The SymbolicData Project

Hans-Gert Gräbe, Andreas Nareike, and Simon Johanning

Abstract We report about a major reengineering of the SymbolicData project [17] and its redesign according to Linked Data principles and RDF technologies that proved to be powerful within modern semantic web approaches. During that redesign the focus of the project changed from a mere data store towards the vision of a Computer Algebra Social Network that aims at technical support of the intercommunity communication needs between different CA subcommunities. The redesign was supported by Saxonian E-Science Initiative grants for A. Nareike and S. Johanning. A first milestone was the release of version 3 of SymbolicData in September 2013.

1 Introduction

A powerful digital research *infrastructure* becomes increasingly important in today's networked and interlinked world. This includes digital support for dissemination of new papers, the refereeing process, conference submissions, and scientific communication within communities. Services such as MathSciNet, arXiv.org, EasyChair.org, or bibsonomy.org have been established and their usefulness is acknowledged by the larger scientific community.

Practical efforts to build up digital support for research infrastructures can be traced back at least until the beginning of the 1990th in different areas of science. E.g., in Computer Algebra (CA) the EU funded projects PoSSo [14] and FRISCO [8] aimed at developing such a research infrastructure in the area of Polynomial Systems Solving.

Although being a challenging engineering task such efforts are rarely acknowledged by the reputational processes of science, and hence are left to the

Hans-Gert Gräbe · Andreas Nareike · Simon Johanning
Universität Leipzig, Germany, e-mail: `(graebe|nareike|johanning)@informatik.uni-leipzig.de`

casual engagement of volunteers. Different to the infrastructural processes in the "large science" for small academic communities as, e.g., the CA community, it is hard to find funding for special infrastructure projects and they have to organize the development of their *intracommunity* communication infrastructure in a different, more distributed way.

Open Source culture offers plenty of experience of how to substitute centrally organized projects by decentralized networked structures and indeed the new focus of SymbolicData version 3 [17] is that of an intercommunity project, providing not only reliable access to data for testing and benchmarking purposes but also technical support for interlinking between different CA subcommunities.

In this paper we present our *e-science environment engineering* approach to redesign SymbolicData in the indicated way with scarce resources (less than a man-year). In such a setting we had to focus on communication with other existing activities and on real engineering aspects along a roadmap predetermined by the international development of e-science standards, in particular by the Dublin Core Metadata Initiative [4], the Virtual International Authority File [20], and the corresponding German GND activities [9]. Our main goal was to produce and operate a valuable piece of software. In the scope of this conference we can only give a short report about our special e-science engineering task with focus on interoperability with the large e-science world.

We assume the reader to be familiar with the basic terminology in that area, in particular with the *Resource Description and Access* (RDA) concepts and standards [15].

2 Background

SymbolicData has been part of CA infrastructural efforts for almost 15 years. It grew out of a Special Session on Benchmarking at the 1998 ISSAC conference, where the participants were faced with a typical situation: within the PoSSo and FRISCO projects volunteers compiled a large database of Polynomial Systems with the goal to make it publicly available for testing and benchmarking of algorithms. At the end of the project people switched to other tasks and it became more and more problematic to access the data. Even worse, badly cloned copies of the data were disseminated and after a while it was even hard to decide, e.g., what exactly 'Katsura-5' means – is it about the example from the well-known series with 5 variables $y_1, \ldots, y_5$ or with 6 variables $x_0, \ldots, x_5$?

The SymbolicData Project started in 1999 on that basis to build a reliable and sustainably available reference of Polynomial Systems data, to extend and update it, to collect metainformation about the records, and also to develop tools to manage the data and to set up and run testing and benchmarking computations on the data. A first implementation with data from Polynomial Systems Solving and Geometry Theorem Proving was realized by Olaf Bachmann and Hans-Gert Gräbe in 1999 and 2000. Data from other fields were added later on by the CoCoA and Singular teams, V. Levandovskyy, and R. Hemmecke, and the Web site `symbolicdata.org` sponsored by the *Fachgruppe Computeralgebra* went online in 2005.

Later on the project joined forces with the Agile Knowledge Engineering and Semantic Web (AKSW) Group at Leipzig University [1] to strongly refactor the data along standard Semantic Web concepts based on the Resource Description Framework (RDF). These efforts were endorsed in 2012/13 by a Saxonian E-Science Initiative grant [6] for A. Nareike and in 2014 for S. Johanning. A first milestone of this reengineering project was the release of SymbolicData version 3 in September 2013.

3 The SymbolicData Infrastructure

Our resources (examples for testing, profiling and benchmarking software and algorithms from different areas of symbolic computation) are publicly available in XML markup, meta data in RDF notation both from a public git repo at `http://github.org/symbolicdata`, and from an OntoWiki [13] based RDF data store at `http://symbolicdata.org/Data`. Moreover, we offer a SPARQL endpoint at `http://symbolicdata.org:8890/sparql` to explore the data by standard Linked Data methods.

The website operates on a standardized installation using an Apache web server to deliver the data, the Virtuoso RDF data store [21] as data backend, a SPARQL endpoint and (optionally) OntoWiki to explore, display and edit the data. This installation can easily be rolled out at a local site to support local testing, profiling and benchmarking. For details we refer to the SymbolicData wiki [17].

The distribution offers also tools for integration with a local compute environment as, e.g., provided by Sagemath [16] – the Python based *SDEval package* [10] by Albert Heinle offers a JUnit like framework to set up, run, log, monitor, and interrupt testing and benchmarking computations, and the *SDSage package* [12] by Andreas Nareike provides a showcase for SymbolicData integration with the Sagemath compute environment.

We follow a development process along the Integration-Manager-Workflow Model. This makes it easy to join forces with the SymbolicData team: Fork the repo to your github account, start development and send a pull request to the Integration Manager if you think you produced something worth to be integrated into the upstream master branch. Even if your contribution is not pulled to the upstream, people can use it, since they can pull it from your to their github repo. This allows even for agile common small feature development – a widely practised way to advance projects hosted at `github.com`.

You are encouraged early to start a discussion about your plans and regularly report your progress on the SymbolicData mailing list.

4 SymbolicData at Work

Preparing SymbolicData version 3 we decided to strengthen the SymbolicData part that is *not* involved with Polynomial Systems Solving. These efforts led to a more consequent distinction between data (owned and maintained by different CA subcommunities) and meta data. Such a distinction is well supported by RDF design principles – the Resource Description Framework is about *description* of *resources*, represented by (globally unique) *resource identifiers* (URIs). We follow the Linked Data best practise to provide URIs accessible by the HTTP internet protocol and to deliver a valuable part of structured information about that resource upon HTTP request.

4.1 Resources

Currently the SymbolicData data collection contains resources from the areas of Polynomial Systems Solving (390 records, 633 configurations), Free Algebras (83 records), G-Algebras (8 records), Geometry Proof Schemes (297 records) and Test Sets from Integer Programming (28 records). These resources are stored per file in a flat XSchema based XML syntax using well established intracommunity syntaxes for the internal data.

Such a concept is not restricted to centrally managed resources, but can easily be extended to other data stores on the web that are operated by different CA subcommunities and offer a minimum of Linked Data facilities. There are draft versions of resource descriptions about Fano Polytopes (8630 records) and Birkhoff Polytopes (5399 records) hosted by Andreas Paffenholz and about Transitive Groups (3605 records) from the Database for Number Fields of Jürgen Klüners and Gunter Malle that point to external resources.

4.2 Metadata and Resource Descriptions

It was one of the great visions of the SymbolicData Project to collect not only benchmark and testing data but also valuable background information about the records in the database as, e.g., information about papers, people, history, systems etc. concerned with the examples in our collection. Within the redesign we developed a general concept of the `sd:Annotation` RDF class to store background information in a unified way. We use that concept in particular to relate bibliographical entries of type `sd:Reference` to different data records.

The management of bibliographical references was completely redesigned with SymbolicData version 3 exploiting RDF and the established Dublin Core ontology [5] to represent bibliographical information in a way that is queryable by standard means and tools. On the other hand, we strongly reduced the part of information about bibliographical references kept inside SymbolicData to a toe-hold since there are comprehensive bibliographical stores available on the web that provide all required information and there is no need to spend efforts to collect such information twice, even if the foreign repositories are not (yet) querable by RDF techniques but provide only access to permanent URLs and to local search engines to extract the required information.

4.3 Towards a Computer Algebra Social Network

From the five stars to be assigned to a Linked Data project according to Tim Berners-Lee's classification [3] SymbolicData earned four stars so far (for offering data in interoperable RDF format on the web and providing a SPARQL querable RDF triple store). For the fifth star one has to build up stable semantic relations to foreign knowledge bases and thus become part of the Linked Open Data Cloud [11].

Much of such interrelation, e.g., a list of interoperability references for people, software and bibliographical data with the Zentralblatt, is on the way. Moreover, we joined forces with the efforts of the board of the German Fachgruppe to store and provide information about people and groups working on CA topics at their new Wordpress driven web site [7]. We developed a first prototype [18] to store this information in RDF format, to extract it by means of SPARQL queries and to view it on the web site using the Wordpress shortcode mechanism via a special Wordpress plugin. We apply this technique to maintain information about

- German working groups in CA,
- projects within the SPP 1489 priority program,
- graduate theses in CA and
- upcoming conferences

at this site [7].

The vision of a Computer Algebra Social Network goes far beyond that: Set up and run within the CA community a semantic aware Facebook like Social Network that provides RDF based tools in such a way that interested people can easily contribute about all topics around Computer Algebra. This sounds quite visionary, but it is in no way utopic. We started to operate a very first prototypical node `http://symbolicdata.org/xodx/` [2] that realizes the challenging concept of a *Distributed Semantic Social Network* [19].

References

1. Agile Knowledge Engineering and Semantic Web (AKSW) Working Group at Leipzig University, http://aksw.org/About.html
2. Arndt, N.: Xodx: A basic DSSN node implementation, http://aksw.org/Projects/Xodx.html
3. Berners-Lee, T.: 5 stars for Open Data. http://5stardata.info/
4. The Dublin Core Metadata Initiative, http://dublincore.org/
5. DCMI Metadata Terms, http://dublincore.org/documents/dcmi-terms/
6. Das eScience-Forschungsnetzwerk Sachsen, http://www.escience-sachsen.de
7. Website of the German Fachgruppe Computeralgebra, http://www.fachgruppe-computeralgebra.de/
8. FRISCO – A Framework for Integrated Symbolic/Numeric Computation, http://www.nag.co.uk/projects/FRISCO.html
9. Gemeinsame Normdatei (GND), http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html
10. Heinle, A.: The SDEval framework. http://symbolicdata.org/wiki/SDEval
11. Linked Data, http://en.wikipedia.org/wiki/Linked_data
12. Nareike, A.: The SDSage package, http://symbolicdata.org/wiki/PolynomialSystems.Sage
13. OntoWiki – A tool providing support for agile, distributed knowledge engineering scenarios, http://aksw.org/Projects/OntoWiki.html
14. The PoSSo Project, http://posso.dm.unipi.it/
15. The Joint Steering Committee for Development of RDA, http://www.rda-jsc.org/rda.html
16. Sage – a free open-source mathematics software system, http://www.sagemath.org
17. The SymbolicData Project Wiki, http://symbolicdata.org
18. The SymbolicData CASN RDF triple store, http://symbolicdata.org/casn/
19. Tramp, S. Frischmuth, P., Ermilov, T., Shekarpour, S., Auer, S.: An Architecture of a Distributed Semantic Social Network. Semantic Web Journal (2012), http://www.semantic-web-journal.net/sites/default/files/swj201_4.pdf
20. The Virtual International Authority File, http://www.oclc.org/viaf.en.html
21. Virtuoso Open-Source Edition, http://virtuoso.openlinksw.com/