* in the RDF terminology, together with the research data. This has one benefit and two drawbacks:

- *Benefit:* A fingerprint can be computed immediately by the commonly used tools or with their slight extension, and can be stored with the resource itself.
- *First Drawback:* Metadata unfold its full expressiveness only if one can search and navigate within it. A storage together with the resource itself implies high extraction costs for metadata navigation and access to the research data collection.
- *Second Drawback:* The very different formats prevent an easy combination of metadata from different communities and even from different sources.

The first drawback can be addressed if the metadata are extracted into a database – either a central one or delivered with the tools for local use – and the commonly used intracommunity tools provide search and navigational functionality within that metadata representation. Such an approach based on a web interface was realized for the *Database for Number Fields* [10] and a tool integration based on a Mongo-DB for the *Polymake Database* [6]. But such a solution has two further drawbacks:

- *Drawback 1a:* The search and navigational functionality is not or only in a restricted way adapted for machine-readable interaction and thus cannot be integrated into more comprehensive search and navigational processes.
- *Drawback 1b:* The search and navigational functionality can't be adapted by the user for its own needs.

A general solution that avoids these drawbacks proposes to extract the metadata information from the resource data and to transform it into RDF. RDF – the Resource Description Framework – is the conceptual basis of Linked Open Data as a worldwide distributed database that can be globally queried and navigated using the SPARQL query language in a similar unified way as SQL allows to navigate in local relational databases.

We applied this approach, first used within the SymbolicData Project to navigate within polynomial systems data, to the data sets on polytopes and on transitive groups newly integrated with SymbolicData version 3, and also within the recompiled version of test sets from integer programming. We store these fingerprints in our RDF data store [16] thus allowing for a unified navigation and even cross navigation within such data using the SPARQL query mechanism as a generic Web service provided by our SPARQL endpoint [17].

We refer to the SymbolicData wiki [18] for detailed information and examples how to use that service.

References

1. Bachmann, O., Gräbe, H.-G. : The SymbolicData Project – Towards an Electronic Repository of Tools and Data for Benchmarks of Computer Algebra Software. Reports on Computer Algebra 27 (2000), Centre for Computer Algebra, University of Kaiserslautern.
2. Bruns, W., Ichim, B., Römer. T., Sieg, R., Söger, C.: Normaliz. Algorithms for Rational Cones and Affine Monoids. https://www.normaliz.uni-osnabrueck.de. [2016-03-08]
3. DFG verabschiedet Leitlinien zum Umgang mit Forschungsdaten. DFG-Magazin "Information für die Wissenschaft" Nr. 66 (2015). http://www.dfg.de/foerderung/info_wissenschaft/2015/info_wissenschaft_15_66/index.html. [2016-05-07]
4. Strategy Report on Research Infrastructures. Roadmap 2016. Published by the European Strategy Forum for Research Infrastructures (ESFRI), Brüssel (2016). http://www.esfri.eu/roadmap-2016. [2016-03-16]
5. FRISCO – A Framework for Integrated Symbolic/Numeric Computation. (1996–1999). http://www.nag.co.uk/projects/FRISCO.html. [2016-02-19]
6. Gawrilow, E., Joswig, M.: Polymake: a Framework for Analyzing Convex Polytopes. In: Kalai, G., Ziegler, G.M. (eds.), Polytopes – Combinatorics and Computation (Oberwolfach, 1997), pp. 43–73, DMV Sem., 29, Birkhäuser, Basel (2000).
7. Gräbe, H.-G., Johanning, S., Nareike, A.: The SymbolicData Project – Towards a Computer Algebra Social Network. In: Workshop and Work in Progress Papers at CICM 2014, CEUR-WS.org, vol. 1186 (2014).
8. Hardin, G.: The Tragedy of the Commons. Science 162 (3859), pp. 1243–1248 (1968). doi:10.1126/science.162.3859.1243.
9. Heinle, A., Levandovskyy, V.: The SDEval Benchmarking Toolkit. ACM Communications in Computer Algebra, vol. 49.1, pp. 1–10 (2015).
10. Klüners, J., Malle, G.: A Database for Number Fields. http://galoisdb.math.uni-paderborn.de/. [2016-03-08]
11. OpenDreamKit: Open Digital Research Environment Toolkit for the Advancement of Mathematics. http://opendreamkit.org/, http://cordis.europa.eu/project/rcn/198334_en.html. [2016-03-16]
12. Paffenholz, A.: Polytope Database. http://www.mathematik.tu-darmstadt.de/~paffenholz/data/. [2016-03-08]
13. The PoSSo Project. Polynomial Systems Solving – ESPRIT III BRA 6846. (1992–1995).
14. Research Infrastructures, including e-Infrastructures. http://ec.europa.eu/programmes/horizon2020/en/h2020-section/research-infrastructures-including-e-infrastructures. [2016-03-16]
15. The SageMath Project. http://www.sagemath.org/. [2016-03-16]
16. The SymbolicData RDF Data Store. http://symbolicdata.org/Data. [2016-03-15]
17. The SymbolicData SPARQL Endpoint. http://symbolicdata.org:8890/sparql. [2016-02-19]
18. The SymbolicData Project Wiki. http://wiki.symbolicdata.org. [2016-03-13]
19. swMATH – a new Information Service for Mathematical Software. http://www.swmath.org/. [2016-03-07]