

Semantic Data Management Initiative

Roundtable Report SemData@Sofia

March 11-12, 2010 in Sofia, Bulgaria

<http://semdata.org/events/2010/sofia>

Abstract

Semantic data management is a paradigm, a common name for a range of techniques, for manipulation and usage of data based on its meaning. The SemData@Sofia Roundtable was a first workshop in a series of events by the Semantic Data Management Initiative which aims at investigating various aspects of semantic databases and data management in the large. The purpose of the roundtable was to provide a ground for presentations, expert discussions and trans-disciplinary collaborations on issues such as semantic repositories, their virtualization and distribution, interoperability, processing, reasoning and accessing linked data, the establishment of semantic data buses or so-called process layers that bridge the gap between the data layer and the application layer, and last but not least the benchmarking of semantic data management solutions. The roundtable was a very interactive event bringing together researchers and practitioners from academia and industry from both the semantics and the database community. Besides playing the role of being the kick-off for a whole series of events and activities, the meeting in Sofia already offered various interesting insights and collaborations. This report is the final summary of the discussions and activities at the roundtable and a pointer towards future endeavors.

List of participants

The workshop was organized by Atanas Kiryakov from Ontotext AD, Dieter Fensel and Reto Krummenacher from STI Innsbruck at the University of Innsbruck. Participation to the Roundtable was by invitation only in order to keep the discussion focused.

| | | | |
|-------------------------|----------------------|-------------------|-------------------|
| Andreas Harth | KIT | Kavitha Srinivas | IBM Watson |
| Atanas Kiryakov | Ontotext | Massimo Paolucci | DOCOMO Euro-Lab |
| Bryan Thompson | Systap | Michael Witbrock | CyCorp |
| Carlos Pedrinaci | Open University | Mitko Iliev | OpenLink Software |
| Damyan Ognyanov | Ontotext | Orri Erling | OpenLink Software |
| Dieter Fensel | STI Innsbruck | Peter Haase | fluid Operations |
| Eleftherios Sidirourgos | CWI | Reto Krummenacher | STI Innsbruck |
| Elena Simperl | KIT | Silvia Karagova | Ontotext |
| Fabrice Huet | INRIA, Univ. de Nice | Spyros Kotoulas | VU Amsterdam |
| Han Sung-Kook | STI Innsbruck | Thanh Tran | KIT |
| Ivan Peikov | Ontotext | Vassil Momtchev | Ontotext |
| John Domingue | Open University | Zoltan Miklos | EPFL |

Program

Day 1: March 11, 2010

Welcome, announcements, introduction of the participants (Atanas Kiryakov)
Introduction to the SemData initiative (Dieter Fensel)

Session 1 - Topics

- Semantic Spaces (Reto Krummenacher)
- Semantic Repositories (Atanas Kiryakov)
- RDBM Versus Graph-based Approaches (Orri Erling)

Session 2 - Vendors

- Kavitha Srinivas (IBM, SHER)
- Bryan Thompson (SYSTAP, Bigdata)
- Damyan Ognyanov, Ivan Peikov (Ontotext, OWLIM)
- Orri Erling (Openlink Software, Virtuoso)

Session 3 - Developers

- Eleftherios Sidirourgos (CWI, MonetDB)
- Spyros Kotoulas (VU Amsterdam, LarKC)
- Fabrice Huet and Reto Krummenacher (INRIA Sophia Antipolis, STI Innsbruck, SOA4All)
- Andreas Harth (KIT, SWSE/YARS)

Day 2: March 12, 2010

Session 4 - First-Tier Users, Engine Extensions

- Michael Witbrock, (Cycorp, LarKC)
- Elena Simperl, Thanh Tran (KIT, PlanetData)
- Carlos Pedrinaci and John Domingue (Open University, SOA4All)

Session 5 - Verticals

- Massimo Paolucci (Docomolab, Mobile Operators)
- Vassil Momtchev (Ontotext, Life Sciences)
- Peter Haase (fluid Operations, Enterprise Data Management)

Session 6 - Summary and next steps

Presentations

Semantic Spaces Middleware (Reto Krummenacher, STI Innsbruck): This first presentation gave an introduction to semantic spaces and semantic middleware in general. Semantic spaces offer Web-style 'persistent publish and read' application integration through a virtualization layer on top of distributed linked data sources. The goal is to abstract from the limitations of (singular) repositories towards a network of spaces in the cloud. In a second part, these basic ideas of data publishing and sharing in machine-to-machine communication were put in the context of running research projects such as LarkC and SOA4All. A scalable data layer is a prerequisite for the implementation of ad hoc coalitions in reasoning systems or to move from service buses to Web-scale service economies. In summary, a move from a data layer towards an integrated layer that supports processes is required.

Semantic Repositories (Atanas Kiryakov, Ontotext):

Semantic Data Management - RDB and Graph-based Approaches (Orri Erling, OpenLink Software): This presentation addressed some of the hot topics around RDF databases and RDB mappings. Mapping RDB is doable, but not very popular nor advisable if an application deals with high numbers of sources, high overlap of schema and high variability of data. In such cases, using RDF is generally the smarter approach. However, also agility will be better, RDF still needs to catch up in terms of performance! To this end, it is all about parallelism and locality: run data from memory, ship functions to data and maximize asynchronicity to avoid network latency, and use ACID intelligently. On a more technical basis, this presentation discussed issues related to the use of column stores for RDF, the use of semantic data models for online applications or highly transactional application for which RDF does not make much sense, and last but not least the lack of real benchmarks that match the real characteristics of graph models (RDF) and the query language (SPARQL 1.1).

SHER Scalable Highly Expressive Reasoner (Kavitha Srinivas, IBM Watson): The presentation started with a short overview of SHER and quickly addressed more detailed aspects of the reasoner. Key features for achieving scalability in SHER are summarization and refinement of instance graphs. The technologies were then applied in a use case of the clinical domain: search in medical literature, text analysis and cleanup, combinations of probabilistic and DL reasoning. Open problems that were named around semantic data management: RDF and RDB, how can existing DB2 infrastructures be exploited? – RDF benchmarks, how appropriate are available benchmarks?

Bigdata® Overview (Bryan Thompson, SYSTAP): The presentation gave an introduction to Bigdata and some of the most recent new features. The offer of Bigdata is a distributed database at petabyte scale with the particular aim of offering a fast integration of heterogeneous data, linked data and structured data. Key differentiators of Bigdata are dynamic sharding, temporal database support, high availability design and the open source development. In terms of RDF storage, Bigdata supports RDFS+ inference, RDF triple-specific

indices, but also quads (named graphs), and datum level provenance as well as metadata queries without complex reification and additional indices.

OWLIM - Basics and New Features (Damyan Ognyanov, Ivan Peikov, Ontotext): The presenters started with a short introduction to the scalable semantic repository and high-performance reasoning engine OWLIM. Furthermore, the presentation included more technical details about various updates that were recently integrated with OWLIM: “smooth” Invalidation, inconsistency checking, RDF node ranking, replication clusters for load balancing of concurrent user requests, failover in replication clusters, and notification services for graph patterns.

The OpenLink Virtuoso Universal Server (Orri Erling, Openlink Software): This presentation gave a technical introduction to the Virtuoso Universal Server and the work of OpenLink for RDF management. The database supports SQL and SPARQL as query languages and has full ACID support. At the level of storage, Virtuoso supports quads and graph-level role-based security. In terms of inference, sub-class/sub-property as well as transitive and inverse properties are supported without load time materialization; i.e. it is all done at runtime with backward chaining. Materialization is however possible via SPARUL. In addition to single-site databases, Virtuoso is cluster-minded with all nodes offering identical services and full ACID. The cluster release includes parallel extensions to SQL, similar to MapReduce. For high availability, partitions can be replicated and automatic exclusion and recovery of failed servers is installed.

The MonetDB/RDF Infrastructure (Eleftherios Sidirourgos, CWI): The presentations offered in insight into ongoing research around MonetDB with respect to RDF data management. The main ideas are to combine established MonetDB technologies such as cracking, recycling and run-time query optimization with new indices for fast transitivity computation of triples (inference) and RDF-specific compression schemas for the global dictionary (for both URIs and literal values).

Parallel Forward Reasoning (Spyros Kotoulas, VU Amsterdam): The main topic of this presentation was RDFS/OWL-Horst materialization on clusters. Two projects were presented in more detail – Marvin and WebPIE. Marvin is designed to compute the materialization on (almost) unstructured networks via local computation and data exchange. WebPIE uses MapReduce to distribute the computation of forward inferences over RDFS and OWL-Horst rule sets. The schema triples are replicated to all nodes, while the instance triples are “streamed”. Some tips derived from this work included: do not maintain full indexes, but only those really required

Overlays Architecture – Some Words About Distribution (Fabrice Huet, INRIA-University of Nice): The work shown in this presentation is based on recent research around distributed semantic spaces in the SOA4All project. Distribution is governed by structured P2P overlays, here a combination of a Chord ring and a 3D-CAN space. RDF triples are stored in the CAN space with lexicographical ordering and hence some degree of namespace-clustering; i.e. no hashing. The Chord ring is used to discover relevant CAN spaces based on the hashed identifier of space URIs. A big share of the presentation was dedicated to a discussion of the pros and cons of

using overlays in combination with distributed RDF – functionality-wise the approach works, however, there are still severe latency problems and it is questionable if the added complexity of the CAN overlay really adds any functionality to real application scenarios. An alternative would be a Chord-based virtualization layer with single-site or clustered RDF repositories.

SWSE/YARS@SemData Sofia 2010 (Andreas Harth, KIT): The presentation began with the claim that Semantic Web systems should be schema-agnostic. Algorithms and systems cannot rely on predefined vocabulary and ontology terms, given that data integrated from the Web exhibits enormous variety in the vocabulary terms used. Another important property of Web data is that the data is distributed over a large number of sources, a fact that has to be taken into account in algorithms operating over such data. For example, authoritative reasoning becomes relevant in a Web context where anyone can say anything – which leads to an explosion of inferred statements. SAOR – Scalable Authoritative OWL Reasoner – is a best-effort RDFS and OWL reasoner at Web data. SAOR accepts as input a set of statements as collect for example by a Web-crawler and produces, by means of forward-chaining, a knowledge base enhanced by the given fragment of OWL reasoning. Further challenges of semantic data management that were shortly addressed were ranking by identifier, benchmarking and faceted browsing/navigation of semantic data.

Scaling up the Data, Scaling up the Semantic (Michael Witbrock, Cycorp): The presentation started with some lessons from the LarkC project: there is a frequent need to use large quantities of data in real-life scenarios and for non-standard inference such as statistical forward inference or “missing value” estimation – there is much more to it as logical inference. To this end, data availability and proximity to computation are big issues. Other important aspects of semantic data management that were addressed in this talk: publishing to the Semantic Web is just as important as reading from it; context is vital for correct inference; and forward inference is good but dangerous.

PlanetData (Elena Simperl, KIT): The objective of this presentation was to introduce a Network of Excellence which will be launched in fall 2010. PlanetData covers the particular topics of SemData and will coordinate research(ers) across Europe, leverage the expertise of different disciplines and create instruments for sustainability. An interesting element of PlanetData that is addressed at individuals and organizations taking part in SemData are the so-called PlanetData Programs. There will be two rounds of calls for associated projects that could be financed on topics that match PlanetData and that foster collaboration between the NoE and external organizations; participants were called to influence the programs and to take part in them.

Schema-agnostic Search Frontends (Thanh Tran, KIT): This presentation was dedicated to search and data source exploration without a priori knowledge of the schema. The presented approach was termed schema-agnostic faceted search and is a possible solution towards providing support for discovery of possible unknown data. The main problem is related to aggregation of distributed data of different granularity, complexity and relevance.

Services and Semantics (Carlos Pedrinaci, John Domingue, Open University): The presentation addressed the issue of semantic data in the world of Semantic Web services. The argument is that linked data is growing significantly faster than the open Web services, and that making available such linked data in smart ways opens up many new business opportunities. Technologically speaking, in the center of attention was a recent development termed iServe which is a publishing platform for Semantic Web services that allows bringing the world of services closer to the world of linked data; i.e. through iServe it is intended to bring Semantic Web services to LOD.

Bringing Semantics at the Street Level (Massimo Paolucci, DOCOMO Euro-Lab): This presentation was about the use of semantics in the context of mobile services and ubiquitous computing. Mobile phones are gateways to services (semantics for services) and information sources (semantics for data and context) with specific requirements. The use of semantics in mobile services is diverse: data interoperability and aggregation, service provisioning, contextual and situational reasoning, and event recognition. Some of the known problems are related to streams (needed to provide context information on the fly), dynamics of data and mobility of users, and quality of data (dirty writes and sensor data).

A very pragmatic view towards: Life Sciences and Health Care Vertical (Vassil Momtchev, Ontotext): The presentation intended to give a better idea of the life science domain which is a very early adopter of semantic technologies. The (public) datasets in life science are extremely large, but still, in bio scenarios complexity is a bigger issue than scale, and applying expressive semantics an even bigger one. Looking at the public data (mostly exposed as part of Linked Life Data / Linked Open Data) there are three main problems to be solved: i) data quality, ii) licensing, and iii) tools for data exploration. Future trends in the domain point towards personalized medicine and electronic health records, new business models for the pharmaceutical industry and the genomics revolution.

Information Workbench for Interacting with Linked Data (Peter Haase, fluid Operations): The business of fluid Operations is to provide management solutions for enterprise clouds. This includes business layer solutions but also infrastructure layer support. The main topic of this presentation was their information workbench which offers integrated management of heterogeneous data sources, and tools for supporting the entire lifecycle of interacting with data. The workbench provides semantic search (schema-agnostic, faceted search and query interpretation) and widget-based user interfaces including mashups with external sources, automated selection of widgets and customization/personalization.

List of activities

On the following pages a list of concrete follow up activities is given. Further general notes about the meeting are given first:

- Potentially related events are the WebDB or CloudDB workshops
- SemData is at the intersection of Semantic Web, Web, DB, Cloud, KM, IR/IE, visualization, services, ML and verticals
- It is important to reach out to these target communities; e.g. the VLDB workshop
- A tutorial element would be nice for learning about state-of-the-art in databases, semantic web services, linked data... Such tutorials could be held in combination with workshops.
- The scoping of the workshop was really good, and having such events on invitation only is a good thing.

1. Instantiation of an Organizing Committee for SemData:

- Atanas Kiryakov, Ontotext
- Dieter Fensel, STI Innsbruck
- Elena Simperl, KIT
- Kavitha Srinivas, IBM Watson
- Reto Krummenacher, STI Innsbruck
- ???

2. Instantiation of a Steering Committee for SemData: The invited participants to the SemData@Sofia Roundtable are the initial steering committee of the initiative.

3. Establishment of an Advisory Board – proposals for representative members of related communities are collected by Atanas Kiryakov. The communities that are recognized to be related to SemData and that at least partly touch or overlap with the roadmap of the initiative are:

- Semantic Web and Linked Data
- Web technology and cloud computing
- Database research
- Information retrieval / information extraction
- Knowledge management and reasoning
- Information visualization
- Machine learning
- (Web) services
- Vertical markets

4. Organization of SemData@ESWC (May 30, Heraklion, Greece) as the second event of the series with on invitation only basis for correspondence with researchers are practitioners

from related communities that were not the core target of SemData and hence not invited to the roundtable. The organization team for SemData@ESWC: Kavitha Srinivas, Grigoris Antoniuo, Elena Simperl, Atanas Kiryakov, and Reto Krummenacher.

5. Organization of SemData@VLDB (September 17, Singapore) as an official VLDB workshop and the special intention to more concretely target the database community. There is a public Call for Paper (<http://semdata.org/events/2010/vldb>) and accepted submission will be included in the thumb drive of the official VLDB 2010 proceedings. The co-chairs of SemData@VLDB are Karl Aberer, Atanas Kiryakov, Reto Krummenacher and Rajaraman Kanagasabai from the Institute for Infocomm Research in Singapore.
6. Organization of SemData@ESTC (December 2010, Vienna) as an event targeted at vendors and verticals. The organizer team is still to be determined.
7. Investigation towards special issues in
 - VLDB journal
<http://www.springer.com/computer/database+management+&+information+retrieval/journal/778>
The VLDB Journal publishes a number of special issues in addition to the regular ones; special issues focus on information that the Editors and the Editorial Board determine to be of importance to the database community. The SemData@VLDB workshop organizers take care of the appropriate steps and timing.
 - Journal of Web Semantics
www.elsevier.com/wps/find/journaldescription.cws_home/671322/description
JWS accepts special issues – there was recently a call published for an issue on Web-scale Semantic Information Processing by Heiner Stuckenschmidt and Jeff Heflin: Submission deadline July 1, 2010.
8. Benchmarking for Semantic Data Management. There is a definite need for more real-world benchmarks for semantic data reflecting better the graphiness of RDF data. At OpenLink Software the Virtuoso team has recently initiated such work with the BotNet Benchmark that simulates an RDF OLTP backend of a Website of a social network (<http://esw.w3.org/BotNetBenchmark>). More such benchmarking activities have to come out of SemData that pay tribute to the particular characteristics of linked data. An initial roadmap/proposal will be established by Orri Erling, Kavitha Srinivas, Bryan Thompson and Atanas Kiryakov.
9. Dagstuhl Seminar 2011/2012: A Dagstuhl seminar is seen to be a very good format for a more intensive and long-lasting meeting on semantic data management. A seminar in 2011 of 2012 should be targeted to have more detailed and high quality discussions on the most central topics of SemData: databases, benchmarking and application layers. There is an upcoming deadline for calls set to April 15, 2010 by the Dagstuhl directorate.

10. Semantic Data Management Track at ESWC2011: To move beyond the organization of workshops and to increase the impact, we should target an own track on semantic data management at next year's ESWC.
11. White Paper: The OC of SemData plus Orri Erling will author a white paper under the initial lead of Dieter Fensel.
12. Summer School 2011: The NoE PlanetData is planning to organize a first summer school on SemData-related topics in June 2011. This school will take place most likely the week before or after ESWC 2011. The plan is to influence the summer school with valuable insights from the SemData community.
13. Open Calls of the PlanetData Network of Excellence: The SemData initiative is asked to actively contribute to the creation of the programs by influencing the topics and scope of the calls. Moreover, the PlanetData programs also offer a platform for smaller project that emerge from the initiative to receive funding.
Timeline
 - 1st call published end of March 2011
 - 1st round of associated projects October 2011 – March 2013
 - 2nd call published end of October 2012
 - 2nd round of associated projects: March 2013 – September 2014
14. Further activities for the future include the potential creation of an own journal or the authoring of a book about semantic data management. Such activities provide other interesting support for the establishment of a community and its impact.

Contact

© Semantic Data Management Initiative: www.semdata.org

Reto Krummenacher

STI Innsbruck, University of Innsbruck

Email: reto.krummenacher@sti2.at

Telephone: +43 512 507 6452