The SYMBOLICDATA Project – a Community Driven Project for the CA Community

H.-G. Gräbe

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

1 Introduction

A central phenomenon of the emerging digital age is the increasing importance of a sustainably and reliably available *digital interconnection infrastructure* for many areas of every day life. This distinguishes the digital age from the computer age that focused on penetration of every day life with *compute power* rather than interconnectedness. With "ubiquitous computing" such a penetration with compute power reached a high level of saturation, but is in no way at its end as is demonstrated by the development of modern sensor and actor systems as "cyber-physical systems" and its applications within "industry 4.0".

During the last years the sensibility to the importance of investments also into a modern digital research infrastructure remarkably increased. It was continuously discussed on the "big" stage of research politics between different stakeholders, see e.g., [11, 12]. The disposition to invest into the development of an appropriate digital infrastructure heavily depends on the visibility of the demand, whereas the demand develops with the productiveness of the available infrastructure – a typical chicken-and-egg problem, that can only be addressed in the socio-technical context of a problem-aware community. Such a community should have a good understanding of the importance of the advancement of its own research infrastructure and the ability to set up a socio-communicative process to coordinate the development of its own demands *and* activities in the desired direction.

Digital infrastructures are not only well suited to exchange research data and make it publicly available, but also proved valuable as technical basis of "social networks" to promote such socio-communicative coordination processes. Nowadays in many cases different channels and means are used for these purposes, but it is due time to combine conceptually and also in practice both aspects of a research infrastructure.

With the advancement of the SYMBOLICDATA Project towards a Computer Algebra Social Network (CASN) we pursued such a concept in a specific context for several years. We started to investigate questions of intra- and intercommunity communication in correlation with practical aspects of the community driven development of a decentrally organized, distributed semantic-aware digital research

infrastructure within the specific research domain of *symbolic and algebraic computations* (CA) coarsely defined by the MSC 2010 classification code 68W30 – a medium sized scientific community, that splits into a number of subcommunities. These CA subcommunities are organized around special research topics and in many cases already managed to organize and consolidate their own intracommunity digital research infrastructures.

In our talk we address relevant questions, observations, and experience of our endeavor to develop and provide technical means to support the emergence of a digital research infrastructure on the intercommunity level. We discuss lessons to be learned from these activities and hurdles and obstructions to generalize intracommunity experience to an intercommunity level within the CA domain.

We propose to deploy a special RDF-based architecture of *CASN nodes* operated by different CA subcommunities and CA groups along the rules of the Linked Open Data Cloud [8]. To ensure interoperability, this should be accompanied by a strong social intercommunity communication process to develop a common *data architecture* of data models and its ontological standards of representation based on well established semantic web concepts and using standard semantic web technology.

2 The SYMBOLICDATA Project as Community Project

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 [10] and FRISCO [3] 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.

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 conceptual 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. 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 [8], and started both conceptual and practical work towards a semantic-aware Computer Algebra Social Network [5].

In March 2016 version 3.1 of the SYMBOLICDATA tools and data was released. On the level of research tools and data the new release contains new resource descriptions ("fingerprints" in the notion of [5]) of remotely available data on transitive groups (*Database for Number Fields* of Gunter Malle and Jürgen Klüners [7]) and polytopes (databases of Andreas Paffenholz [9] within the *polymake* project [4]), a recompiled and extended version of test sets from integer programming – work by Tim Römer (*normaliz* group [2]) – and an extended version of the *SDEval benchmarking environment* – work by Albert Heinle [6].

The main development is coordinated within the SYMBOLICDATA *Core Team* (Hans-Gert Gräbe, Ralf Hemmecke, Albert Heinle) with direct access to our public github account `https://github.com/symbolicdata`. We refer to the SYMBOLICDATA Wiki [13] for more details about the project and the new release.

3 The CA Community and its Subcommunities

During the last years the SYMBOLICDATA Project adjusted its focus to address 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 [5]) of the different data sets at our central RDF store 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 operated via our SPARQL endpoint.

- 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 favor of RDF or XML markup for their primary sources.

4 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 (social) "web of people".

A first roadmap towards such a CASN and our experimental setting was described in [5] and developed further during the last years. We try not to "reinvent the wheel" but to address step by step 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".

We realized that this network itself is an "overlay network" that connects a greater number of research networks of individuals around special topics with own lightweight research infrastructures. It is an interesting challenge for semantic concepts to support the requirements of intercommunity communication to exchange semantic content on different levels and different levels of detail.

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 within the SYMBOLICDATA main RDF store and the experimental setting of the semantic support of the website of the German Fachgruppe [14] was reorganized as a first CASN node.

References

[1] O. Bachmann, H.-G. Gräbe: 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] W. Bruns, B. Ichim, T. Römer, R. Sieg, C. Söger: Normaliz. Algorithms for Rational Cones and Affine Monoids. https://www.normaliz.uni-osnabrueck.de. [2016-03-08]

[3] FRISCO – A Framework for Integrated Symbolic/Numeric Computation. (1996–1999). http://www.nag.co.uk/projects/FRISCO.html. [2016-02-19]

[4] E. Gawrilow, M. Joswig: Polymake: a Framework for Analyzing Convex Polytopes. In: G. Kalai, G.M. Ziegler (eds.), Polytopes – Combinatorics and Computation (Oberwolfach, 1997), pp. 43–73, DMV Sem., 29, Birkhäuser, Basel (2000).

[5] H.-G. Gräbe, S. Johanning, A. Nareike: 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).

[6] A. Heinle, V. Levandovskyy: The SDEval Benchmarking Toolkit. ACM Communications in Computer Algebra, vol. 49.1, pp. 1–10 (2015).

[7] J. Klüners, G. Malle: A Database for Number Fields. http://galoisdb.math.uni-paderborn.de/. [2016-03-08]

[8] The Linked Open Data Cloud. http://lod-cloud.net/. [2016-05-20]

[9] A. Paffenholz: Polytope Database. http://www.mathematik.tu-darmstadt.de/~paffenholz/data/. [2016-03-08]

[10] The PoSSo Project. Polynomial Systems Solving – ESPRIT III BRA 6846. (1992–1995). http://research.cs.ncl.ac.uk/cabernet/www.laas.research.ec.org/esp-syn/text/6846.html. [2016-03-16]

[11] Research Infrastructures, including e-Infrastructures. http://ec.europa.eu/programmes/horizon2020/en/h2020-section/research-infrastructures-including-e-infrastructures. [2016-03-16]

[12] 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]

[13] The SYMBOLICDATA Project. http://wiki.symbolicdata.org. [2016-05-20]

[14] Website of the German Fachgruppe Computeralgebra. http://www.fachgruppe-computeralgebra.de/. [2016-03-06]