The SymbolicData Project in a Computer Algebra Social Network Perspective. Some Architectural Considerations

Hans-Gert Gräbe

Leipzig University, Leipzig, Germany
graebe@informatik.uni-leipzig.de

Abstract. On March 1, 2016, version 3.1 of the SymbolicData database was released. With the new release the SymbolicData Project offers new, recompiled and extended data and introduces an adjusted git repo structure. The *main goal* of the new release was directed towards an architectural redesign of the CASN subproject with the following main tasks:

- Enlarge the SymbolicData People database both in the number of instances and with valuable additional information for author disambiguation – one of the great challenges of all catalogue systems.
- Strengthen the notion of a local CASN node maintained by a CA substructure as basis of an upcoming federated network of such nodes.
- Reorganize the CASN data collected so far according to these adjusted conceptional basis using established semantic web best practices.

In this paper we explain the conceptional background of such a redesign in more detail and put it in the context of some architectural considerations.

1 Introduction

The allocation of resources for a sustainably available research infrastructure seems to be a great challenge in particular to smaller scientific communities. The SymbolicData Project witnesses the peaks and troughs of such efforts. It grew up from the Special Session on Benchmarking at the 1998 ISSAC conference in a situation where the research infrastructure built up within the PoSSo [15] and FRISCO [4] 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 road-map to advance that matter.

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 within the Singular group concepts, tools and data structures for a structured representation and storage of this data and prepared a large number (about 500) of instances from *Polynomial Systems Solving* and *Geometry Theorem Proving* to be available within this research infrastructure.

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" [7] 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.

During the last ten years with Open Access, Open Data and the emerging semantic web the general understanding of the importance of such community-based efforts to develop common research infrastructures matured. This development was accompanied with conceptual, technological and architectural standardization processes that had also impact on the development of concepts and data structures within the SymbolicData Project. 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.

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

- new resource descriptions ("fingerprints" in the notion of [6]) of remotely available data on transitive groups (*Database for Number Fields* of Gunter Malle and Jürgen Klüners [10]) and polytopes (databases of Andreas Paffenholz [13] within the *polymake* project [5]),
- a recompiled and extended version of test sets from integer programming – work by Tim Römer (*normaliz* group [1]),
- an extended version of the *SDEval benchmarking environment* – work by Albert Heinle [8] and
- a partial integration (SymbolicData People database, databases of upcoming and past conferences) of data from the CASN – 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 data,
- *code* – code directory with master and develop branches,
- *maintenance* – code chunks from different tasks and demos 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 old repo *symbolicdata* is deprecated and was removed from the github account, so please adjust your local repo structure. 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 [22] for more details about the new release.

All changes reported so far are mentionable advances of the SymbolicData Project. Nevertheless the *main goal* of the new release was directed towards an architectural redesign of the CASN subproject with the following main tasks:

- Enlarge the SymbolicData People database both in the number of instances and with valuable additional information for author disambiguation – one of the great challenges of all catalogue systems, see, e.g., VIAF [23].
- Strengthen the notion of a local CASN node maintained by a CA substructure as basis of an upcoming federated network of such nodes – in a first step such a node exposes valuable information in RDF as files for download and use in a local RDF store.
- Reorganize the CASN data collected so far according to these adjusted conceptional basis using established semantic web best practices.

In this paper we explain the conceptional background of such a redesign in more detail and put it in the context of some architectural considerations.

2 Semantic Web as Web of Data

> The *Semantic Web* is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF).
> According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". The term was coined by Tim Berners-Lee for a web of data that can be processed by machines. ... In 2006, Berners-Lee and colleagues stated that: "This simple idea ... remains largely unrealized". (Source: [16])

In this statement optimism and pessimism are close together – the optimistic perspective concerns the potential of the ongoing development and the pessimistic one the time horizon of that development to mature.

The core of the vision of a "semantic web as web of data" is an architectural change of the web itself. Web 1.0 started and matured as web of interlinked presentations, in the first time as hyperlinked static HTML pages connected via the HTTP hypertext protocol. Modern web engines deliver dynamically generated HTML code (in many cases with additional Javascript driven visualization code) and provide interlinking of specially prepared information pieces between different presentation layers – the "content". Such web applications are typically designed along a 3-tier architecture with data layer, business or model layer and presentation layer. All modern web frameworks and web design pattern as, e.g., the Model-View-Controller (MVC) or the Presentation-Abstraction-Control (PAC) pattern, support and presuppose such a layered architecture. The main drawback of such an architecture is its tight coupling within a local web server and its direction towards information supply *ready for use*, i.e., provision of *interpreted data*, mainly controlled by the needs and world perception of the information *providers.*

Unfortunately, information reveals usefulness often enough in new, unexpected contexts, not foreseen and even not predictable by the information providers. This is the starting point of the vision of the semantic web – use existing protocols, in particular HTTP, to connect the data layers directly and to combine "uninterpreted" data from different sources in machine readable form under the control (and interpretation) of the *data user*, not the data provider.

3 Data and Software Architectures

A deeper analysis of the problem reveals that there are no (interesting) "uninterpreted" data. A bit stream of data exchanged between two computers can, e.g., represent a picture, if the computer recognizes the image format (analyzing the file name extension or detecting an appropriate pattern prefix in the bit stream – a standard to be agreed upon *before* exchanging the first picture), and the computer can "interpret" the bit stream starting a special algorithm to render the picture. The picture itself is "reinterpreted" once more by the user who called that picture with a certain goal in mind. Such multilevel interpretation processes are ubiquitous in the web and it is one of the difficult problems to harmonize such interpretational frames within social communication rooms.

Thus one of the central challenges of digital communication is the coordination of conceptual and notational conventions on a level of detail that can be algorithmically processed by digital machines. Nowadays the most successful semantic web projects address domains with sophisticated taxonomies and systematics well established already in a predigital era that have "only" to be fully formalized for digital use.

Such requirements led to a completely new job profile within computer science – in addition to *software architects*, who are responsible for a reasonable program architecture within a single application or application system, nowadays we have *data architects* to design powerful comprehensive data models ("ontologies") for interpretational frames that facilitate cooperation between applications from different domains or, more precisely, between the users of those applications.

4 Semantic Web as Web of People

The systematic development of such ontologies can be supported by tools but is in its core a socially triggered process. Thus from the perspective of a data architect the web is not only a web of data but also a web of people driven by different interests and goals using and producing this data. The "web of data" metaphor masks not only the users of this data and their goals but also the coordination processes required to use the "web of data" in a sound way.

RDF as language concept and framework offers its strength in such a domain – due to its generic concept RDF is appropriate to be used for modelling processes at both the data and the metadata level and can be used to describe not only data (resources in the RDF terminology) but also data structures (resource descriptions in the RDF terminology), metadata structures (languages or meta-meta models in other terminologies) and even more elaborated abstractions.

5 The `MathHub.info` Project

During the last years semantic web concepts largely influenced also the research infrastructure of science. Within the domain of mathematics there are big projects on the way as WDML [14, 26] or EuDML [3] and also smaller ones as the semantification of the MSC2010 index [12] or the swMATH project [18].

Beyond such domain-specific efforts there is not only *Google Scholar* but there are more combined e-science projects with global scope on the way to support and restructure scholarly communication using semantic technologies as, e.g., VIVO[1], VIAF[2] or OCLC[3].

It is a great challenge to smaller scientific communities to adopt such developments for their own scientific communication processes and to join forces with other scientific communities to get own requirements publicly recognized.

[1] `http://vivoweb.org`. "... VIVO supports recording, editing, searching, browsing and visualizing scholarly activity. VIVO encourages research discovery, expert finding, network analysis and assessment of research impact. ...". [24]

[2] "The Virtual International Authority File (VIAF) is an international service designed to provide convenient access to the world's major name authority files. ...". [23]

[3] "A global library cooperative that provides shared technology services, original research and community programs for its membership and the library community at large". `http://oclc.org`

Michael Kohlhase reported in two contributions [11, 9] to the session *Projects and Surveys* at CICM 2012 and CICM 2014 about efforts to "develop a general framework – the Planetary system – for social semantic portals that support users in interacting with STEM[4] documents ..." [11] and started with `MathHub.info` to build up such a research infrastructure.

Although being a very interesting approach to enrich established STEM technologies (in particular LaTeX and arXiv) semantically it seems that it could not attract a broader audience so far. Kohlhase described the design goal of the system in [9] as follows:

> `MathHub.info` must satisfy two conflicting goals: On the one hand, it must be so generic that it is open to all logics and implementations; on the other hand, it must be aware of the semantics of the formalized content so that it can offer meaningful services.

The experience within the SymbolicData Project indicates that such a design goal is very ambitious and requires a strategy of *social* communication on a long run to get its results accepted by the target community.

Running a research infrastructure and providing reliable access to it is a cross cutting concern [2] orthogonal to the core research concerns of any special interest groups. It is "nice to have" and "hard to get" even if it has plenty of advantages that Kohlhase described in [9] in such a way:

> We claim that `MathHub.info` will resolve two major bottlenecks in the current state of the art. It will provide a permanent archiving solution that not all systems and user communities can afford to maintain separately. And it will establish a standardized and open library format that serves as a catalyst for comparison and thus evolution of systems.

From an architectural point of view `MathHub.info` tries to integrate elements of a local working place environment and access elements to a global infrastructure that has yet to emerge. The main focus, see fig. 1 in [9], lies on the functionality of a local working place environment. The structural aspects of the formation of a global data infrastructure remain hidden in the boxes "MMT" and "GitLab" in that figure. Thus the design focuses on *software architectural* aspects and does not address the (social) requirements of *data architecture* building processes.

Ten years of semantic web experience indicate that data architectural aspects are in the core of the building processes of the "web of data" and require social organization in such a way that the (yet emerging) global data structure fits with a large number of different software frameworks and architecture models nowadays in use at local sites.

Lets end this section of critical analysis with a clear statement about the valuable contribution of Kohlhase's work that he summarized in [9] as follows:

> Already now, `MathHub.info` is unique in its class in that it gives a unified interface to multiple theorem prover libraries together with linguistic

[4] STEM is a shortcut for "Science, Technology, Engineering, Mathematics".

> and educational resources. Now that the ground work has been laid, we anticipate the rapid integration of new semantic services, editing support and new content.

Kohlhase provides an elaborated system of tools and data that addresses the needs of the theorem prover CA subcommunity and thus is a good candidate for a node within an emerging Computer Algebra Social Network (CASN). Although we propose a roadmap towards such a CASN and realized first steps, such an idea goes far beyond the attempts and goals of the SymbolicData Project and requires common efforts of a much larger community.

6 Extending the SymbolicData data store

During the last years the SymbolicData Project adjusted its focus to address also more general technical and social aspects of a semantically enriched research infrastructure within the domain of Computer Algebra based on RDF for representation of intercommunity and relational information.

Such a change of the focus had its impact on several earlier design decisions of the data store itself. Enlarging the database of SymbolicData we gained the following experience:

- The CA community consists of several subcommunities with own concepts, notational conventions, semantic-aware tools and established communication structures.

  There is no need to duplicate such structures but to support the subcommunities to enrich semantically these communication processes.
- We provide structural metadata (“fingerprints” in the notion of [6]) of the different data sets at our central RDF store [20] but not necessarily duplicate the data itself.

  Thus we rely on sustainably available research infrastructures of CA subcommunities and restrict our activities to a central search and filter service on the metadata level to find and identify data. This service is based on a generic semantic web concept, the SPARQL query language, and can be accessed via our SPARQL endpoint [21].

  We applied this principle to the newly integrated data sets on polytopes and on transitive groups and also within the recompiled version of Test Sets from Integer Programming. Data are hosted by the *polymake* group [5], within the *Database for Number Fields* [10] and by the *normaliz* group [1].
- RDF is a useful and meanwhile well established standard for metadata and relational information, but there is no need and one cannot expect from CA subcommunities to give up established notational conventions in favour of RDF or XML markup.

  Semantic-aware tools of the subcommunities are well tuned for the established notational conventions, and representation of data in a different format requires additional transformation effort to use it.

Moreover, one can use MathML or OpenMath standards and tools for the casual exchange of data. Note nevertheless, that the notational conventions of a subcommunity use many shortcuts that are valid only in a special interpretational frame (the “general nonsense” of the field, well known to the specialists) that is hard and probably useless to formalize, since practical use of data from a special field requires a minimum of semantic-awareness of the user itself.

7 About the CASN Architecture

The CASN subproject tries to embed aspects of the maintenance of the SymbolicData data store into a more general process of formation of a semantically enriched social network of academic communication within the CA community in the sense of a “web of people” mentioned above.

A first roadmap towards such a CASN and our experimental setting was described in [6] and developed further during the last years. We try not to “reinvent the wheel” but to address the already existing “CA memory” – a huge number of very loosely related web pages about conferences, meetings, working groups, projects, private and public repositories, private and public mailing lists etc. Hence the main focus towards CASN is to develop a framework based on modern semantic technologies for a decentralized network that increases the awareness of the different parts of that already existing “CA network”.

As a coarse architectural concept to establish such a network we propose

- to operate a central RDF store with SPARQL endpoint providing the full bandwidth of Linked Open Data services and
- to convert nodes of the “CA memory” into CASN nodes providing part of their data in structured RDF format for easy access and exchange.

SymbolicData version 3.1 is a first step in that direction since

- several data from the formerly separate CASN RDF store are now integrated with the SymbolicData main RDF store [20] and
- the experimental setting of the semantic support of the website of the German Fachgruppe [25] was reorganized as a first CASN node.

CASN Integration into the SymbolicData RDF Store

The CASN Integration into the SymbolicData RDF Store covers the following topics:

- The RDF store provides information about scientific activities of people mainly extracted from conference announcements. The personal information is stored as instances of the RDF type `foaf:Person` with (as subset of) keys `foaf:name`, `foaf:homepage` and `sd:affiliation` (a literal). Due to privacy reasons we do not provide `foaf:mbox` (email) values.

This list is steadily enlarged and used as URI reference for reports about different activities (invited speakers, conference organizers etc.). As of March 2016 the SymbolicData People database contains 1036 `foaf:Person` entries from the CA scientific community that can be explored via the SymbolicData SPARQL endpoint [21] and also using the *CA People Finder* at the SymbolicData info page [19].

In August 2014 we compiled a first alignment of the SymbolicData People database with the ZBMath author database to evalute the potential of a community-based author disambiguation and could resolve 348 matchings out of 678 persons. The result is available as `ZBMathPeople` RDF graph in our database.

- The RDF store provides information about upcoming CA conferences from several mailing lists, usually up to 20 entries with references to the SymbolicData People database.

  The information is extracted via SPARQL query and displayed both in the Wordpress based site of the German Fachgruppe [25] and at the SymbolicData info page [19].
- The RDF store provides information about past CA conferences with references to the SymbolicData People database about speakers and organizers (as far as available).

  Most of the entries were moved from the upcoming CA conferences list to that list for archiving purposes and aligned according to their archival status. As of March 2016 there are 139 records of past CA conferences.

  A short set of information is extracted via SPARQL query and displayed at the SymbolicData info page [19].
- As another feature we started to provide semantic annotations to a subset of news (beyond conference announcements) posted on several mailing lists as instances of RDF type `sioc:BlogPost`. We operate a special mailing list `sd-announce` with archive and forward interesting news to that archive for URI reference if the original mailing list is not archived.

  Such an annotation contains an excerpt of that message in a standardized way that can be explored at the SymbolicData info page [19]. The concept can easily be extended to the concept of an CASN news channel.

The CASN node of the German Fachgruppe

The CASN node of the German Fachgruppe contains

- a list of (extended) FOAF-Profiles used to render, e.g., the page

  `http://www.fachgruppe-computeralgebra.de/fachgruppenleitung/`,
- lists of current and former members of the board of the German Fachgruppe,
- (not up to date) information about German CA working groups,
- standardized information about the SPP 1489 projects with references to the SymbolicData People database,

– keyword enriched information about scientific publications in the CA-Rundbrief of the German Fachgruppe using the `dcterms` ontology,
– a survey on successfully defended CA dissertations in the scope of the Fachgruppe (a joint effort with the CA-Rundbrief) and
– a (also not up to date) list of CA books.

The data is available in RDF format for direct download from a web directory

`http://www.fachgruppe-computeralgebra.de/rdf`

as part of an an upcoming global RDF data structure and can be harvested and processed within a local RDF store. The data is used in different PHP-based presentations that are summarized at

`http://www.fachgruppe-computeralgebra.de/symbolicdata/`

Best practice code how to embed such information into a Wordpress based website using the *EasyRDF* PHP library and also the code to operate the SymbolicData info website [19] is available from our git repos *maintenance* and *web*.

The SymbolicData CASN Node

We also operate a rudimentary SymbolicData CASN Node at the publicly accessible directory

`http://symbolicdata.org/rdf/`.

In the subdirectory `Conferences` we provide detailed information about five CA conferences (SPP annual meeting in Bad Boll 2014 and the CICM conferences 2012–2015) as proof of concept of standardized detailed conference reports using the *Semantic Web Conference Ontology* [17]. The records provide information about the general venue, programme committes, tracks and talks of the conference.

For the CICM conferences we exploited and transformed the publicly available XML-based representation developed by Serge Autexier to render the conference pages at `http://cicm-conference.org`.

References

1. 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]
2. Cross Cutting Concern. From Wikipedia, the Free Encyclopedia.
`https://en.wikipedia.org/wiki/Cross-cutting_concern`.
3. EuDML. The European Digital Mathematical Library. `https://eudml.org/` [2014-09-23]
4. FRISCO – A Framework for Integrated Symbolic/Numeric Computation.
`http://www.nag.co.uk/projects/FRISCO.html` [2014-02-19]

5. 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).
6. 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).
7. Hardin, G.: The Tragedy of the Commons. Science 162 (3859), pp. 1243–1248 (1968). doi:10.1126/science.162.3859.1243.
8. Heinle, A., Levandovskyy, V.: The SDEval Benchmarking Toolkit. ACM Communications in Computer Algebra, vol. 49.1, pp. 1–10 (2015).
9. Iancu, M., Jucovschi, C., Kohlhase, M., Wiesing, T.: System Description: MathHub.info. In: Watt, S.M., Davenport, J.H., Sexton, A.P., Sojka, P., Urban, J. (eds.), Intelligent Computer Mathematics. LNCS vol. 8543, pp. 431–434 (2014).
10. Klüners, J., Malle, G.: A Database for Number Fields. http://galoisdb.math.uni-paderborn.de/ [2016-03-08]
11. Kohlhase, M.: The Planetary Project: Towards eMath3.0. In: Jeuring, J., Campbell, J.A., Carette, J., Dos Reis, G., Sojka, P., Wenzel, M., Sorge, V. (eds.), Intelligent Computer Mathematics. LNCS vol. 7362, pp. 448–452 (2012).
12. Lange, C., Ion, P., Dimou, A., Bratsas, C., Corneli, J., Sperber, W., Kohlhase, M., Antoniou, I.: Reimplementing the Mathematics Subject Classification (MSC) as a Linked Open Dataset. In: Jeuring, J., Campbell, J.A., Carette, J., Dos Reis, G., Sojka, P., Wenzel, M., Sorge, V. (eds.), Intelligent Computer Mathematics. LNCS vol. 7362, pp. 458-462 (2012).
13. Paffenholz, A.: Polytope Database. http://www.mathematik.tu-darmstadt.de/~paffenholz/data/ [2016-03-08]
14. Pitman, J., Lynch, C.: Planning a 21st Century Global Library for Mathematics Research. Notices of the AMS, August 2014.
15. The PoSSo Project. http://posso.dm.unipi.it/ [2014-02-19]
16. Semantic Web. From Wikipedia, the Free Encyclopedia. https://en.wikipedia.org/wiki/Semantic_Web [2016-03-07]
17. The Semantic Web Conference Ontology. http://data.semanticweb.org/ns/swc/swc_2009-05-09.html.
18. swMATH – a new Information Service for Mathematical Software. http://www.swmath.org/ [2016-03-07]
19. The SymbolicData Info Page. http://symbolicdata.org/info [2016-03-08]
20. The SymbolicData RDF Data Store. http://symbolicdata.org/Data [2016-03-15]
21. The SymbolicData SPARQL Endpoint. http://symbolicdata.org:8890/sparql [2014-02-19]
22. The SymbolicData Project Wiki. http://wiki.symbolicdata.org [2014-09-13]
23. VIAF – the Virtual International Authority File. http://viaf.org/ [2016-03-07]
24. VIVO – an Open Source Software and Ontology for Representing Scholarship. https://wiki.duraspace.org/display/VIVO/VIVO [2016-03-07]
25. Website of the German Fachgruppe Computeralgebra. http://www.fachgruppe-computeralgebra.de/ [2014-03-06]
26. World Digital Mathematics Library (WDML). http://www.mathunion.org/ceic/wdml/ [2014-09-23]