# The SymbolicData Project – a Community Driven Project for the CA Community

Hans-Gert Gräbe

**Abstract.** The development of digital research infrastructures is a big challenge in particular to small scientific communities with restricted resources. In this paper 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 in the area of Computer Algebra. Our contribution sums up 15 years of experience within the SymbolicData Project both as intracommunity project serving the needs of the polynomial systems solving subcommunity and as intercommunity project at the CA level "at large". Within that volunteers' project we also practically developed different parts of such a digital research infrastructure. As the semantic web matures RDF based technologies and the Linked Data Cloud play a more and more important role also within our project. We discuss lessons learned from these activities and hurdles and obstructions to generalize intracommunity experience to an intercommunity level within the CA domain.

**Mathematics Subject Classification (2010).** Primary 68W30.

**Keywords.** symbolic data, research infrastructure, research social web.

## 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 application 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 arena" of research politics between different stakeholders, see e.g., [14, 15]. 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 should 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 – data *and* communication – of a research infrastructure.

With the advancement of the SYMBOLICDATA Project [20] towards a Computer Algebra Social Network (CASN) we pursued such a concept in a specific context for several years. We started to investigate intra- and intercommunity communication processes 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 this paper 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 learned from these activities and hurdles and obstructions to generalize intracommunity experience to an intercommunity level within the CA domain.

During the last decade RDF – the Resource Description Framework – emerged as de facto semantic web standard. Nowadays there is a well established *Semantic Web Stack*[1] of HTTP based protocols that support the different tasks within semantic web communication. We spent much effort to transform the SYMBOLICDATA database to that emerging standard.

To advance the development of a distributed digital CA social network 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 [9] and already set up two such nodes at

- `http://symbolicdata.org/rdf` – the SYMBOLICDATA CASN node with additional information in draft version, in particular detailed records about several CA conferences,
- and `http://www.fachgruppe-computeralgebra.de/rdf` – the CASN node of the German Fachgruppe with RDF files providing additional information, in particular about articles published in their Computeralgebra Rundbrief.

Such a distributed infrastructure complements the centrally managed SYMBOLICDATA infrastructure consisting of

- our *github repos* [17] with RDF data dumps, code of our tools, and best practice examples,
- a Virtuoso based RDF store [18] serving the data for inspection and SPARQL querying,
- a SPARQL endpoint [19] to query the data, and
- a wiki [20] with detailed additional information about our project.

To ensure interoperability, this technical development process should be accompanied by a strong social intercommunity communication process to agree upon 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 Database

The SYMBOLICDATA Project took its origin from the polynomial systems solving subcommunity and early extended its scope to geometry theorem proving and integer programming as two of the major application areas of polynomial systems solving at that time. We refer to [6] for a more detailed overview of the goal, aims, concepts and history of the SYMBOLICDATA Project.

In [5] we gave a detailed overview of the data modelling and RDF bindings for polynomial systems and polynomial ideals data and metadata. Currently the SYMBOLICDATA data collection

[1] `https://en.wikipedia.org/wiki/Semantic_Web_Stack`.

contains 390 records and 633 configurations from the areas of polynomial systems solving, 83 free algebra records, 8 G-algebra records and 297 GeoProofSchemes.

During the last years we enlarged the SYMBOLICDATA database with data from other CA subcommunities following a slightly different policy – data are stored by the subcommunity itself, metadata are extracted and stored in our database. Such a distinction is well supported by RDF design principles since the Resource Description Framework is about *description* of *resources*, represented by (globally unique) *resource identifiers* (URIs). It is best Linked Data practise to provide URIs in such a way, that they are accessible by the HTTP internet protocol and a valuable part of structured information about the resource is delivered upon HTTP request to that URI.

Note that the distinction between *resources* (managed by CA subcommunities using intracommunity standard methods of representation and access) and *resource descriptions* (important for technically supported interchange between such subcommunities and for intercommunity communication) leads away from XML based design principles that mainly focus on the distinction between information (XML records) and information structure (described with XSchema). To use XML-based storage formats was one of the design decisions within earlier SYMBOLICDATA versions. In our new version we offer a greater variety of storage formats for research data and shift the focus from the data format to the provision of a *semantic aware parser* component that can be integrated with other tools. This gives CA subcommunities more freedom to provide their well maintained and curated data for intercommunity use. Within the SYMBOLICDATA project we concentrate on the extraction and management of *metadata*.

Thus the SYMBOLICDATA database was enlarged with metadata of 8630 Fano polytope records [12] and 5399 Birkhoff polytope records [11] from the polymake project [4] hosted by Andreas Paffenholz, 3605 transitive group records from the Database for Number Fields [8] by Jürgen Klüners and Gunter Malle and 49 test set records from the Normaliz collection [2].

## 3. Querying the Database – Two Examples

Representation of metadata in RDF format together with the SPARQL querying language are well suited to explore data and nowadays even emerged to a certain standard within the Linked Data world. Let's demonstrate the search and filter facility on the metadata of two different collections by two examples[2]. We assume the reader to be familiar with RDF and SPARQL basics.

**Example 1:** This is an example from the Fano polytope metadata. The following listing in RDF Turtle notation with our special name space

```
sd: <http://symbolicdata.org/Data/Model#>
```

for the SYMBOLICDATA model elements (classes and predicates) describes a typical Fano polytope metadata record:

```
<http://symbolicdata.org/Data/FanoPolytope/F.3D.0001>
  a sd:FanoPolytope ;
  sd:inZIPFile <http://polymake.org/polytopes/.../fano-v3d.tgz> ;
  sd:hasFileName "F.3D.0001.poly" ;
  sd:pointed "true" ;
  sd:feasible "true" ;
  sd:hasDimension "3"^^xsd:integer ;
  sd:lineality_dim "0" ;
  sd:n_vertices "6"^^xsd:integer ;
  sd:n_facets "5"^^xsd:integer ;
  sd:lattice "true" ;
```

[2] See also `http://wiki.symbolicdata.org/MoreQueries`.

```
  sd:cone_ambient_dim "4"^^xsd:integer ;
  sd:cone_dim "4"^^xsd:integer ;
  sd:facet_width "5"^^xsd:integer ;
  sd:compressed "false" ;
  sd:essentially_generic "false" ;
  sd:volume "31/3" ;
  sd:lattice_volume "62" ;
  sd:n_interior_lattice_points "1"^^xsd:integer ;
  sd:n_lattice_points "34"^^xsd:integer ;
  sd:n_boundary_lattice_points "33"^^xsd:integer .
```

The example belongs to the `fano-v3d.tgz` collection of 3-dimensional Fano polytopes by Andreas Paffenholz [12]. The collection itself is contained in a single tgz-file and is provided for download from the URL given as second entry in the above listing. The third entry provides the file name of the core data of the given example within that collection.

All other lines contain numerical or boolean metadata information about the given example. We apply the golden rule of *self-explanatory identifiers* to associate semantic meaning with syntactic predicate names as recommended within the RDF world. This helps the "semantic aware user" to understand the given metadata record without further elaboration. Note that RDF provides means to describe semantics of predicates in more detail and we provide such additional information, see our wiki [20].

The metadata in the above example was extracted from the source file, transformed into RDF notation and stored as one of the metadata records in the `FanoPolytopes.ttl` file by our fellow Andreas Nareike.

If metadata are separated from core data and collected into a single RDF graph one can SPARQL query the data and find all examples with given properties. Similar to SQL for relational data SPARQL provides an expressive syntax for search and filter purposes on RDF data. Moreover, queries are executed using only the HTTP protocol and hence they can easily be integrated into more complex programming contexts.

Let's, e.g., query URI, volume, cone dimension, number of facets and the file name for all 4-dimensional Fano polytopes in Andreas Paffenholz' collection `fano-v4d.tgz` with lattice volume $v < 100$ and at most 7 facets. The `FanoPolytopes` collection is served by our RDF store [18], hence the result can be obtained running the following query within the web interface [19] of the SYMBOLICDATA SPARQL endpoint.

```
PREFIX sd: <http://symbolicdata.org/Data/Model#>
select ?uri ?v ?cd ?nfc ?fn
from <http://symbolicdata.org/Data/FanoPolytopes/>
where {
  ?uri a sd:FanoPolytope .
  ?uri sd:cone_dim ?cd .
  ?uri sd:n_facets ?nfc .
  ?uri sd:hasFileName ?fn .
  ?uri sd:lattice_volume ?v .
  filter ((xsd:integer(?v)<100) and (xsd:integer(?nfc)<7))
}
```

The query returns 6 records, one with 4 facets, two with 5 facets and three with 6 facets.

**Example 2:** The following test sets example

```
http://symbolicdata.org/Data/TestSet/graph10_001_k5free
```

wee took from the Normaliz collection provided in 2015 by Tim Römer (entry 4 below). It is part of a collection of counterexamples to a conjecture of Sturmfels and Sullivant (more about that in entry 3 below). The remaining lines contain mainly numerical metadata for the given example.

```
<http://symbolicdata.org/Data/TestSet/graph10_001_k5free>
  a sd:TestSet, sd:Normalization ;
  rdfs:label "TestSet/graph10_001_k5free" ;
  rdfs:comment """Verification of a conjecture of Sturmfels and
    Sullivant on normality of cut polytopes for some specific
    graphs [SS], [BHIKS], [O], [S]."""  ;
  sd:hasOrigin """Normaliz Collection, Tim Roemer 2015-03.
    From Example 8 of the Normaliz collection"""  ;
  sd:hasNumberOfExtremeRays "512"^^xsd:integer ;
  sd:hasSupportingHyperplanes "352"^^xsd:integer ;
  sd:numberOfRows "512"^^xsd:integer ;
  sd:numberOfColumns "48"^^xsd:integer ;
  sd:hasNumberOfHilbertBasisElements "512"^^xsd:integer ;
  sd:hasRank "25"^^xsd:integer ;
  sd:hasNormalizPrimaryBase
    <http://.../normaliz/InterChallExamples/cut_monoids.zip> ;
  sd:hasNormalizSDBase
    <http://symbolicdata.org/.../graph10_001_k5free.zip> .
```

The last two lines contain additional information about the core data itself – the last line gives the URL to a zip-file with the files `graph10_001_k5free.in` (input in Normaliz format) and `graph10_001_k5free.out` (containing among others listings of the 512 elements of the Hilbert basis and the 512 extremal rays in Normaliz syntax) and the second last line points to an external source with (among others) the primary base of the given example.

To explore the *graph examples* from the Normaliz test suite one can issue a SPARQL query that lists label, rank, number of columns, number of rows, number of extreme rays, number of Hilbert Basis elements, an external link to the Normaliz database and the link to the SD test sets collection. Several entries in the following query are listed as *optional* to include also records where the given parameter is not available. This is the case if some of the parameters are hard to compute.

```
PREFIX sd: <http://symbolicdata.org/Data/Model#>
select distinct ?l ?r ?nc ?nr ?ner ?nhbe ?npb ?nsb
from <http://symbolicdata.org/Data/TestSets/>
where {
  ?p a sd:TestSet .
  ?p rdfs:label ?l .
  optional { ?p sd:hasRank ?r . }
  optional { ?p sd:numberOfColumns ?nc . }
  optional { ?p sd:numberOfRows ?nr . }
  optional { ?p sd:hasNumberOfExtremeRays ?ner . }
  optional { ?p sd:hasNumberOfHilbertBasisElements ?nhbe . }
  optional { ?p sd:hasNormalizPrimaryBase ?npb . }
  optional { ?p sd:hasNormalizSDBase ?nsb . }
  filter regex(?p, "graph")
}
order by ?l
```

The query returns 19 records, 18 of them from the `InterChallExamples/cut_monoids.zip` series and another one `TestSet/semigraphoid5`. The latter can be filtered out using, e.g., the regular expression `"/graph"` instead.

## 4. Fingerprints and Semantic Aware Indexing

In the previous section we demonstrated the ease and power of RDF and SPARQL querying to manage, search and filter metadata. From the internal perspective of a research community this is a special aspect of every research data collection. For this purpose core data is usually enriched with metadata that collect important relevant information of the individual data records and is stored together with the core data in a compact way. We denote such metadata for an individual data record as its *fingerprint*.

Similar to a hash function a fingerprint function computes a compact metadata record (*resource description* in the RDF terminology) to each individual core data record (*resource* in the RDF terminology). As with a hash function one can use the fingerprints to (almost) distinguish different data records within the given collection and to match new records with given ones. But there is an essential difference between (classical) hash functions and well designed fingerprints: fingerprint functions exploit not only the textual representation of the data record as meaningless syntactical character string but convey *semantically* important information or even compute such information from the string representation. Fingerprints are in this sense *semantic-aware* and can even be designed in such a way that they map ambiguities in the textual representation of records (e.g., polynomial systems given in different polynomial orders and even in different variable sets) to *semantic invariants*.

The design of appropriate fingerprint signatures is an important *intracommunity* activity to structure its own research data collections. Such fingerprint signatures are also very useful for the *intercommunity* usage of research data collections, since they allow to navigate within the (foreign) research data collection without presupposing the full knowledge of the "general nonsense" of the target research domain, i.e., the informal background knowledge required freely to navigate as scientist in that domain. Hence well designed fingerprint signatures are to be considered also as a first class service of a special research community to a wider audience to inspect their research data collections without using the community-internal tools to access the resources themselves.

Usually the research data collections (resources in the RDF terminology) of a certain community are stored in a specially designed community-internal format, often as plain text (e.g., the Normaliz collection [2]), in a special XML notation (e.g., the polymake collection [4]) or as SQL database (e.g., the Database for Number Fields [8]). Such formats usually employ special formal semantics agreed within the community as an effective way to store domain specific input and output data and used by commonly developed tools with appropriate parsing functionality.

Usually such formats are extended to store research metadata, i.e., fingerprints or *resource descriptions* in the RDF terminology, together with the research data. For SPARQL querying one has to extract those fingerprints and to transform the data in an appropriate RDF notation. In the previous section we demonstrated the advantage of such an approach, that suggests a subdivision of duties between the intracommunity efforts to compile well designed fingerprints and the intercommunity level to provide appropriate concepts and tools to upload, manage, search and filter such fingerprint metadata and to combine such data from different subcommunities. Similar to unit testing or benchmarking environments the latter effort is a *cross cutting concern*, since it can be technically solved once and afterwards reused many times within different scientific domains.

We applied this approach, first used within the SymbolicData Project to navigate within polynomial systems data, to the data sets on polytopes and on transitive groups newly integrated with SymbolicData version 3, and also within the recompiled version of test sets from integer

programming. We store these fingerprints in our RDF data store [18] thus allowing for a unified navigation and even cross navigation within such data using the SPARQL query mechanism as a generic Web service provided by our SPARQL endpoint [19].

## 5. 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 [13] and FRISCO [3] projects – the polynomial systems database – was going to break down. After the end of the projects' funding 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 [9], and started both conceptual and practical work towards a semantic-aware Computer Algebra Social Network.

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 of remotely available data on transitive groups (*Database for Number Fields* by Gunter Malle and Jürgen Klüners [8]) and polytopes (databases by Andreas Paffenholz [10] 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 [7].

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 our wiki [20] for more details about the project and the new release.

## 6. Towards a Computer Algebra Social Network

All parties want to have a powerful digital research infrastructure but are rarely enough 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, and 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 researchers within the Computer Algebra community and its severe subcommunities similar to other social networks.

Note that 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 should take into account such a diversity and develop a decentralized solution based on modern semantic technologies that increases 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 way.

### 6.1. 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]. 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 [19] 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 demand such nodes to be uniformly structured nor do such nodes require big web resources. LOD sources are self-explanatory by design and their 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 the workload 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 publicly accessible directory of RDF files containing valuable information as provided by the CASN sample node[4] of the SYMBOLICDATA Project. As proof of concept in the subdirectory `Conferences` we provide prototypically detailed information about five CA 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

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

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.

### 6.2. 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 [19]. This information is used by the German CA Fachgruppe at one hand to present an online list of upcoming conferences and at the other 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 already for several years and compiled from it a list of (at the moment 173) *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.

### 6.3. The SYMBOLICDATA People Database

The concept of URI (Unique Resource Identifier) 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. Whereas 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 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, thesis supervisors and reviewers 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) if they are interested in such a cooperation.

### 6.4. 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.

### 6.5. The CA Systems Project

In summer 2016 we also intensively discussed perspectives of the swMATH project [16]. 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., 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 from 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.

## 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`. [2016-11-19]

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

[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 – From Data Store to Computer Algebra Social Network. In: Computeralgebra-Rundbrief vol. 55, pp. 22-26 (2014).
[6] 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).
[7] A. Heinle, V. Levandovskyy: The SDEval Benchmarking Toolkit. ACM Communications in Computer Algebra, vol. 49.1, pp. 1–10 (2015).
[8] J. Klüners, G. Malle: A Database for Number Fields. http://galoisdb.math.uni-paderborn.de/. [2016-11-08]
[9] The Linked Open Data Cloud. http://lod-cloud.net/. [2016-11-20]
[10] A. Paffenholz: Polytope Database. http://www.mathematik.tu-darmstadt.de/~paffenholz/data/. [2016-11-08]
[11] A. Paffenholz: Lists of Combinatorial Types of Birkhoff Faces. http://polymake.org/polytopes/paffenholz/www/birkhoff.html [2016-11-20]
[12] A. Paffenholz: Smooth Reflexive Lattice Polytopes. http://polymake.org/polytopes/paffenholz/www/fano.html [2016-11-20]
[13] 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. [2016-11-16]
[14] Research Infrastructures, including e-Infrastructures. http://ec.europa.eu/programmes/horizon2020/en/h2020-section/research-infrastructures-including-e-infrastructures. [2016-11-16]
[15] 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. [2016-11-16]
[16] swMATH – an Information Service for Mathematical Software. http://swmath.org/about_contact. [2016-11-16]
[17] The SYMBOLICDATA Github Account. https://github.com/symbolicdata. [2016-11-15]
[18] The SYMBOLICDATA RDF Data Store. http://symbolicdata.org/Data. [2016-11-15]
[19] The SYMBOLICDATA SPARQL Endpoint. http://symbolicdata.org:8890/sparql. [2016-11-20]
[20] The SYMBOLICDATA Project. http://wiki.symbolicdata.org. [2016-11-20]

Hans-Gert Gräbe
Computer Science Department
Leipzig University
Augustusplatz 10
D 04109 Leipzig
Germany
e-mail: graebe@informatik.uni-leipzig.de