Semantic-aware Fingerprints of Symbolic Research Data

Hans-Gert Gräbe

Leipzig University, Germany
http://bis.informatik.uni-leipzig.de/HansGertGraebe

ICMS-2016, Berlin, 2016-07-13

Introductory Remarks

- *Information Services for Mathematics* addresses a more complex target compared to *Mathematical Software*.
- Mathematical software is only part of a whole *infrastructure for mathematical research* that nowadays goes much beyond the classically hawked *paper and pencil* or *chalk and blackboard*.
- The themes *software*, *services*, *models*, *and data* point to at least four dimensions to enhance the mathematical research infrastructure in the era of ubiquitous computing and increasingly important digital interconnectedness within the Digital Universe (DU).

Real World (RW) and Digital Universe (DU)

(D1) RW: Real world *tasks* – *Resources* in the RDF terminology. A **task** is a real world *process* with a *goal* triggered by *interested* people.

(D2) DU: Associated task related *descriptions*. Requires *pointers* to RW task details, i.e., **digital identities** (DI) as URIs (unique resource identifiers) in the most formalized way required to be processed by computers.

(D3a) DU: Many descriptions relate to the same RW task. *Problem:* Match related DIs in communication.

(D3b) DU: Many descriptions relate different RW task. *Problem:* Express such relations as relations between DIs in the DU.

Real World and Digital Universe

RDF – the *Resource Description Framework* – with its models, standards, protocols and web architecture is nowadays an established standard to process such problems on a technical level. Together with the *Open Culture Paradigm* it is the basis for the ever growing *Linked Open Data Cloud* (LOD).
Links:

- `https://www.w3.org/RDF/`
- `http://www.w3.org/standards/techs/rdf`
- `http://lod-cloud.net/`

(D4) RW: The digitally supported social communication processes about relations between DIs in the DU should have *impact on the RW performance of coupled tasks* within a cooperative environment.

Information Services for Mathematics

Information Services for Mathematics have to address at least four dimensions:

(I1) *Research Data* – papers, conference announcements, mailing lists, web sites ... (data streams).

Although it seems to be completely within the DU, it is a good advice to differentiate between *resources* (D1) – as rather a part of the RW – and *resource descriptions* (D2).

Resource descriptions are required to *structure* resources for search and filter processes (D3b). RDF is best suited for such a task.

Information Services for Mathematics

(I2) *Research Input Data* – well curated publicly available data stocks with data relevant for a whole research community (data stores).

This is a new phenomenon within the upcoming DU with great impact on research problems, research methods and research paradigms, in particular within the *digital humanities*.

Important for the coherence of research questions addressed by the community and thus for the formation of a specific research community itself around its central research problems.

Information Services for Mathematics

(I3) *Mathematical Software* – written to run computer simulations.

If software is not used in such a way it is of less academic interest. Moreover, computer simulations often require the interplay of several *scientific packages* bundled within an *application*, hence we propose to use *computer simulation* as the broader notion.

Note that computer simulation (as a special kind of "experiment") relates to an epistemological dimension not covered by (D1–D3) above.

The public availability of newly developed *simulation methods, procedures and techniques* is relevant for the traceability of the proposed scientific approaches and increasingly accompanies classical forms of description of scientific advancement by academic papers.

Information Services for Mathematics

(I4) Publicly available *output data* – important for the independent reproduction of results and thus of essential importance for the process of academic quality assurance.

Output data is the starting point for new research questions and thus output data mutates to input data.

In most of the cases such a mutation is mediated by a community-internal interpersonal transformation process that transforms the often large output data (or a whole bundle of such data) into (one or several) more compact input data adapted to the new research question(s).

Information Services for Mathematics

Efforts to secure a research infrastructure for mathematical data at large at the level (I2) or even (I4) are lost in the brushwood of everlasting (for at least a decade) debates about reliable formal but semantically expressive formats as MathML or OpenMath for data resulting from calculi, that are already highly formalized – at least at an informal level – by the internal nature of the research topics themselves.

The situation reminds the Tower of Babel Project, since subcommunities are digitally already well established, developed their own formalizations for their own research data at (I2) level and apply such formalizations very successful within their intracommunity communication processes.

The SymbolicData Project

The SymbolicData Project

(S1) is an inter-community project with roots in the activities of different Computer Algebra Communities to develop concepts and tools for profiling, testing and benchmarking Computer Algebra Software (CAS) – level (I2),

(S2) aims at interlinking these and other scientific activities between different subcommunities of the CA community using modern Semantic Web concepts – tasks (D3a) and (D3b) and

(S3) during the last years concentrated efforts to set up the technical basis for a *CA Social Network infrastructure* within the Linked Open Data Cloud – level (I1).

Research Data and Metadata

- For management, search and filter functionality research data is usually enriched with *metadata* that collect important relevant information of the individual data records in a compact manner.

  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 data record (*resource* in the RDF terminology).
- As with a hash function one can use the fingerprints to distinguish different data records within the given collection and to match new records with given ones.

Example

Fingerprints for ideals in polynomial rings:

```
<http://symbolicdata.org/Data/Ideal/Gerdt-93a> ...
  sd:hasDegreeList "2,3,4" ;
  sd:hasLengthsList "3,3,4" ;
  sd:relatedPolynomialSystem
       <http://symbolicdata.org/Data/IntPS/Gerdt-93a> ;
  a sd:Ideal .
```

Polynomial systems and ideals – the semantic complexity of a seemingly easy question.

Fingerprints are usually stored together with the resource itself (as for polytopes in *polymake*, contribution by A. Paffenholz, and for integer programming examples in *normaliz*, contribution by Tim Römer) or within a database (transitive groups by Klüners and Malle).

Fingerprint Signatures

- 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.

Fingerprint Signatures

- 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.

Fingerprints and the LOD

For navigational and filter tasks it is necessary to extract fingerprints into a common database. Best practice uses an RDF representation to integrate that information into the Linked Open Data Cloud and to offer SPARQL querying the metadata.

Example SPARQL query for polynomial systems:

```
PREFIX sd:  <http://symbolicdata.org/Data/Model#>
select ?a
from <http://symbolicdata.org/Data/PolynomialSystems/>
where {
?a a sd:Ideal .
?a sd:hasDegreeList "2,3,4" .  }
```

See `http://symbolicdata.org/Presentations/icms16-examples.txt` for more examples.

Fingerprints and the LOD

During the last years the SymbolicData Project concentrated on collecting such fingerprint information from different CA sources – the research data are maintained by the subcommunity, the fingerprints allow for easy navigation within that data.

Different approach: Direct integration of the resources themselves into a general software system as, e.g., SageMath. Drawbacks: Restricted search functionality, no direct integration into the Linked Open Data Cloud possible.

The SymbolicData Project

The SymbolicData provides

Data and Fingerprints:

- Polynomial Systems Solving
- Geometry Theorem Proving
- Free Algebras
- G-Algebras

Fingerprints:

- Test Sets from Integer Programming (T. Römer)
- Fano Polytopes (A. Paffenholz)
- Birkhoff Polytopes (A. Paffenholz)
- Transitive Groups (J. Klüners, G. Malle)

The SymbolicData Project

Tools:

SDEval Package (Albert Heinle)

- Aim: Set up, run, log, monitor standardized Computations on SD data series in a reliable way
- Technology: Python standalone on top of the OS
- `http://wiki.symbolicdata.org/SDEval`

SDSage Package (Andreas Nareike)

- Aim: Call the new Polynomial Systems format from SageMath
- Technology: SageMath Python Package
- `http://wiki.symbolicdata.org/PolynomialSystems.Sage`

Tools and data are designed to be used both on a local site for special testing and profiling purposes to manage a central repository at `http://www.symbolicdata.org`

SymbolicData Infrastructure

- Github organizational account `http://github.com/symbolicdata`
- A project wiki at `http://symbolicdata.org`
- A mailing list
- Web access to the XML resources
- A centrally operated Virtuoso based RDF data store for meta data
- Organized along Linked Data Principles
- Regular dumps of RDF data in Turtle format
- A SPARQL endpoints to query the data
- Advise for local installation of tools and data based on Virtuoso and a local Apache Web server

Towards a CA Social Network

During the last years we consolidated the following infrastructure:

- We collect, update, and serve relevant information about CA people, upcoming and past CA conferences through our central RDF store.
- We set up local CASN nodes at
  - `http://symbolicdata.org/rdf`
  - `http://fachgruppe-computeralgebra.org/rdf`

  with more information in Linked Open Data format as best practices.
- We operate `http://symbolicdata.org/info` as example how to integrate such information into local web pages, see also

  `http://fachgruppe-computeralgebra.org/symbolicdata`

Links

- `http://wiki.symbolicdata.org` – the SD Wiki
- `http://symbolicdata.org/XMLResources` – the SD XML Resources
- `http://symbolicdata.org/RDFData` – the SD RDF Data Turtle Files
- `http://symbolicdata.org/Data` – the SD OntoWiki view on the RDF data, including the CASN data
- `https://github.com/symbolicdata` – the SD organizational github account with several git repos