The SymbolicData Project – Maturing the Computer Algebra Social Network Perspective

H.-G. Gräbe (Universität Leipzig)

graebe@informatik.uni-leipzig.de

Published in *Computeralgebra Rundbrief* 59, October 2016, pp 17–21.

Introduction

In [1, 2, 3, 4] we described the roots and goals of the SymbolicData Project and also basic Linked Data and RDF principles. In this note we report about advances of the project during the last year. Our main efforts were directed to strengthen and consolidate the Computer Algebra Social Network (CASN) part.

In March 2016 we released SymbolicData version 3.1 including

- new data from different CA subcommunities,
- a reorganized git repo structure,
- close integration of the CASN part into the main project and
- a set of examples based on the *EasyRDF* PHP library and the Bootstrap web framework to show how to set up web presentations of the data in a very simple way.

The focus of the SymbolicData project moved from a project mainly centered around activities of the German CA Fachgruppe to a broader international scope. This opening started with a presentation of the project in the Work in Progress section of the 2014 CICM conference. It was furthermore discussed with the SIGSAM chair at the ACA 2015 conference in Kalamata, where *Hans-Gert Gräbe* gave a presentation in combination with a hands on session on SymbolicData[1]. The contacts were deepened by *Albert Heinle* and *Victor Levandovskyy* with a paper [5] about the SDEval framework in the *ACM Communications in Computer Algebra*.

In 2016 we had the opportunity to present and discuss various aspects of SymbolicData and more general perspectives of a digital research infrastructure for the CA community at the session *Information Services for Mathematics: Software, Services, Models, and Data* at ICMS-2016 in Berlin (organized by Wolfram Sperber, FIZ Karlsruhe, and Michael Kohlhase, Jacobs University Bremen) and at the session *Information Services for Mathematical Software, Models, and Research Data* at ACA-2016 in Kassel (organized by Hans-Gert Gräbe, University of Leipzig, Albert Heinle, University of Waterloo, and Wolfram Sperber, FIZ Karlsruhe). A more detailed report on these discussions will appear in a forthcoming publication.

Enlarging the SymbolicData Database

With SymbolicData 3.1 we consolidated the integration of several data sources into our main data and metadata collection that were available so far only in a draft version.

With that integration the conceptual design of the SymbolicData database changed from a data store to a metadata store – the new data collections provide (only) metadata information about the core research data that is hosted and maintained in a separate remote data store by a CA subcommunity. The CA subcommunity provides expertise to maintain the research data in a semantically correct way and provides metadata in “raw form”. The SymbolicData team collects that metadata, transforms it into RDF and prepares it for search and filter processing.

Note that such a design change was enabled by the consistent transformation of the metadata collected so far along RDF principles during preparation of SymbolicData version 3. RDF requires a strong distinction between data and metadata whereas the data structure design of CA research data usually handles metadata as “data extension” and stores it together with the primary data in a common file.

Our new conceptual approach along that lines supports the formation of an *interlinked distributed research data infrastructure* within the CA community and between its several subcommunities. We realized that in many cases CA subcommunities (in particular subcommunities developing specialized research CA software systems) have well established research data infrastructures with no need to be duplicated, but interlinking these infrastructures is yet a challenge.

In the following subsections we describe the advances at the “data frontier” in more detail.

Fingerprints for Polytopes and Groups

In [2] we announced draft versions of RDF based resource descriptions (i.e., metadata or *fingerprints*, see [3]) of Fano and Birkhoff polytopes collected by Andreas Paffenholz as part of the *polymake* data store and of transitive groups from the *Database for Number Fields* collected and stored in a similar way by Jürgen Klüners and Gunter Malle.

The draft versions were extracted from the primary data sources and transformed into RDF based *fingerprints* by *Andreas Nareike* in 2013. The metadata was

[1] See our publications and presentations overview at http://wiki.symbolicdata.org/Publications for more details.

provided as part of the primary data in different formats within the respective remote collections. The derived metadata is now integrated into the SymbolicData main database. Each such metadata record contains a link to the corresponding data record within the remotely maintained research data store provided by the respective CA subcommunity.

As main advantage such separated metadata can be queried in a common, uniform and well established way using the SPARQL query language. This is another W3C web standard with many tools and concepts, which were mainly developed for a performant management of big data given in the RDF semantic web format and for integration with other applications.

SPARQL plays a similar role for querying the worldwide distributed and interlinked semantic web data store as SQL plays for querying local databases. Using that technology one can navigate within such data, restructure it for efficient search or identify a given example within the database. We refer to our wiki[2] for more background and some example queries.

Transforming Test Sets into the Normaliz Format

In a similar, we transformed and enlarged the data on integer programming – the SymbolicData test sets collection.

The old SymbolicData test sets collection was compiled by *Raymond Hemmecke* several years ago along the former SymbolicData rules – develop a data model, an XML-binding for data storage to represent this special data type, and an RDF ontology for metadata (fingerprints and maintenance information).

The former exclusive usage of XML-bindings for data representations was inspired by the success and wide usage of XML as a unified way to represent data in other application areas at that time, in particular influenced by the upcoming MathML standard. Meanwhile XML is much less prominent for exchange practices of structured data, and within the redesign of SymbolicData we decided to accept and use data also in other formats.

Such a decision also was inspired by the observation that specialized CA software comes with well a defined input data format, and within subcommunities using a common software the data are stored and exchanged in just that format. Hence for such a subcommunity (as, e.g., the polytope subcommunity around *polymake* or the integer programming subcommunity around *Normaliz*) there is no need to develop another standard for data exchange – such a standard would hardly be accepted. Note that the situation is different within the polynomial systems solving subcommunity since there exist several major software systems with different input formats, as *Singular*, *Macaulay2*, *Magma*, *GB* or *CoCoA* – to name only the most important ones.

For the new test sets collection we use and store data in the *Normaliz* exchange format and thus prepared the data in a similar way as for polytopes and transitive groups. This work was done by *Tim Römer* who transformed also the "legacy" test sets into the new format. All content written in the old format was cleaned up from the repos and the web pages were adjusted.

Towards a Computer Algebra Social Network

Motivation

All parties want to have a powerful digital research infrastructure, but they are rarely willing or able to invest in it. It is a complex social challenge to organize active goal-oriented cooperation in such an area outside the scientific reputation process. We learned over the years not only to concentrate on the collection of scientific *data* but also on structured and semantically enriched information about the scientific and social *processes* to produce this data.

Several years ago the SymbolicData Project extended its scope to analyze and support the exchange of such information in a structured way. Our vision is a distributed and tool based network of semantic aware nodes corresponding to the (small and big) nodes of the real CA research network. Such a *Computer Algebra Social Network* (CASN) should be a semantically enriched digital infrastructure for a social network of scientific research and scientific researchers within the Computer Algebra community and its several subcommunities similar to other social networks as, e.g., Facebook.

Note that the starting point for such a CASN is at least in two ways different from the Facebook starting point. First, there is no Marc Zuckerberg nor such an amount of money to push the project. Second, there is already a digital "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. The CASN design has to take such a diversity into account and develop a decentralized solution based on modern semantic technologies. This solution must increase the awareness of the different parts of that already existing "CA network" and supports the *exploration* of that network to get useful deep results in an easy manner.

CASN Nodes

For a proper CASN design it is essential to exploit the potential of concepts, tools and standards of the fast growing distributed Linked Open Data (LOD) Cloud[3]. Pascal Hitzler emphasizes the importance of such a coordinated conceptual approach to set up an interoperably interlinked digital universe, since "with the omnipresence and availability of data from different times, locations, perspectives, topics, cultures, resoluti-

[2] See, e.g., `http://wiki.symbolicdata.org/MoreQueries`.

[3] `http://lod-cloud.net`

ons, qualities, and so forth, *exploration* becomes an additional (4th) paradigm of science” [6].

As a first step towards a digital network within the CA community capable to *explore* social and scientific relations

- we operate the RDF based SymbolicData main data store together with its SPARQL endpoint [8] to query centrally maintained data and
- propose to convert other nodes of the “CA memory” into CASN nodes that provide part of their data in structured RDF format.

RDF principles neither require such nodes to be uniformly structured nor running on big web resources. LOD sources are self-explanatory by design and its structure can be explored with appropriate RDF tools by interested third parties at run time to prepare to fetch the information in a structured way. Hence efforts to present and explore data within such a CASN network can be shared in a wide scope between data providers and data consumers.

In a first version such a node can be even only a directory with valuable RDF files publicly accessible in the web as provided by the CASN sample node[4] of the SymbolicData Project. As proof of concept we provide detailed information about five CA conferences in the subdirectory `Conferences` using the (meanwhile outdated) *Semantic Web Conference Ontology*[5].

The CASN node of the German CA Fachgruppe[6] is designed along a more advanced concept. During the revision of concepts towards SymbolicData 3.1 we consequently redesigned this data to form a proper CASN node with publicly accessible but locally maintained RDF sources of (almost) all structured information displayed on the web site of the German CA Fachgruppe. This information is explored by a special plugin and rendered in its Wordpress based web presentation[7]. Hence one can explore both the “pure” information in standard RDF notation to embed it into third party web workflows as *interlinked data* and in the “old fashion” as *hyperlinked text*. Note that the technical realization is unpretentious – the RDF data is stored as plain files in RDF/XML format in the CASN node and the plugin uses the *EasyRDF* PHP library and the Wordpress Shortcode mechanism for rendering. No advanced technique as RDF store or SPARQL endpoint has to be set up. The code is mirrored as best practice example in our *maintenance* Git repo.

CA Conferences

As another service within the CASN we maintain a list of *Upcoming Conferences*. The data about conferences is extracted from several sources, transformed into RDF format and delivered by our main SPARQL endpoint [8]. This information is used by the German CA Fachgruppe on one hand to present an online list of upcoming conferences and on the other hand to generate the conference announcement section of the printed version of their CA Rundbrief. The RDF database contains more advanced information about conferences as, e.g., submission deadlines or program committees.

We run this service in a draft version for several years already and compiled from it a list of (at the moment 166) *Past Conferences*. In summer 2016 this data was enhanced with additional data about past conferences supplied by the SIGSAM web team and extended by a *Conference Series* concept from the SIGSAM collection. The SIGSAM collection provides structured information about such conference series (description and publication rules) in an (almost) unstructured way that was transformed to structured RDF using predicates `sd:description` and `sd:publicationRules`. Not to duplicate information without reason we use the standard predicate `rdfs:seeAlso` to link with the corresponding part in the SIGSAM conference series web page for additional information.

The SymbolicData People Database

The concept of the Unique Resource Identifier (URI) as part of the RDF standard provides a generic way to disambiguate people and artifacts. More precisely, each such URI, considered as *digital identity*, is the entry point from the real world to the digital universe, and any statement within the digital universe can be followed and traced back using digital technology only up to such (combinations of) URIs. URIs are bound to real world entities by more complex socio-political and technical “agreements”. To shape politically such “agreements” is the real core of digital privacy.

The use of URIs provides an easy way to assign digital facts to special digital identities and thus solve the *disambiguation problem* – a great problem in the text oriented “hyperlinked universe” that required powerful text mining so far. One of the great challenges to academic content providers within the transformation of their digital universes is *author disambiguation*. Such disambiguation is required to, e.g., assign URIs of publications to the correct author URIs. Most of the academic content providers come up with own solutions for their own universe, i.e., for the provider’s internal data collection that counts as its main “capital”. Interoperability between providers remains a great challenge since it requires to interlink data sources that are very private from a business point of view. While this Gordian knot is hard to cut from a provider’s position, a comparatively small scientific community could solve that interoperability challenge by a common effort – develop its own People Database, i.e., its own URI system for people and provide dictionaries to the part of the URI systems of the

[4] `http://symbolicdata.org/rdf/`

[5] `http://data.semanticweb.org/ns/swc/swc_2009-05-09.html`

[6] `http://www.fachgruppe-computeralgebra.de/rdf/`

[7] `http://www.fachgruppe-computeralgebra.de/`

different providers relevant to their academic scope.

This is the goal of the SymbolicData People Database for the CA community. As one of the benefits of such a disambiguation one can track reputation and merits more precisely querying the whole SymbolicData database or even interlinking it with other RDF based sources within the Linked Open Data Cloud. Moreover, people within the CASN can systematically provide and update information about their own scientific activities.

Currently the SymbolicData People Database contains more than 1200 entries, i.e., digital identities of scientists that are active in the area of Computer Algebra. These URIs were mainly extracted from program committee lists of different conferences or (in a restricted scope) from lists of authors of accepted papers.

As standard information we provide personal information as instance of `foaf:Person` with (a 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, program committees etc.) in other parts of our CASN database.

As proof of concept we aligned our URIs in a common task with the “Zentralblatt” with their author disambiguation system and produced more than 300 `sd:hasZBMathAuthorID` matches. This work was done in 2014 on an early version of the SymbolicData People Database and can be queried from our RDF store, too. In a near future we plan to update that alignment with “Zentralblatt”. The concept can easily be extended to other content providers (in particular to the ACM people database or the MathSciNet author disambiguation system) that are interested in such a cooperation.

The CA Dissertations Project

The CA Rundbrief of the German CA Fachgruppe maintains a section with reports about dissertations in Computer Algebra finished in working groups within the Fachgruppe. We made the metadata available also in RDF within the CASN node of the Fachgruppe and display it at their web site. Within the discussions with SIGSAM in summer 2016 we realized that there is a large data pool of similar information collected by SIGSAM for years that could be integrated into a common database of dissertations in Computer Algebra. For the moment we moved the existing RDF data about dissertations to the SymbolicData main data store and aligned the presentation in the web site of the German CA Fachgruppe accordingly.

The CA Systems Project

In summer 2016 we also intensively discussed perspectives of the swMATH project [7]. In particular we considered ways to popularize it to a larger audience (within the CA community) and discussed to what extend RDF principles and LOD alignment could support such a popularization. We agreed that it would be helpful to represent a core part of the swMATH metadata in RDF, provide URIs with a consistent naming scheme, and publish this data as Linked Open Dataset to achieve better visibility within the semantic web community. Such a metadata extraction also makes the alignment with other overviews on CA systems as, e.g., the one maintained by SIGSAM, much easier.

A first prototypical draft version of such an RDF based overview on *CA systems* extracted from the swMATH database was compiled during our discussions in summer 2016 and is available in our RDF store. We also set up a prototypical view on that data within the SymbolicData info pages[8].

Additionally, we discussed whether the swMATH data model has to be redesigned better to reflect subtleties as the relation between CA systems and CA packages or different versions of the same system. All these questions require much deeper analysis. Since RDF can be used in a consistent way to express modeling aspects a Linked Open Dataset as just described could support also such a discussion.

Advances in the SymbolicData Infrastructure

In October 2015 we converted our main git repo[9] to an organizational account. With SymbolicData version 3.1 we reorganized our git repo structure and set up several new repos with different maintenance rules.

- *data* – the data repo with a single master branch mainly to backup recent versions of data, no versioning,
- *code* – code directory with master and develop branches, under versioning,
- *maintenance* – code chunks from different tasks and demos how to work with RDF based data, no versioning,
- *publications* – as a backup store of the LaTeX sources of SymbolicData publications, only master branch, no versioning,
- *web* – as an extended backup store of the SymbolicData web site that provides useful code to learn how RDF based data can be presented in the web.

The main development is coordinated by the SymbolicData *Core Team* (Hans-Gert Gräbe, Ralf Hemmecke, Albert Heinle) with direct access to the organizational account.

The repos *maintenance* and *web* are intended to show best practice code for using the RDF based data

[8] http://wiki.symbolicdata.org/info

[9] https://github.com/symbolicdata

of the SymbolicData project. In particular, the *maintenance* repo contains a mirror of the Wordpress plugin code used by the German CA Fachgruppe and the transformation code developed by Andreas Nareike in 2013 for polytopes and groups databases. To use the code you may fork the repo, but there is almost no reason to pull code back. If you have a valuable contribution please contact the Core Team to discuss how that contribution can be added to the project.

The repo *data* is mainly for backup purposes. If you plan to add valuable data to the project please contact the Core Team to discuss how that contribution can be added. We provide help to put the data in an appropriate Linked Open Data format.

The repo *publications* is used mainly for reference and not intended for public additions. We provide LaTeX sources of our papers and slides and also information about the review processes of our work since reviews provide many valuable suggestions for the further development of our project. The repo *code* contains several coding subprojects concerning SymbolicData tools for various purposes.

Literatur

[1] Hans-Gert Gräbe, Simon Johanning, Andreas Nareike. The SymbolicData Project – from Data Store to Computer Algebra Social Network. *Computeralgebra-Rundbrief*, 55:22–26, 2014.

[2] Hans-Gert Gräbe, Simon Johanning, Andreas Nareike. The SymbolicData Project – Towards a Computer Algebra Social Network. *Workshop and Work in Progress Papers at CICM 2014*. CEUR-WS.org, vol. 1186, 2014.

[3] Hans-Gert Gräbe. Semantic-aware Fingerprints of Symbolic Research Data. In Gert-Martin Greuel, Thorsten Koch, Peter Paule, Andrew Sommese (Eds.). *Mathematical Software – ICMS 2016*. Volume 9725 of Lecture Notes in Computer Science, page 411–418, 2016.

[4] Hans-Gert Gräbe. The SymbolicData Project – a Community Driven Project for the CA Community. Talk given at the ACA 2016 Session "Information services for mathematical software, models, and research data." `http://symbolicdata.org/Papers/aca-16.pdf`. [2016-09-11]

[5] Albert Heinle, Viktor Levandovskyy. The SDEval Benchmarking Toolkit. *ACM Communications in Computer Algebra*, 49.1:1–10, 2015.

[6] Pascal Hitzler, Krzysztof Janowicz. Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, 4.3:233–235, 2013.

[7] swMATH – an Information Service for Mathematical Software. `http://swmath.org/about_contact`. [2016-09-16]

[8] The SymbolicData SPARQL Endpoint. `http://symbolicdata.org:8890/sparql`. [2016-09-11]