The SymbolicData Project – from Data Store to Computer Algebra Social Network

H.-G. Gräbe, S. Johanning, A. Nareike
(Universität Leipzig)

graebe@informatik.uni-leipzig.de
nareike@informatik.uni-leipzig.de
simonjohanning@googlemail.com

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. For mathematics in the large the vision of a *21st Century Global Library for Mathematics Research* (GDML) [2] matured during the last years.

Smaller academic communities, e. g. the Computer Algebra (CA) community, are challenged to organize their *intracommunity* communication infrastructure in a similar way. Infrastructural efforts are rarely acknowledged by the reputational processes of science, however, and are hence left to the casual engagement of volunteers unless led by leading-edge scientists of the community.

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 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 explain the technological basics of an RDF based semantic web of mathematics and discuss social interlinking to the *vision in large* of an emerging GDML.

RDF Basics

We refer to [6] for an overview of the goal, aims, concepts, infrastructure and history of the SymbolicData project, in particular for a more detailed motivation to use RDF concepts and Linked Data principles. The use of such principles and notations for the metadata of our data store is one of the core advances with version 3 of SymbolicData. In this section we give a brief overview of such concepts.

RDF is a data model which stores information as *triples*

```
s p o .
```

which are considered a *sentence* with subject s, predicate p and object o. Subjects and predicates are URIs (Uniform Resource Identifiers) while objects (or 'values') can be a URI or a literal (a string in quotes). There are shortcut notations of RDF, e. g. RDF-XML, JSON, or Turtle, and plenty of tools and parsers for the different formats.

As an example we consider the resource at

```
http://symbolicdata.org/XMLResources/
   IntPS/Czapor-86c.xml
```

that represents the XML record about the polynomial system

$$x^2 + y\,z\,a + x\,d + g$$
$$y^2 + x\,z\,b + y\,e + h$$
$$z^2 + x\,y\,c + z\,f + k$$

known as the *Czapor-86c* example. Such a polynomial system can be interpreted differently as polynomial ideal in different polynomial rings (see below).

We are going to store properties of the interpretation of that polynomial system as ideal

$$I \subset \mathbb{Q}[x, y, z, a, b, c, d, e, f, g, h, k]$$

(an *ideal configuration*, associated with that record) as new RDF subject. First, a new URI has to be assigned to such a new subject (using a sound naming scheme, not discussed here):

```
<http://symbolicdata.org/Data/Ideal/
   Czapor-86c.Flat>
```

or `sdideal:Czapor-86c.Flat` for short.

Now we can assign metainformation to that subject. Some of these properties can be extracted directly from the XML resource, others have to be calculated. Here is the record in Turtle notation:

```
sdideal:Czapor-86c.Flat a sd:Ideal ;
rdfs:comment "Flat variant of Czapor-86c" ;
  sd:hasDegreeList "3,3,3" ;
  sd:hasLengthsList "4,4,4" ;
  sd:relatedPolynomialSystem
    sdpol:Czapor-86c ;
  sd:hasVariables
    "x,y,z,a,b,c,d,e,f,g,h,k" .
```

This record contains 8 triples in Turtle shortcut notation. Since all triples share the same subject (the first line), Turtle compacts the notation by separating property–value pairs to the same subject with a semicolon. The `sd:`, `sdp:`, `sdideal:`, `sdpol:` and `sdxml:` prefixes are abbreviations for name-space prefixes[1].

The configuration above refers to the Polynomial System `sdpol:Czapor-86c`, described by the RDF record

```
sdpol:Czapor-86c
  a sd:IntegerPolynomialSystem ;
  sd:createdAt "1999-03-26" ;
  sd:createdBy sdp:Graebe_HG ;
  sd:relatedXMLResource
    sdxml:Czapor-86c.xml .
```

In other words, the *Czapor-86c.Flat* configuration is derived from the 'true' *Czapor-86c* example, the ideal

$$I' \subset S' = \mathbb{Q}(a,b,c,d,e,f,g,h,k)[x,y,z]$$

generated by the same polynomials in a different polynomial ring S'. Geometrically the latter represents a complete intersection of three (generic affine) quadrics over the field of rational functions in the given parameters. There is also a record on that configuration:

```
sdideal:Czapor-86c a sd:Ideal ;
  sd:createdAt "1999-03-26" ;
  sd:createdBy sdp:Graebe_HG ;
  sd:hasDegreeList "2,2,2" ;
  sd:hasLengthsList "4,4,4" ;
  sd:hasDegree "8"^^xsd:integer ;
  sd:hasDimension "0"^^xsd:integer ;
  sd:hasParameters "a,b,c,d,e,f,g,h,k" ;
  sd:parameterize sdideal:Czapor-86c.Flat ;
  sd:hasVariables "x,y,z" .
```

Czapor-86c is derived from the *Czapor-86c.Flat* configuration by parameterization with respect to the given parameters (see below for details). Note that the degree lists of *Czapor-86c.Flat* and *Czapor-86c* are different. The record contains some more metainformation: S'/I' is zero dimensional and has degree 8.

To operate on RDF data the databases (called *RDF Graphs*) have to be uploaded into an *RDF triple store*. SymbolicData uses the Virtuoso triple store [18] that provides also a SPARQL endpoint [14] to query the data, see [6] or our wiki [16] for more background information.

Modelling Polynomial Systems

RDF provides language tools to express metainformation in an interoperably searchable way that have to be applied in a *domain-specific way* to model topics from CA subcommunities. We explain such aspects of SymbolicData modelling again on the topic of Polynomial Systems. Similar considerations are required to model any other part of the SymbolicData database and the SymbolicData wiki [16] gives details about modelling other data (Free Algebras, G-Algebras, Geometry Proof Schemes etc.).

Polynomial Systems XML resources store lists of polynomials in distributive normal form with integer coefficients together with a complete list of variables (and, for modular systems, the modular base domain $GF(p)$). Hence even modular polynomial systems can be semantically considered as set $F = \{f_1, \dots, f_s\}$ of polynomials in $S = \mathbb{Z}[x_1, \dots, x_n]$ in the indeterminates $x_1, \dots, x_n$ listed in the record.

Polynomial Systems Solving considers polynomial systems in different contexts. We have already shown how to construct the *Czapor-86c* ideal from the *Czapor-86c.Flat* resource. This is an example for a standard kind of interpretation where we divide indeterminates $x_1, \dots, x_n$ into disjoint subsets $u_1, \dots, u_k$ and $z_1, \dots, z_m$ and consider the ideal

$$I' \subseteq S' = R(u_1, \dots, u_k)[z_1, \dots, z_m]$$

generated by the images of $f_1, \dots, f_s$ in S' over the base coefficient field R. Here $u_1, \dots, u_k$ are considered as parameters and $z_1, \dots, z_m$ as variables. R is usually the field $\mathbb{Q}$ of rationals or a modular field $GF(p)$ (other settings are possible). S has the universal property that the canonical map on the indeterminates extends to a ring homomorphism $S \to S'$ in a unique way. We call such an interpretation of a Polynomial System resource as ideal generators in a polynomial rings S' *(ideal) configuration*. (Such configurations can be derived not only from Polynomial System resources but from other configurations as well.)

While SymbolicData version 2 would store each new configuration as a new resource, version 3 foresees (polynomial-time) transformations to express their relation to the basic resource.

For example, the `sd:homogenize` transformation derives a new configuration by homogenization (with respect to standard grading). Given the configuration F in S' and a new variable h we generate the homogenized polynomials $F^h = \{f_1^h, \dots, f_s^h\}$ in

$$S'' = R(u_1, \dots, u_k)[z_1, \dots, z_m, h]$$

and the ideal I'' generated by F^h in S''. There is a natural ring homomorphism $\phi : S'' \to S'$ mapping $h \to 1$ and the polynomials $f_1^h, \dots, f_s^h$ are called the *pull-back polynomials* of $f_1, \dots, f_s$ with respect to ϕ. Note that the pull-back ideal $\phi^{-1}(I')$ contains the pull-back polynomials but is not necessarily generated by them.

The transformations `sd:flatten`, `sd:parameterize` and `sd:substitute` are defined by similar concepts. Compared to former SymbolicData versions the only restriction is that semantic-aware tools (that 'know' what a polynomial is) are required to generate configurations from given basic ones. We believe

[1] Note that for more clarity we use a sloppy notation here that is not fully compliant with the RDF notational standards.

this is not a real restriction since for serious computations on Polynomial Systems semantic-aware tools are required in any case. Such tools also have to provide a Polynomial Systems parser to input the basic XML examples. With your favorite CA software being aware of polynomial semantics it should be easy to implement the transformation modes required to obtain the different configurations. As a proof-of-concept, A. Nareike compiled the *sdsage package* [10] to be integrated with the Sagemath system [12].

Navigation within the Polynomial Systems Data

Special semantic knowledge is also required for navigation and identification of data. This topic is particularly important for intercommunity communication since one cannot expect researchers to be familiar with common practices of the different subcommunity. For a case-in-point let us again turn to Polynomial Systems Solving.

It is one of the challenges to check whether a Polynomial System configuration obtained from an external source is contained in the database, since the 'same' configuration may be given by polynomials with different variable sets and in different term orders. Thus for navigational purposes *fingerprints* of Polynomial System configurations are required that are independent of variable names and term orders. For a polynomial $0 \neq f \in S' = R(u_1, \ldots, u_k)[z_1, \ldots, z_m]$, invariants may be derived from the set $T(f)$ of terms. Every such polynomial has a distributive normal representation

$$f = \sum_{\alpha \in \mathbb{N}^m} c_\alpha \cdot z^\alpha, \qquad \begin{array}{l} c_\alpha \in R(u_1, \ldots, u_k), \\ z^\alpha = z_1^{\alpha_1} \cdot \ldots \cdot z_m^{\alpha_m}, \end{array}$$

and $T(f) = \{z^\alpha : c_\alpha \neq 0\}$ is independent of the term order (but not of the variable names). There are two invariants that are well-defined for f regardless of variable names and orders – the number $|T(f)|$ of terms (the *length* of the polynomial f) and the pattern of the total degrees $(\deg(z^\alpha) : c_\alpha \neq 0)$ of the terms in $T(f)$. In particular, for $0 \neq f$ the maximum degree $\deg(f) = \max(\deg(z^\alpha) : c_\alpha \neq 0)$ is well-defined.

We use ordered lists of polynomial lengths and of maximum degrees as fingerprints of configurations and provide them precompiled as part of the metadata for a given configuration. Such a fingerprint can easily be computed by almost all semantic-aware tools. While configurations with different fingerprints are definitely distinct, there can be different examples with the same fingerprint. Although such fingerprints could be refined there was no need so far, since the examples with equal fingerprints are rare and can easily be inspected by hand.

Knowledge Frames and Social Framing

The SYMBOLICDATA project can be seen as part of the worldwide efforts towards a *World Digital Mathematical Library* (WDML). The most viable contributions on that way are the EuDML project [3], funded by the European Union during 2010–2013, and the WDML project [19], triggered by the US National Research Council and heavily supported by the IMU – the International Mathematical Union – and the Alfred P. Sloan Foundation. There are other activities on the way in that direction, in particular by Wolfram Research Inc., developing *Wolfram Alpha* as one of the most mature online resources of structured mathematical knowledge.

One of the big challenges within these projects is the question "How to structure and represent mathematical knowledge?" Such a question is not only (and probably not even in the first plan) a question about mathematics but also about mathematicians, their common efforts, social relations, and the way how they organise common work. Minsky's well established concept of *knowledge frames* as "artificial intelligence data structure used to divide knowledge into substructures by representing 'stereotyped situations'" [4] suggests that such "stereotyped situations", in our case different research contexts with complex social structures and interrelations, play an important role in the way how knowledge structures emerge and evolve. A. and M. Kohlhase [8] emphasized the importance of such practices for mathematicians and formalized them in the (slightly different) notions of *theory graphs* and *framing* (the latter considered as 'reframing' in a Minsky inspired terminology [7]).

Computer Algebra at the border of mathematics and computer science played an important and special role during any computer triggered technological breakthrough in mathematics in the last decades. It is a first class challenge to the CA community to organize the changes towards a GDML for their own community and to contribute to the changes at large.

The setting of (not only) the SYMBOLICDATA Project to address the needs of the CA community and their severe subcommunities – considered as a "major social frame" with many intimately related "smaller social frames" – is a first class playground

- to study and explore own cooperate contexts and practices under the special aspects and formalization requirements of modern semantic web technologies,
- to develop and communicate best practices along the IMU recommendations [11],
- and to advance towards a better understanding of their requirements and consequences,

since we have researchers well educated both in mathematics and computer science and the scientific CA community is small enough to allow for social relations in a less institutionalized way.

Towards CASN — a Computer Algebra Social Network

Such a challenge meets other visions of the SYMBOLICDATA Project, in particluar to collect not only benchmark and testing data but also valuable background information about the records in the database, e. g. information about papers, people, history, systems, concerned with the examples in our collection. RDF is particularly suited for this, for it provides both the concept of typeless URIs within a typed world to point to resources of different types in a uniform way and it allows linking to foreign URIs in other databases to build up a semantic network with many nodes where the node at `symbolicdata.org` is only one in the multitude of nodes of such a (distributed) Computer Algebra Social Network.

There are plenty of activities towards such an 'e-science world', in particular by the *Association of German Libraries* [5], by the *Zentralblatt Mathematik* [13], by the MKM community [9] and others.

As explained in the introduction it is a great challenge to small scientific communities to adopt such developments for its own scientific communication processes and to join forces with other scientific communities to get own requirements publicly recognised. A first step in such a direction is a more detailed description of ongoing scientific processes using standard RDF terminology.

SymbolicData started such efforts within the CA scientific community, collecting and presenting

- information about scientific activities of people – as of Sept. 2014 the SymbolicData People database contains `foaf:Person` entries of 708 people from the CA scientific community that can be explored via the SymbolicData SPARQL endpoint [14] – it is mirrored at [15],
- information about upcoming conferences – the information is extracted via SPARQL query to [15] from `http://symbolicdata.org/casn/UpcomingConferences/` and displayed in the Wordpress based site of the German Fachgruppe [20],
- information about past conferences and conference reports with references to the SymbolicData People database about speakers and organizers,
- information about German CA working groups and 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 – the information is displayed in the Wordpress based site of the German Fachgruppe [20],
- and semantic annotations to news in the blog of the German Fachgruppe as instances of RDF type `sioc:BlogPost` and `bibo:Document` attached to the blog post URL.

A particularly interesting project was started in cooperation with *Zentralblatt Mathematik* (ZBMath) towards a better author disambiguation. During the last years ZBMath spent much effort for better author disambiguation primarily using language processing technology to retrieve data tracks of people from their stock of abstracts, see [13]. The SymbolicData references within our People database offer additional insight into people activities (and – if properly recognised by the community – allows for active support and influence on a process that touches both vivid personal and scientific interests of the community), and we started to align SymbolicData People URIs with ZBMath URIs in the *ZBMath Person Matches* table at `http://symbolicdata.org/Data/ZBMathPeople/`.

Despite the importance of the ongoing 'Big Data' harvesting processes for the moment this and other activities suffer from little attention of a broader CA audience. In our talk at the CICM-14 conference [6] we asked: "How turn passive users (listed in the SymbolicData People database) into active ones?" A first hurdle is authentication. Nowadays there are three different ways to solve that problem:

1. The 'classical one': Set up a local database with user/password; users have to administer plenty of such application/user/password records (of course best using RDF technology).
2. The 'Google one': Use the (OAuth based) authentication service of one of the 'big players'. Your advantage: one password fits all, but who knows what the guys are tracking . . .
3. The WebID approach: Use your browser certificate and a FOAF profile managed by your own to provide access.

   You register with SymbolicData your FOAF profile, we will take the challenge stored there and offer it mixed up with our challenge to your browser. If the browser (with your certificate, imported into the browser from your source) returns the correct answer we know that's you sitting in front of it.

We started to set up such an infrastructure along the FOAF standards as 'proof of concept':

- The SymbolicData People database contains 'lightweight' FOAF profiles.
- For any person interested to join actively the CASN we generate a `foaf:PersonalProfileDocument` that relates this 'lightweight' FOAF profile with your own FOAF profile as `foaf:primaryTopic`.
- For a limited number of people from the German Fachgruppe (yet as passive users) we created such profiles and use them to display information about the different boards of the German Fachgruppe at their website [20].
- You can take over your own profile at any convenient web place under your control and extend it to meet the authentication requirements and tell us about that place.

  We change your PersonalProfileDocument and you are ready actively to join the CASN process.

The vision of a Computer Algebra Social Network goes far beyond that:

- Maintain at your local site up-to-date information about your working group and its people in a consistent RDF format as e. g. the AKSW team [1] does at `http://aksw.org/Team.html`.
- Maintain your FOAF profile at a personal page containing all important public up-to-date information about your activities in a consistent RDF format as e. g. the AKSW team member Natanael Arndt does at `http://aksw.org/NatanaelArndt.html`.
- Organize a regular harvesting process within the scientific community for such information to feed common pages at sites as e. g. `http://www.computeralgebra.de`, and to substitute part of the information centrally stored at SymbolicData today by decentrally managed one.
- Set up and run within the scientific community a semantic-aware Facebook-like Social Network and contribute to it about all topics around Computer Algebra using tools that express your contributions in an RDF-based syntax.

The last point sounds quite visionary but it is in no way utopic. The AKSW team provides a first prototype of a tool that realizes the challenging concept of a *Distributed Semantic Social Network* [17] to be running, in contrast to Facebook, on a multitude of independent nodes all over the world. Since the concept works also with a single node and can be extended later on, we set up such a node at `symbolicdata.org` for testing, see our CASN wiki page [16] for more information. Even if all this is very pre-alpha yet, the future is already on the way. Don't miss the train.

References

[1] The Agile Knowledge Engineering and Semantic Web Group at Leipzig University. `http://aksw.org/About.html` [2014-02-19]

[2] Developing a 21st Century Global Library for Mathematics Research. Report of the Committee on Planning a Global Library of the Mathematical Sciences. The National Academies Press 2014. `http://www.nap.edu/catalog.php?record_id=18619` [2014-09-24]

[3] EuDML – the European Digital Mathematical Library. `https://eudml.org/` [2014-09-23]

[4] Frame at Wikipedia. `http://en.wikipedia.org/wiki/Frame_(artificial_intelligence)` [2014-09-23]

[5] Gemeinsame Normdatei (GND). Informationsseite der Deutschen Nationalbibliothek. `http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html` [2014-09-13]

[6] H.-G. Gräbe, A. Nareike, S. Johanning. The SymbolicData Project — Towards a Computer Algebra Social Network. In: Workshop and Work in Progress Papers at CICM 2014. CEUR-WS.org vol. 1186. `http://ceur-ws.org/Vol-1186/#paper-21` [2014-09-13]

[7] S. Kaufman, M. Elliott, D. Shmueli: Frames, Framing and Reframing. In: G. and H. Burgess (eds.): Beyond Intractability. Conflict Information Consortium, University of Colorado, Boulder. Posted: September 2003. `http://www.beyondintractability.org/essay/framing` [2014-09-23]

[8] A. and M. Kohlhase: Spreadsheet Interaction with Frames: Exploring a Mathematical Practice. In: Intelligent Computer Mathematics, Lecture Notes in Computer Science Volume 5625, 2009, pp 341–356.

[9] Mathematical Knowledge Management. The MKM interest group. `http://www.mkm-ig.org/` [2014-09-13]

[10] A. Nareike. The SymbolicData sdsage package. `http://symbolicdata.org/wiki/PolynomialSystems.Sage` [2014-02-28]

[11] J. Pitman, C. Lynch: Planning a 21st Century Global Library for Mathematics Research. Notices of the AMS, August 2014.

[12] Sage – a free open-source mathematics software system. `http://www.sagemath.org` [2014-02-19]

[13] U. Schöneberg, W. Sperber. POS Tagging and its Applications for Mathematics. In: Intelligent Computer Mathematics. Lecture Notes in Computer Science, Volume 8543, 2014, pp 213–223.

[14] The SymbolicData SPARQL Endpoint. `http://symbolicdata.org:8890/sparql` [2014-02-19]

[15] The CASN SPARQL Endpoint. `http://symbolicdata.org:8891/sparql` [2014-09-13]

[16] The SymbolicData Project Wiki. `http://wiki.symbolicdata.org` [2014-09-13]

[17] S. Tramp et al.: An Architecture of a Distributed Semantic Social Network. In: Semantic Web Journal 2012, Special Issue on The Personal and Social Semantic Web.`http://www.semantic-web-journal.net/sites/default/files/swj201_4.pdf` [2014-09-23]

[18] Virtuoso Open-Source Edition. `http://virtuoso.openlinksw.com/` [2014-02-19]

[19] World Digital Mathematics Library (WDML). `http://www.mathunion.org/ceic/wdml/` [2014-09-23]

[20] Website of Fachgruppe Computeralgebra `http://www.fachgruppe-computeralgebra.de/` [2014-03-06]