

# Semantic Data Management

Edited by

Grigoris Antoniou<sup>1</sup>, Oscar Corcho<sup>2</sup>, Karl Aberer<sup>3</sup>, Elena Simperl<sup>4</sup>,  
and Rudi Studer<sup>5</sup>

- 1 University of Huddersfield – Huddersfield, UK, [g.antoniou@hud.ac.uk](mailto:g.antoniou@hud.ac.uk)
- 2 Universidad Politécnica de Madrid – Madrid, ES, [ocorcho@fi.upm.es](mailto:ocorcho@fi.upm.es)
- 3 École Polytechnique Fédérale de Lausanne – Lausanne, CH,  
[karl.aberer@epfl.ch](mailto:karl.aberer@epfl.ch)
- 4 Karlsruhe Institute of Technology – Karlsruhe, DE, [elena.simperl@kit.edu](mailto:elena.simperl@kit.edu)
- 5 Karlsruhe Institute of Technology – Karlsruhe, DE, [studer@kit.edu](mailto:studer@kit.edu)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 12171 “Semantic Data Management”. The purpose of the seminar was to have a fruitful exchange of ideas between the semantic web, database systems and information retrieval communities, organised across four main themes: scalability, provenance, dynamicity and search. Relevant key questions cutting across all of these themes were: (i) how can existing DB and IR solutions be adapted to manage semantic data; and (ii) are there new challenges that arise for the DB and IR communities (i.e. are radically new techniques required)? The outcome was a deeper, more integrated understanding of the current state of the art on semantic data management and a the identification of a set of open challenges that will inform the three communities in this intersection.

**Seminar** 22.–27. April, 2012 – [www.dagstuhl.de/12171](http://www.dagstuhl.de/12171)

**1998 ACM Subject Classification** H.2 Database Management, D.3.1 Formal Definitions and Theory

**Keywords and phrases** Semantic data, Semantic Web, Linked Data, Large-scale data management, Dynamicity and stream processing, Provenance and access control, Information retrieval and ranking

**Digital Object Identifier** 10.4230/DagRep.2.4.39

## 1 Executive Summary

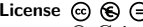
*Grigoris Antoniou*

*Oscar Corcho*

*Karl Aberer*

*Elena Simperl*

*Rudi Studer*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Grigoris Antoniou, Oscar Corcho, Karl Aberer, Elena Simperl, Rudi Studer

The Semantic Web represents the next generation World Wide Web, where information is published and interlinked in order to facilitate the exploitation of its structure and semantics (meaning) for both humans and machines. To foster the realization of the Semantic Web, the World Wide Web Consortium (W3C) developed a set of metadata (RDF), ontology languages (RDF Schema and OWL variants), and query languages (e.g., SPARQL). Research in the past years has been mostly concerned with the definition and implementation of



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license

Semantic Data Management, *Dagstuhl Reports*, Vol. 2, Issue 4, pp. 39–65

Editors: Grigoris Antoniou, Oscar Corcho, Karl Aberer, Elena Simperl, and Rudi Studer



DAGSTUHL  
REPORTS

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

these languages, the development of accompanying ontology technologies, and applications in various domains. This work has been very successful, and semantic web technologies are being increasingly adopted by mainstream corporations and governments (for example by the UK and USA governments) and by several Science communities (for example, Life Sciences or Astronomy). Moreover, semantic technologies are at the core of future developments, e.g. in the UK Open Data Institute. However, compared to more traditional solutions, semantic technologies often appear to be immature, and current tools lag behind in terms of efficiently handling of large data sets. What are additionally needed include solid data management concepts, architectures, and tools that follow the paradigms of more traditional database (DB) and information retrieval (IR) systems. Semantic data management refers to a range of techniques for the manipulation and usage of data based on its meaning. The aim of this workshop was to discuss in-depth a number of crucial issues, with particular emphasis on the fruitful exchange of ideas between the semantic web, database systems and information retrieval communities. Relevant key questions cutting across all topics covered were: (i) how can existing DB and IR solutions be adapted to manage semantic data; and (ii) are there new challenges that arise for the DB and IR communities (i.e. are radically new techniques required)? For the purposes of this workshop, and for this report, we understand semantic data simply as data expressed in RDF, the lingua franca of linked open data and hence the default data model for annotating data on the Web. The workshop was organized along the following key themes:

1. **Scalability.** In order to make semantic technologies take on the targeted market share, it is indispensable that technological progress allows semantic repositories to scale to the large amount of semantic data that is already available and keeps growing. It is essential to come close to performance parity with some of the best DB solutions without having to omit the advantages of a higher schema flexibility compared to the relational model. Moreover, the exploitation of semantic data on the Web requires managing the scale that so far can only be handled by the major search engine providers. However, this should be possible without losing the advantages of a higher query expressivity compared to basic key-value stores and IR solutions.
2. **Provenance.** An important aspect when integrating data from a large number of heterogeneous sources under diverse ownership is the provenance of data or parts thereof; provenance denotes the origin of data and can also include information on processing or reasoning operations carried out on the data. In addition, provenance allows for effectively supporting trust mechanisms and policies for privacy and rights management.
3. **Dynamicity.** Another important property of many (semantic) data is its dynamicity. While some data, such as public administration archives or collections of text documents might not change too frequently, other data, coming from sensors, RSS, user-generated content (e.g. microblogging), etc., might evolve on a per millisecond basis. The effects of such changes have to be addressed through a combination of stream processing, mining, and semantics-based techniques.
4. **Search and Ranking.** The large and growing amount of semantic data enables new kinds of applications. At the same time, more data means that ultimately, there might be more results produced from it that one can or desires to inspect. Data and results to concrete information needs vary in the degree of relevance. Effective ranking mechanisms that incorporate the information needs as well as contextual information into account can deliver and rank pertinent results and help the users to focus on the part of the data that is relevant.

Each day of the workshop was scheduled to deal with one of the four key themes, in the order that has been presented above. For every theme there was an initial tutorial presentation that aimed at setting the context and vocabulary to be used throughout the day, as well as expose the current state of the art in the theme and the open challenges. Such tutorial presentation also helped to introduce the more in-depth research topics dealt with later on by the other presenters. Some of the main results of the discussions held during the workshop were the following, separated again by themes:

### Scalability (day 1)

Issues that have been dealt with so far in the semantic data management community are how to store semantic data, how to index it and how to deal with processing and optimizing complex queries. Solutions that have been proposed work for the small datasets and scale that we know from the previous years. Meanwhile, the amount of data published on the Web including linked data and RDFa data associated with Web pages has exploded, resulting in billions of RDF triples – a number that is increasing very fast. Existing techniques and solutions no longer scale to this scenario. It requires the adoption of mainstream and proven DB and IR technologies and possibly, new radical solutions. Some of the main questions addressed at the workshop were (a) how DB solutions capable of dealing with large amounts of data with flexible schema or even schema-less data can be adopted for semantic data management. Likewise, IR solutions to data management have proven to be robust and scale to the Web. (b) How can this body of work be adopted and extended to deal with not only keyword search but also complex queries? In particular, the specific topics that require interests are (c) storage and indexing of large-scale semantic data, (d) parallel SPARQL query processing and optimization, (e) federated query processing on linked data, (f) top-k and approximate query processing, (g) workload-adaptive dynamic load balancing and optimization and (h) semantic data management in the cloud.

The theme was introduced by Philippe Cudre-Mauroux (University of Fribourg) through a tutorial that reviewed some of the recent techniques to support vertical (scale-up) and horizontal (scale-out) scalability for Semantic Web systems. Several presentations followed that were focused not only on demonstrating how some scalability aspects had been addressed in different systems, but also at clarifying what scalability means for semantic data management, and how it relates to the DB community. These were presentations from Avigdor Gal (Technion), Frank van Harmelen (VU), Peter Haase (fluidOps), Juan Sequeda (UT) and Charalampos Nikolau (NKUA). These were complemented with presentations from use cases in the financial and medical domains, from Steve Harris (Garlik) and Satya Sahoo (Case Western Reserve University).

The rest of presentations and discussions of the day, including part of the research groups, were focused on benchmarking. Two specific presentations on benchmarking were delivered by Kavitha Srinivas (IBM) on the characterisation of benchmarks, and Jeff Heflin (Lehigh University), on the experience with LUBM. These were the basis of the working group that was held later on benchmarking, and were important for several discussions throughout the week.

Some of the conclusions obtained from the discussions were related to the relationships between relational models and graph data models, and how difficult it was to express this difference clearly across communities. In particular, it was rather clear that the semantic web community is still too constrained in the work related to trying to make the underlying

technology store and access RDF, rather than exposing the benefits of semantics. In fact, work and discussions should not be focused on comparing both worlds, but on characterising the workloads and usage that drive the use of each type of technology, and the tipping point of when to use each. A message of obtaining scalability not only in seconds, but also in months (in terms of the benefits that semantic technology can bring in in issues like schema heterogeneity and data integration) were also pointed out.

## Provenance (day 2)

Provenance is an important theme in the database systems and Web communities. There are two main research trends related to provenance management: workflow and data provenance. In the case of data provenance, the underlying model deals with declarative queries that include operations such as joins and unions, whereas workflow provenance models typically describe procedural workflows and involve operations that are treated as black boxes because of their complexity. The W3C Linking Open Data effort has boosted the publication and interlinkage of large amounts of heterogeneous RDF datasets on the Semantic Web and the result of this work is the Linked Data paradigm. In this context, the big challenges that must be addressed are related to the evaluation of linked data properties such as quality, trustworthiness, reliability to support a variety of applications including data integration and fusion, and mashups. In order to be able to evaluate the aforementioned qualities it is critical to record the provenance which denotes the origin of data that can also include information on processing or reasoning operations carried out on the data.

Paul Groth (VU) was in charge of the initial tutorial where the basic concepts of provenance were explained, and was accompanied next by Olaf Hartig (HU Berlin), Luc Moreau (University of Southampton) and Paolo Missier (Newcastle University) on the introduction to the W3C PROV family of specifications. Such tutorial and introduction to W3C activities were a good starting point to understand the current status of work on provenance by the semantic data management community. Besides, results from the recent Dagstuhl seminar on Provenance were also shared with the attendees.

The first presentation after the tutorial was focused on how provenance is used in BioPortal to provide access control in SPARQL. This was presented by Manuel Salvadores (Stanford University). Then the rest of presentations were focused on technologies and approaches to represent and exploit provenance in different contexts. These were presentations by Bryan Thompson (SYSTAP), which proposed a novel manner to represent provenance for triples or quads, and James Cheney (University of Edinburgh), who presented database wiki. There was also one presentation on workflow provenance done by Kerry Taylor (CSIRO).

Some of the conclusions obtained from the provenance sessions were related to the fact that this is a clear community with clear problems, and already quite well organised (for instance, through the corresponding W3C group on provenance). Some of the current limitations, and calls for action were on the need to make provenance data sets available and understandable, and address some of the interoperability issues that currently exist between different languages and models for describing provenance. Some of the open research issues that need to be addressed are (a) the definition of provenance models that take into account schema information to represent provenance information independently of the underlying applications (b) the extension of the existing query languages to support implicit provenance (d) querying and reasoning with provenance data (e) efficient storage of provenance data in the presence of data evolution (f) the use of provenance in applications

such as view maintenance and updates, access control, debugging, attribution of responsibility, assessment of trustworthiness and information quality (g) presentation and visualization of provenance information, (h) provenance for probabilistic RDF data, and (i) provenance in scientific workflows in order to get insights on the quality and integrity of results, and to better understand scientific results, while also reaching the more general business workflow community. The possibility of organising a panel at one of the subsequent BPM conferences was considered as a nice follow-up. Another nice analysis is provided by Paul Groth in his blog, at <http://thinklinks.wordpress.com/2012/05/01/dagstuhl-semantic-data-management/>

### **Dynamicity (day 3)**

Several areas of work address the problems related to the dynamicity of data sources, many of them common across types of data sources and leading to general research issues such as: (a) generating models for describing and understanding the data sources and the evolution of their data in time, considering not only primary data sources, but also derived ones, whose changes normally depend on the ones that they are based on. There are also specific problems related to the intrinsic characteristics of each data source. For example, problems in sensor and stream data management are mainly related to: (b) the efficient processing and storage of this type of data, (c) to the standardization of their representation using semantic models, (d) to the integration of heterogeneous sources, and to their analysis in order to detect patterns, trends, complex events, but also anomalies, by means of different types of techniques, including learning and summarization techniques.

Kerry Taylor (CSIRO) was in charge of presenting the initial tutorial explaining what we normally understand by dynamicity and how it is tackled on the Web (mainly Data Stream Management Systems and Complex Event Processing) and on the Semantic Web (with approaches focused on ontology versioning as streams of "slow" changes, and streaming extensions of SPARQL that have appeared lately). She also described current efforts in providing semantics to sensor networks and their related data, such as the W3C Semantic Sensor Network (SSN) Ontology, and finished her keynote with a discussion on some of the upcoming challenges in this area: pushing semantics into little devices, catching up with the work on semantics that is being performed by the Open Geospatial Consortium (OGC) and working on effecting change, not just responding to it.

Then we had several sessions on presentations about dynamicity in the context of data streams and sensor data, from Spyros Kotoulas (IBM Research), Manfred Hauswirth (DERI), Oscar Corcho (UPM) and Ivana Podnar Zarko (University of Zagreb), on the need to consider the dynamic nature of endpoints and Linked Data URIs while querying, from María Esther Vidal (Universidad Simón Bolívar) and Olaf Hartig (HU Berlin), and on dynamicity in scientific data, services and workflows, from José Manuel Gómez-Pérez (iSOCO).

In the context of semantic sensor data and data stream processing, several use cases were proposed, many of which where in the context of Smart Cities (traffic, emergency management, black water distribution, energy prediction, etc.) and in environmental sensing. Some of the open challenges in this area are related to how to make sure that most of the processing and RDF management could be done at the sensor level, instead of the middleware level, given the limitations of these devices. how to handle private and public data, static and dynamic data, provenance, access rights, etc. The need for appropriate benchmarks and experimental facilities was also highlighted, and the need to standardise the semantic stream query language to be used and the Linked stream standards.

In the context of dynamicity of endpoints and Linked Data, the need for adaptive semantic data management techniques was highlighted, and several current proposals in this direction were presented. Besides, a novel approach for querying the Web of Linked Data, focused on dereferencing URIs instead of using directly endpoints (e.g., as in the SQUIN system), was presented and heavily discussed among attendees.

Finally, in the context of dynamicity in scientific domains (workflows, data sources, services), the need to identify and handle appropriate evolution and lifecycle models for all these artifacts, so as to ensure reproducibility in Science, was presented and discussed, together with lessons learned from some ongoing work on scientific workflow preservation, such as the need to monitor decay, to understand data with a purpose, and to encapsulate parts of workflows and services.

Some of the main conclusions of the discussions held during this day were related to **the need to cover all types of streaming data sources** (from sensor data to social media) as new types of data sources that may benefit from the use of semantics in different forms. In the context of sensor data sources, work is clearly underway by different groups and through joint initiatives at W3C, and some of the next challenges are related to the needs of the Internet of Things community, in scenarios like, for instance, Smart Cities. Another conclusion was that the Web is dynamic, and we have to be able to **provide solutions that are able to adapt to this dynamicity**, including query support when no triple stores are available in Web servers, we have to provide a better characterisation of the balance between caching and non-caching, and intermediate solutions, and we have to cope with run-time discovery of semantic data sources, query planning to them, etc., as well as data and processes decay.

## Search and Ranking (day 4)

In the IR community, the central question has always been the one about relevance. Which results are relevant to the query posed by the user? Recently, this question has attracted interests of DB researchers. Instead of retrieving documents, which constitute the primary subject in the IR field, query processing in the DB community is about computing complex results tuples. Correspondingly, different notions of relevance have been investigated. There exists a broad range of applications for semantic data which might involve simple IR-like queries as well as complex tasks in the fashion of DB-style relational queries, online analytical processing and reasoning. Also from the data perspective, semantic data shares both DB- and IR-like characteristics. Semantic data found on the Web might be clean and highly-structured just like relational data in the database or messy, schema-less and possibly, contain a large-amount of text just like data in an IR system. For ranking results and enabling users to focus on relevant semantic data, we need to study, adopt and extend existing notions of relevance proposed by IR and DB researchers. Some of the specific questions discussed during the workshop were (a) how can IR-approaches such as the probabilistic model or the most recent state of the art of generative models be applied to the problem of semantic data ranking? Is this enough or (b) do we also need to take the explicit structure and semantics of the data into account, and (c) can we leverage some existing DB-approaches for that? Do we need hybrid models combining IR and DB solutions to deal with the structured and unstructured nature of semantic data? (d) What other new challenges are introduced by semantic data and correspondingly, what new concepts are required to address them?

Duc Thanh Tran (KIT) was in charge of the initial tutorial explaining what we understand

by semantic search (using semantics in structured data, conceptual and lexical models) and why it is important to consider it. He explained several of the techniques currently used for semantic search in the state of the art and discussed about several research directions, among which he pointed out to keyword search over structured/semantic data, supporting complex information needs (long tail), and requiring efficient traversal algorithms and specialized indexes, and various models for ranking). Among the selected challenges identified in this talk we can cite the need for benchmarking, the need to deal with hybrid content management, and ranking hybrid results, the need to study the querying paradigm to use for complex retrieval tasks (keywords, natural language, facets), and the possibility of using semantics or providing a broader search context.

This presentation was followed by several other presentations on the use of probabilistic models to check consistency between links of different datasets, from Edna Ruckhaus (Universidad Simón Bolívar), on ranking SPARQL results, from Ralf Schenkel (Universität de Saarlandes), and an example application of search mechanisms applied to a concrete case study on crystals, from Kerry Taylor (CSIRO). Later on some additional presentations were done on some of the current limitations of SPARQL in terms of complexity, from Marcelo Arenas (PUC).

The last two presentations, and part of the discussion, went about the role of crowdsourcing for search and for the provisioning of metadata, with talks from Gianluca Demartini (University of Fribour) and Wolf-Tilo Balke (TU Braunschweig).

One of the main conclusions from this day and the discussions held during the day were related to the fact that this theme is the one that shows a stronger relationship between the three communities at which this workshop was addressing (IR/DB/SW), and hence it is one of the areas where there is a larger possibility of more **cross-fertilisation** and of finding more research and further development opportunities. Some additional examples to semantic search will be gathering evidences from documents, combining data sources while answering queries, applying provenance in IE pipelines, etc.

Another important conclusion and step forward that was identified was related to the context of **crowdsourcing and semantics**, which is well extended in the IR community. It was clear throughout discussions that we cannot talk yet about a Semantic Read/Write Web in a general manner, since the SW community is still operating in a Read/Web1.0 context (a few publishers only) instead of in a Read/Write Web2.0 model (lots of producers). Some opportunities may arise in this context in the application of this paradigm for complex tasks like link validation, improving the quality of Linked Data, etc. What it was not so clear was whether it was easy as well to provide support from semantic technologies to some of the existing crowdsourcing activities that are being held now outside of the SW community. In this sense, there was a proposal to organise in the future a workshop or roadmapping activity on this topic.

## 2 Table of Contents

### Executive Summary

*Grigoris Antoniou, Oscar Corcho, Karl Aberer, Elena Simperl, Rudi Studer* . . . . 39

### Overview of Talks

Scalability in Semantic Data Management  
*Philippe Cudre-Mauroux* . . . . . 48

When Scalability Meets Uncertainty  
*Avigdor Gal* . . . . . 48

Semantics and Big Data challenges: Beyond Volume  
*Peter Haase* . . . . . 49

SPARQL as fast as SQL  
*Juan Sequeda* . . . . . 49

Building Scalable Semantic Geospatial RDF Stores  
*Charalampos Nikolaou* . . . . . 49

Awakening Clinical Data: Semantics for Scalable Medical Research Informatics  
*Satya S. Sahoo* . . . . . 50

Building better RDF benchmarks  
*Kavitha Srinivas* . . . . . 50

Provenance: some helpful concepts  
*Paul Groth* . . . . . 50

An Overview on W3C PROV-AQ: Provenance Access and Query  
*Olaf Hartig* . . . . . 51

Access Control in SPARQL, The BioPortal Use Case  
*Manuel Salvadores* . . . . . 51

Reification made easy  
*Bryan Thompson* . . . . . 52

Reaping the Rewards: What is the provenance saying?  
*Kerry Taylor* . . . . . 52

Database Wiki and provenance for SPARQL updates  
*James Cheney* . . . . . 53

Dynamicity: Sensor Networks, Streaming Data, and Designing for Change  
*Kerry Taylor* . . . . . 54

Linking the real world  
*Manfred Hauswirth* . . . . . 54

Semantic data streams: does it make sense?  
*Oscar Corcho* . . . . . 55

Probabilistic Processing of Sliding Window Top-k Queries  
*Ivana Podnar Zarko* . . . . . 55

Adaptive Data Management Techniques for Federations of Endpoints  
*Maria-Esther Vidal* . . . . . 56




Conceiving the Web of Linked Data as a Database <i>Olaf Hartig</i> . . . . .	57
Scientific Data Management – From the Lab to the Web <i>José Manuel Gómez-Pérez</i> . . . . .	57
Semantic Search Tutorial <i>Duc Thanh Tran</i> . . . . .	57
Probabilistic Models and Reasoning for Link Consistency Checking and Validation <i>Edna Ruckhaus</i> . . . . .	58
Ranking SPARQL Results <i>Ralf Schenkel</i> . . . . .	58
The Search for Crystals <i>Kerry Taylor</i> . . . . .	59
Paths in semantic search: A back and forth story <i>Marcelo Arenas</i> . . . . .	60
Getting Semantics from the Crowd <i>Gianluca Demartini</i> . . . . .	60
Efficient Crowdsourcing for Metadata Generation <i>Wolf-Tilo Balke</i> . . . . .	61
<b>Working Groups</b> . . . . .	61
<b>Participants</b> . . . . .	65

### 3 Overview of Talks

This section is organised according to the four themes of the workshop, and following a chronological order. For every theme, there was a presenter in charge of providing an overview of the current status of the theme, so as to get all participants up to speed. Afterwards, several talks were scheduled that went into depth into specific aspects of the theme, leading later to working group breakouts.

#### 3.1 Scalability in Semantic Data Management

*Philippe Cudre-Mauroux (University of Fribourg, CH)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Philippe Cudre-Mauroux


**Main reference** M. Wylot, J. Pont, M. Wisniewski, P. Cudré-Mauroux, “dipLODocus[RDF] – Short and Long-Tail RDF Analytics for Massive Webs of Data,” ISWC 2011, LNCS, Vol. 7031, pp. 778–793, Springer-Verlag, 2011.

**URL** [http://dx.doi.org/10.1007/978-3-642-25073-6\\_49](http://dx.doi.org/10.1007/978-3-642-25073-6_49)

Scalability is still one of the major issues of the Semantic Web. In this short tutorial, I will review some of the recent techniques to support vertical (scale-up) and horizontal (scale-out) scalability for Semantic Web systems.

#### 3.2 When Scalability Meets Uncertainty

*Avigdor Gal (Technion – Haifa, IL)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Avigdor Gal


There is a tight relationship between semantics, scalability, and uncertainty. To start off, whenever there’s data there’s semantics. Our continuous attempt to include more and more semantics in our data to make it more usable impacts scalability. Whenever semantics is missing, uncertainty grows. Therefore, there is a trade-off between scalability and uncertainty when semantics of data is managed.

In the seminar I presented the relationships between the three concepts, using schema matching as a case in point. In schema matching, we aim at matching between elements of data representations. Typically, such elements are designed without giving much thought to accurate semantics definition. This, in turn, increases uncertainty, and when trying to manage it, scalability comes into play.

The European project NisB (<http://www.nisb-project.eu/>) aims at providing scalable methods to deal with uncertainty in the matching process.

### 3.3 Semantics and Big Data challenges: Beyond Volume


*Peter Haase (fluidOps, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Peter Haase

Big Data means more than volume and vertical scale. Our community has a role to play in Big Data. There are many things that we can contribute: Ontologies as conceptual models to access big data; Integration of diverse, heterogeneous data sources; and (Light-weight) inference for data analytics. There is also some areas that we need to work on: Ability to deal with streaming data, and Horizontal scalability and data virtualization

### 3.4 SPARQL as fast as SQL


*Juan Sequeda (University of Austin at Texas, US)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Juan Sequeda

The Semantic Web promise to achieve web-wide data integration requires the inclusion of legacy relational data as RDF, which, in turn, requires the execution of SPARQL queries on the legacy relational database. Two independent studies have shown that current Relational Database to RDF (RDB2RDF) systems that execute SPARQL on relational databases are several orders of magnitude slower than executing the semantically equivalent SQL query directly on the RDBMS. In this presentation, I explore the hypothesis: existing commercial relational databases already subsume the algorithms and optimizations needed to support effective SPARQL execution on existing relationally stored data. I will present two important optimizations: detection of unsatisfiable conditions and self-join elimination, which are implemented by commercial RDBMS, such that, when applied, SPARQL queries execute at nearly the same speed as semantically equivalent native SQL queries, providing strong evidence of the validity of the hypothesis.

### 3.5 Building Scalable Semantic Geospatial RDF Stores

*Charalampos Nikolaou (National and Kapodistrian University of Athens, GR)*


License  Creative Commons BY-NC-ND 3.0 Unported license  
© Charalampos Nikolaou

**Joint work of** Koubarakis, Manolis; Kyzirakos, Kostis; Karpathiotakis, Manos; Nikolaou, Charalampos; Garbis, George; Sioutis, Michael; Bereta, Konstantina;

We present our geospatial extensions of RDF and SPARQL, called stRDF and stSPARQL respectively. stRDF extends RDF with the ability to represent geospatial and temporal information using linear constraints. stSPARQL extends SPARQL for querying stRDF data. In their new versions, stRDF and stSPARQL use formats (WKT and GML) that have become standards by the Open Geospatial Consortium (OGC) for the encoding of spatial information. As a result, stSPARQL is very close to GeoSPARQL, the standard query language of OGC for querying geospatial information in RDF. Furthermore, we discuss the implementation of our open source system Strabon, which is a scalable semantic geospatial RDF store for stRDF/stSPARQL. We present some preliminary experimental evaluation results based on real linked geospatial datasets and synthetic ones. Last, we briefly discuss RDFi, our extension of RDF with the ability to represent and query RDF data with incomplete information.

### 3.6 Awakening Clinical Data: Semantics for Scalable Medical Research Informatics

*Satya S. Sahoo (Case Western Reserve University, US)*


License  Creative Commons BY-NC-ND 3.0 Unported license  
© Satya S. Sahoo

Health care data is growing at an explosive rate, with highly detailed physiological processes being recorded, high resolution scanning techniques (e.g. MRI), wireless health monitoring systems, and also traditional patient information moving towards Electronic Medical Records (EMR) systems. The challenges in leveraging this huge data resources and transforming to knowledge for improving patient care, includes the size of datasets, multi-modality, and traditional forms of heterogeneity (syntactic, structural, and semantic). In addition, the US NIH is emphasizing more multi-center clinical studies that increases complexity of data access, sharing, and integration. In this talk, I explore the potential solutions for these challenges that can use semantics of clinical data – both implicit and explicit, together with the Semantic Web technologies. I specifically discuss the ontology-driven Physio-MIMI platform for clinical data management in multi-center research studies.

Further Details: [http://cci.case.edu/cci/index.php/Satya\\_Sahoo](http://cci.case.edu/cci/index.php/Satya_Sahoo)

### 3.7 Building better RDF benchmarks

*Kavitha Srinivas (IBM TJ Watson Research Center – Hawthorne, US)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Kavitha Srinivas

RDF is most useful when the data exhibit the following characteristics (a) have variable schema, (b) there is data sparsity, (c) queries are complex, so key based lookups will fail, (d) queries are such that denormalizing the data so a key based lookup is not an option, (e) inference is needed. It would be useful to characterize RDF benchmarks along these characteristics. Even better would be to treat these as different dimensions along which a given benchmark could be varied. We have demonstrated how this might be accomplished for one dimension (that of sparsity) in a SIGMOD 2012 paper but it would be useful to think of how to do this for other dimensions.

### 3.8 Provenance: some helpful concepts

*Paul Groth (Free University – Amsterdam, NL)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Paul Groth

Knowing the provenance (the origin, source or history) of information and things is vital in many domains. In open information systems, like the web, that integrate information, it is imperative to know what the information sources are, what information the sources contain, and how the information has been combined. In science, provenance enables reproducibility. In the blogosphere, provenance enables credit to be spread. In government, provenance enables transparency. In this talk, a brief overview of the research in provenance is given

structured using three dimensions: content, management, and use. In this context, I identify some helpful concepts including the following examples: provenance is by its nature about the past; it extends across multiple systems and thus interoperability is required; it can be used for trust calculation but is not itself trust and there may be multiple views of provenance.

### 3.9 An Overview on W3C PROV-AQ: Provenance Access and Query

*Olaf Hartig (HU Berlin, DE)*

**License** © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license  
 © Olaf Hartig  
**Joint work of** Klyne, Graham; Groth, Paul; Moreau, Luc; Hartig, Olaf; Simmhan, Yogesh; Myers, James; Lebo, Timothy; Belhajjame, Khalid; Miles, Simon;  
**Main reference** G. Klyne, P. Groth (eds.), “PROV-AQ: Provenance Access and Quer,” W3C Working Draft, 19 June 2012.  
**URL** <http://www.w3.org/TR/prov-aq/>

This short talk introduces the "Provenance Access and Query" (PAQ) document which is part of the PROV family of documents developed by the W3C Provenance Working Group. The purpose of PAQ is to describe how to locate, retrieve, and query provenance information on the Web. The talk will briefly introduce the following main contributions of PAQ: 1) A simple mechanisms for discovery and retrieval of provenance information, and 2) More advanced discovery service and query mechanisms. Finally, we will point out some of the open issues of the current version of PAQ.

### 3.10 Access Control in SPARQL, The BioPortal Use Case


*Manuel Salvadorés (Stanford University, US)*

**License** © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license  
 © Manuel Salvadorés  
**URL** <http://www.stanford.edu/~manuelso/presentations/dagstuhl.html>

BioPortal is a repository of biomedical ontologies – the largest such repository, with more than 300 ontologies to date. This set includes ontologies that were developed in OWL, OBO and other formats, as well as a large number of medical terminologies that the US National Library of Medicine distributes in its own proprietary format. We have published the RDF version of all these ontologies at <http://sparql.bioontology.org>. This dataset contains 190M triples, representing both metadata and content for the 300 ontologies. We use the metadata that the ontology authors provide and simple RDFS reasoning in order to provide dataset users with uniform access to key properties of the ontologies, such as lexical properties for the class names and provenance data. The dataset also contains 9.8M cross-ontology mappings of different types, generated both manually and automatically, which come with their own metadata.

### 3.11 Reification made easy

*Bryan Thompson (SYSTAP – Greensboro, US)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Bryan Thompson

Joint work of Thompson, Bryan; Personick, Mike


Many domains require the ability to make statements about statements. This shows up in graph databases as link attributes which are used to attach weights. It shows up in entity link extraction as confidence values assigned by the extractor. It is required in intelligence analysis in order to have provenance for each "fact" that trace the evidence for and against those facts as analysts reason about them. It is required for security in systems which require datum level security.

We propose a new indexing scheme for RDF data which makes it possible to efficiently index the data and answer queries about both ground statements and metadata about those statements.

There was strong interest in this topic during the seminar. An combined academic and industry group was formed during the seminar consisting of Olaf Hartig (Humbolt University), Tran Thanh (KIT), Orri Erling and Yrjana Rankka (OpenLink), and Bryan Thompson and Mike Personick (SYSTAP, LLC). The group has formalized the model theory for the proposed approach and reconciled it with the RDF model theory and the SPARQL algebra. Both OpenLink and SYSTAP are pursuing implementations and several other vendors have expressed an interest in the approach. Our goal is to publish the output of this collaboration and to develop and submit a proposal for standardization to the W3C.

### 3.12 Reaping the Rewards: What is the provenance saying?


*Kerry Taylor (CSIRO, AU)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Kerry Taylor

As the scientific community is beginning to take workflow provenance seriously, we are beginning to build substantial collections of provenance traces, often at a fine level of detail required for repeatable execution. We can anticipate the outcomes of the W3C Provenance working group will enable cross-domain querying of provenance repositories. I ask whether we can exploit the properties of semantic representations of provenance to enable the expression of similarity and difference amongst provenance traces at multiple levels of abstraction, identifying the level of abstraction that conveniently highlights similarity. I sketch an approach we are developing to this problem.

### 3.13 Database Wiki and provenance for SPARQL updates

James Cheney (University of Edinburgh, GB)

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© James Cheney

**Joint work of** Cheney, James; Halpin, Harry; Buneman, Peter; Lindley, Sam; Mueller, Heiko

**Main reference** P. Buneman, J. Cheney, S. Lindley, H. Mueller, “The database Wiki project: a general-purpose platform for data curation and collaboration,” SIGMOD Rec. 40, 3 (November 2011), 15–20, 2011.

**URL** <http://dx.doi.org/10.1145/2070736.2070740>

My presentation briefly covered two topics relevant to the seminar: curated Web-based databases and provenance for SPARQL updates. Condensed abstracts of relevant publications [1, 2, 3] follow:

Databases and wikis have complementary strengths and weaknesses for use in collaborative data management and data curation. Relational databases, for example, offer advantages such as scalability, query optimization and concurrency control, but are not easy to use and lack other features needed for collaboration. Wikis have proved enormously successful as a means to collaborate because they are easy to use, encourage sharing, and provide built-in support for archiving, history-tracking and annotation. However, wikis lack support for structured data, efficiently querying data at scale, and localized provenance and annotation. To achieve the best of both worlds, we are developing a general-purpose platform for collaborative data management, called DBWIKI. Our system not only facilitates the collaborative creation of structured data; it also provides features not usually provided by database technology such as annotation, citability, versioning, and provenance tracking. This paper describes the technical details behind DBWIKI that make it easy to create, correct, discuss, and query structured data, placing more power in the hands of users while managing tedious details of data curation automatically.


The (Semantic) Web currently does not have an official or de facto standard way exhibit provenance information. While some provenance models and annotation techniques originally developed with databases or workflows in mind transfer readily to RDF, RDFS and SPARQL, these techniques do not readily adapt to describing changes in dynamic RDF datasets over time. Named graphs have been introduced to RDF motivated as a way of grouping triples in order to facilitate annotation, provenance and other descriptive metadata. Although their semantics is not yet officially part of RDF, there appears to be a consensus based on their syntax and semantics in SPARQL queries. Meanwhile, updates are being introduced as part of the next version of SPARQL. We explore how to adapt the dynamic copy-paste provenance model of Buneman et al. to RDF datasets that change over time in response to SPARQL updates, how to represent the resulting provenance records themselves as RDF using named graphs, and how the provenance information can be provided as a SPARQL end-point.

#### References

- 1 Peter Buneman, James Cheney, Sam Lindley and Heiko Müller *DBWiki: a structured wiki for curated data and collaborative data management*. SIGMOD Conference 2011: 1335–1338
- 2 Peter Buneman, James Cheney, Sam Lindley and Heiko Müller: *The Database Wiki project: a general-purpose platform for data curation and collaboration*. SIGMOD Record 40(3): 15–20 (2011)
- 3 Harry Halpin and James Cheney. *Dynamic Provenance for SPARQL Updates Using Named Graphs*. Workshop on Theory and Practice of Provenance (TaPP 2011). USENIX, 2011. [http://static.usenix.org/events/tapp11/tech/final\\_files/Halpin.pdf](http://static.usenix.org/events/tapp11/tech/final_files/Halpin.pdf)

### 3.14 Dynamicity: Sensor Networks, Streaming Data, and Designing for Change

*Kerry Taylor (CSIRO, AU)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Kerry Taylor

**Joint work of** Lefort, Laurent; Henson, Cory; Taylor, Kerry (eds)

**Main reference** L. Lefort, C. Henson, K. Taylor, P. Barnaghi, M. Compton, O. Corcho, R. Garcia-Castro, J. Graybeal, A. Herzog, K. Janowicz, H. Neuhaus, A. Nikolov, A., K. Page, “Semantic Sensor Network XG Final Report,” W3C Incubator Group Report. June 2011.

**URL** <http://www.w3.org/2005/Incubator/ssn/XGR-ssn/>

In this tutorial introduction to the state of the art and the challenges that dynamicity creates for the semantic web, I summarise progress and ask whether there are opportunities that are being missed. The web is becoming increasingly dynamic, with interactive services and near-real time data services.

I present general techniques and tools that have arisen from the database community and also the event-driven architecture community and introduce the semantics-based achievements that have, roughly speaking, mirrored or built on those pre-semantics approaches. This includes both native RDF stream reasoners, and several wrappers over pre-semantic tools that rewrite queries to enrich basic time-aware services with semantics. There are several proposals for SPARQL extensions for this purpose. But are we missing opportunities to exploit the semantics of dynamics that have not been explored because of the mirroring approach?

I review the work of the recent SSN-XG [1], an incubator group of the W3C that developed an OWL ontology for describing sensor networks and their observations.


I find that although the needs for semantics in a dynamic Web of Data are being addressed, there is very little work in the in the developing Internet of Things, where pushing the semantics down to small and mobile devices, and dealing with actuation in addition to sensing deserves attention.

#### References

- 1 Lefort, L., Henson, C., Taylor, K., Barnaghi, P., Compton, M., Corcho, O., Garcia-Castro, R., Graybeal, J., Herzog, A., Janowicz, K., Neuhaus, H., Nikolov, A., and Page, K., *Semantic Sensor Network XG Final Report*, W3C Incubator Group Report. Available at <http://www.w3.org/2005/Incubator/ssn/XGR-ssn/>, June 2011.

### 3.15 Linking the real world

*Manfred Hauswirth (National University of Ireland – Galway, IE)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Manfred Hauswirth

**Joint work of** Hauswirth, Manfred; Le-Phuoc, Danh; Nguyen-Mau, Hoan Quoc; Xavier Parreira; Josiane

**Main reference** M. Hauswirth, D. Le-Phuoc, H.Q. Nguyen-Mau, J. Xavier Parreira, “A middleware framework for scalable management of linked streams,” *Web Semantics: Science, Services and Agents on the World Wide Web*, 2012, in press.

**URL** <http://dx.doi.org/10.1016/j.websem.2012.06.003>





Until now the virtual world of information sources on the World Wide Web and activities in the real world have always been separated. However, knowledge accessible on the Web (the virtual world) may influence activities in the real world and vice versa, but these influences are usually indirect and not immediate. We still lack means to interconnect and link this



information in a meaningful, general-purpose, and simple way. Additionally, information comes in the form of streams which complicates the data management from the IoT up to backend information systems. In this talk, I will present ongoing work and open research problems in this domain which aims at enabling novel ways for humans to interact with their environment and facilitating interactions among entities of the physical world (with a specific focus on data management problems).

### 3.16 Semantic data streams: does it make sense?

*Oscar Corcho (Universidad Politécnica de Madrid, ES)*

**License**     Creative Commons BY-NC-ND 3.0 Unported license  
© Oscar Corcho

**Joint work of** Corcho, O.; Calbimonte, J.P.

**Main reference** J.P. Calbimonte, H. Jeung, O. Corcho, K. Aberer, “Enabling Query Technologies for the Semantic Sensor Web,” *International Journal of Semantic Web and Information Systems*, Volume 8, Issue 1, 2012.

**URL** <http://dx.doi.org/10.4018/jswis.2012010103>

Several works have shown in the past that it is possible to generate and publish data streams using semantic technologies. These data streams may be generated from environmental sensors, traffic sensors, humans, etc. While technology and infrastructure are being made available, as will be described in this talk, there is a need to reflect on which are the main challenges and limitations, not only from a technological point of view but also from a wider perspective. This talk will also reflect on such limitations and challenges.

### 3.17 Probabilistic Processing of Sliding Window Top-k Queries

*Ivana Podnar Zarko (University of Zagreb, HR)*

**License**     Creative Commons BY-NC-ND 3.0 Unported license  
© Ivana Podnar Zarko

A sliding window top-k (top-k/w) query monitors incoming data stream objects within a sliding window of size  $w$  to identify the  $k$  best-ranked objects with respect to a given scoring function over time. Processing of such queries is challenging because, even when an object is not a top-k/w object at the time when it enters the processing system, it might become one in the future and thus a set of potential top-k/w objects has to be stored in memory.


The talk presents the first probabilistic approach to top-k/w processing in the literature which offers significantly improved runtime performance compared to existing deterministic algorithms, especially for large values of  $k$ , at the expense of a small and controllable probability of error. The algorithm is designed for processing random order data streams.

#### References

- 1 Kresimir Pripuzic, Ivana Podnar Zarko, and Karl Aberer. *Distributed processing of continuous sliding-window k-NN queries for data stream filtering*. World Wide Web 14 (5-6); pp. 465-494, 2011
- 2 Kresimir Pripuzic, Ivana Podnar Zarko, and Karl Aberer. *Top-k/w publish/subscribe: A publish/subscribe model for continuous top-k processing over data streams*. Information Systems, 2012

### 3.18 Adaptive Data Management Techniques for Federations of Endpoints

Maria-Esther Vidal (*Universidad Simón Bolívar, VE*)

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Maria-Esther Vidal

Joint work of Vidal, Maria-Esther; Montoya, Gabriela; Acosta, Maribel

URL <http://code.google.com/p/defender-portal/>


Emerging technologies that support networks of sensors or mobile smartphones are making available an extremely large volume of data or Big Data; additionally, in the context of the Cloud of Linked Data, a large number of huge RDF linked datasets have become available, and this number keeps growing. Simultaneously, although scalable and efficient RDF engines that follow the traditional optimize-then-execute paradigm have been developed to locally access RDF data, SPARQL endpoints have been implemented for remote query processing. Given the size of existing datasets, lack of statistics to describe available sources, and unpredictable conditions of remote queries, existing solutions are still insufficient. First, the most efficient RDF engines rely their query processing algorithms on physical access and storage structures that are locally stored; however, because of the size of existing linked datasets, loading the data and their links is not always feasible. Second, remote linked data query processing can be extremely costly because of the lack of query planning; also, current techniques are not adaptable to unpredictable data transfers or data availability, thus, executions can be unsuccessful. To overcome these limitations, query physical operators and execution engines need to be able to access remote data and adapt query execution schedulers to data availability. In this talk we present the basis of adaptive query processing frameworks defined in the database area, and their applicability in the Linked and Big Data context where data can be accessed through SPARQL endpoints. In this talk we limitations of existing RDF engines, adaptive query processing techniques, and how traditional RDF data management approaches can be well-suitable to runtime conditions, and extended to access a large volume of data distributed in federations of SPARQL endpoints. We compare adaptive query processing techniques implemented by ANAPSID [1] and compare with respect to the ones provided by existing federated engines. e.g., FedX [2]; the FedBench benchmark [3] was used during the experiments. We show that in some cases execution time can be sped up by up to three orders of magnitude; results of the study can be found at <http://code.google.com/p/defender-portal/>.

#### References

- 1 Maribel Acosta and Maria-Esther Vidal and Tomas Lampo and Julio Castillo and Edna Ruckhaus *ANAPSID: an adaptive query processing engine for SPARQL endpoints*. Proceedings of the 10th international conference on The semantic web – Volume Part I, 2011.
- 2 Andreas Schwarte and Peter Haase and Katja Hose and Ralf Schenkel and Michael Schmidt *FedX: Optimization Techniques for Federated Query Processing on Linked Data*. Proceedings of the 10th international conference on The semantic web – Volume Part I, 2011.
- 3 Michael Schmidt and Olaf Gorlitz and Peter Haase and Gunter Ladwig and Andreas Schwarte and Thanh Tran. *FedBench: A Benchmark Suite for Federated Semantic Data Query Processing*. Proceedings of the 10th international conference on The semantic web – Volume Part I, 2011.

### 3.19 Conceiving the Web of Linked Data as a Database

*Olaf Hartig (HU Berlin, DE)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Olaf Hartig

**Main reference** O. Hartig, J.-C. Freytag, “Foundations of Traversal Based Query Execution over Linked Data,” in Proc. of the 23rd ACM Conference on Hypertext and Social Media; Milwaukee, WI, USA, pp. 43–52, 2012.

**URL** <http://doi.acm.org/10.1145/2309996.2310005>

The World Wide Web (WWW) currently evolves into a Web of Linked Data where content providers publish and link their data as they have done so for Web documents since 20 years. While the execution of SQL-like queries over this emerging dataspace opens possibilities not conceivable before, querying the Web of Linked Data poses novel challenges. Due to the openness of the WWW, it is impossible to know all data sources that might contribute to the answer of a query. To tap the full potential of the Web, traditional query execution paradigms are insufficient because those assume a fixed set of potentially relevant data sources beforehand. In the Web context these data sources might not be known before executing the query.

In this talk we discuss how the Web of Linked Data –conceived as a database– differs from traditional database scenarios. Furthermore, we introduce a novel query execution paradigm that allows the execution engine to discover potentially relevant data during the query execution. The main idea of this paradigm is to traverse data links during query execution to discover data and data sources that may contribute to the query result.

### 3.20 Scientific Data Management – From the Lab to the Web

*José Manuel Gómez-Pérez (iSOCO, ES)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© José Manuel Gómez-Pérez

The digital universe is booming, especially metadata and user-generated data. This raises strong challenges in order to identify the relevant portions of data which are relevant for a particular problem and to deal with the lifecycle of data. Finer grain problems include data evolution and the potential impact of change in the applications relying on the data, causing decay. The management of scientific data is especially sensitive to this. We present the Research Objects concept as the means to identify and structure relevant data in scientific domains, addressing data as first-class citizens. We also identify and formally represent the main reasons for decay in this domain and propose methods and tools for their diagnosis and repair, based on provenance information. Finally, we discuss on the application of these concepts to the broader domain of the Web of Data: Data with a Purpose.

### 3.21 Semantic Search Tutorial

*Duc Thanh Tran (KIT, DE)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Duc Thanh Tran

This tutorial on semantic search reviewed the most relevant aspects of this area.

### 3.22 Probabilistic Models and Reasoning for Link Consistency Checking and Validation

*Edna Ruckhaus (Universidad Simón Bolívar, VE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Edna Ruckhaus

Joint work of Ruckhaus E; Vidal ME

Currently, links among datasets in the LOD cloud are neither complete nor correct and links may be missing or inconsistent. Such is the case of the life sciences domain where datasets on diseases, drugs and clinical trials have been published in the LOD cloud. In this work, we combine Bayesian Networks and Probabilistic Reasoning in order to study and correct the incompleteness and inconsistencies of links in the LOD cloud. Bayesian networks are suitable for studying the probability of occurrence of a certain link among datasets or value within a single dataset, while probabilistic reasoning about similarity is used. We propose a two-fold approach: (1) a Conflict Detector that discovers structural and semantic linked data conflicts through queries to an RDF Bayesian network, and (2) a Linked Data Cleanser that uses Probabilistic Similarity Logic (PSL) in order to produce information on similar items that allows the solution of these conflicts.

### 3.23 Ranking SPARQL Results

*Ralf Schenkel (Universitat des Saarlandes, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Ralf Schenkel

The amount of semantic data available for applications is increasing rapidly. Querying this large amount of data, and especially finding the most important or most useful results to a query, is a difficult and important problem. While specific ranking methods are available for specific applications, there is a clear need for a generic ranking method, similar to document ranking in Web search.

This talk presents two such ranking methods. First, it proposes ranking query results based on the frequency of the corresponding facts in a large corpus, such as the web. Using standard ranking methods from IR, this allows to rank important facts at the top of the results. It is also easily possible to extend the method to include keyword conditions or query relaxation.

Second, the talk presents a method to rank documents that contain a fact and therefore can serve as witness for the correctness of the fact. It combines document authority, confidence of the extraction patterns, and fact coverage to a document score.

### 3.24 The Search for Crystals

Kerry Taylor (CSIRO, AU)

**License** © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license  
© Kerry Taylor

**Joint work of** Newman, J.; Bolton, E. E.; MÄller-Dieckmann, J.; Fazio, V. J.; Gallagher, D. T.; Lovell, D.; Luft, J. R.; Peat, T.S. ; Ratcliffe, D.; Sayle, R.A.; Snell, E. H.; Taylor, K.; Vallotton, P.; Velanker S.;von Delft, F.

**Main reference** J. Newman, E. E. Bolton, J. Müller-Dieckmann, V. J. Fazio, D. T. Gallagher, D. Lovell, J. R. Luft, T. S. Peat, D. Ratcliffe, R. A. Sayle, E. H. Snell, K. Taylor, P. Vallotton, S. Velanker, F. von Delft, "On the need for an international effort to capture, share and use crystallization screening data," Acta Crystallographica Section F: Structural Biology and Crystallization Communications, 68:3, March 2012.

**URL** <http://dx.doi.org/10.1107/S1744309112002618>

With the rapid adoption of formal ontologies for modelling, especially in the life sciences, there is a new opportunity for data mining over instance data that leverages the ontology to provide meaningful features for learning that are not directly present in the instance data.

In this work we tackle the problem of the paucity of expert knowledge in the domain of protein crystallisation, a bottleneck step in the process of life sciences research by which protein function is determined. In practice, high throughput chemical laboratory machines are used to sample the space of experimental conditions in a batch. After weeks or months, sometimes a crystal, or something that is considered to be close to a crystal, is formed and this outcome is used to refine the choice of conditions for the next batch of experiments. An expert group has been formed to collaboratively develop an ontology to capture the domain knowledge and to provide a basis for sharing experimental data amongst the laboratories. See [1] for a more developed problem description.


Building on the tradition of ILP for learning in logical formalisms, we are developing an approach for searching for patterns in experimental data supported by background knowledge represented as an ontology, currently EL+ with concrete domains. We have developed a large database of rdf instance data, and are experimenting with techniques to search the space of descriptions of clusters of experiments with similar outcomes. We find that coverage checking under the OWA and over a lot of data is particularly difficult in our work and are developing ways to improve this step.

#### References

- 1 J. Newman, E. E. Bolton, J. MÄller-Dieckmann and V. J. Fazio, D. T. Gallagher, D. Lovell, J. R. Luft T. S. Peat, D. Ratcliffe, R. A. Sayle, E. H. Snell, K. Taylor, P. Vallotton, S. Velanker, F. von Delft, *On the need for an international effort to capture, share and use crystallization screening data*. Acta Crystallographica Section F: Structural Biology and Crystallization Communications, 68:3, March 2012

### 3.25 Paths in semantic search: A back and forth story

*Marcelo Arenas (Universidad Catolica de Chile, CL)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Marcelo Arenas


**Joint work of** Arenas, Marcelo; Conca, Sebastian; Perez, Jorge

**Main reference** Marcelo Arenas, Sebastian Conca and Jorge Perez, “Counting beyond aYottabyte, or how SPARQL 1.1 Property Paths will prevent adoption of the standard,” in Proc. of the 21st Int’l Conf. on World Wide Web (WWW 2012), Lyon, France, pp. 629–638, 2012.

SPARQL -the standard query language for querying RDF- provides only limited navigational functionalities, although these features are of fundamental importance for graph data formats such as RDF. This has led the W3C to include the property path feature in the upcoming version of the standard, SPARQL 1.1. In this talk, we review the efforts of the W3C to define the semantics of property paths in SPARQL 1.1. In particular, we show that the initial proposal for this semantics, which was based on counting, is infeasible in practice, and we discuss some drawbacks of the current proposal that still includes a mild form of counting. We also argue in this talk that there are interesting problems to solve in this area, like standardizing some functionalities for returning paths in the answer to a query.

### 3.26 Getting Semantics from the Crowd

*Gianluca Demartini (University of Fribourg, CH)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Gianluca Demartini

**Joint work of** Demartini, Gianluca; Difallah, Djellel Eddine; Cudrè-Mauroux, Philippe

**Main reference** G. Demartini, D.E. Difallah, P. Cudrè-Mauroux, “ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking,” in Proc. of 21st Int’l Conf. on World Wide Web (WWW 2012): 469–478.


**URL** <http://doi.acm.org/10.1145/2187836.2187900>

Semantics data needs social computing techniques to obtain high quality and scalability at the same time. Specifically, in this talk we claim that crowdsourcing can be exploited to improve several semantic web tasks. As an example of social computing application to semantic web, we present ZenCrowd: a system that exploits probabilistic techniques and crowdsourcing for large-scale entity linking. It takes advantage of human intelligence to improve the quality of the links by dynamically generating micro-tasks on an online crowdsourcing platform.

Our system, ZenCrowd, identifies entities from natural language text using state of the art techniques and automatically connects them to the Linked Open Data cloud. We show how one can take advantage of human intelligence to improve the quality of the links by dynamically generating micro-tasks on an online crowdsourcing platform. We developed a probabilistic framework to make sensible decisions about candidate links and to identify unreliable human workers. We evaluated ZenCrowd in a real deployment and show how a combination of both probabilistic reasoning and crowdsourcing techniques can significantly improve the quality of the links, while limiting the amount of work performed by the crowd.

### 3.27 Efficient Crowdsourcing for Metadata Generation

*Wolf-Tilo Balke (TU Braunschweig, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Wolf-Tilo Balke

Rich and correct metadata still plays a central role in accessing data sources in a semantic fashion. However, at the time of content creation it is often virtually impossible to foresee all possible uses of content and to provide all interesting index terms or categorizations. Therefore semantic retrieval techniques have to provide ways of allowing access to data via missing metadata, which is only created when needed, i.e. at query time. Since the creation of most such metadata will to some degree depend on human judgement (either how to create it in a meaningful way or by actually providing it), crowdsourcing techniques have recently raised attention.

By incorporating human workers into the query execution process crowd-enabled databases already can facilitate intelligent, social capabilities like completing missing data at query time or performing cognitive operators. Typical examples are ranking tasks, evaluating the correctness of automatically extracted information, or judging the similarity or subjective appeal of images. But for really creating metadata for probably large data sources, the number of crowd-sourced mini-tasks to fill in missing metadata values may often be prohibitively large and the resulting data quality is doubtful. Instead of simple crowd-sourcing to obtain all values individually, in this talk utilizing user-generated data found in the Social Web is discussed

By exploiting user ratings semantically meaningful perceptual spaces can be built, i.e. highly-compressed representations of opinions, impressions, and perceptions of large numbers of users. Then, using few training samples obtained by expert crowd sourcing, missing metadata can be extracted automatically from the perceptual space with high quality and at low costs. First experiments show that this approach actually can boost both performance and quality of crowd-enabled databases, while also providing the flexibility to expand schemas in a query-driven fashion.

## 4 Working Groups

Working groups were created on the first two days of the workshop in order to allow discussions on topics related to the first two main areas identified for the workshop: scalability and provenance. The decision on the working groups to be run was done collaboratively among the workshop participants, who proposed, during the morning, topics that they were interested in discussing, and then these topics were clustered and reorganised by the day leader so as to generate the working groups. After the working group meeting, a reporting session was held where the main outcomes of the working group were reported.

In the following sections we provide a brief summary of the main outcomes reported on these sessions.

### Benchmarking

*Reported by Oscar Corcho.* This was not the only working group dealing with this topic (in fact, it was dealt with for every theme), what shows the importance of benchmarking for a community that is getting sufficiently mature. It was agreed in general that existing

benchmarks have served the semantic data management community in its initial stages, but that there is a need to improve them adding more types of queries, more realistic datasets, more possibilities of configuring them, adequate infrastructure and processes, etc.

The group started identifying a list of existing benchmarks (the list is not intending to be exhaustive, but aims at showing the heterogeneity and variety of benchmarks available in the state of the art – for a more up-to-date list the following URL can be consulted: <http://www.w3.org/wiki/RdfStoreBenchmarking>). The following list presents some benchmarks focused on SPARQL, in increasing order of complexity, taking into account the size of the query plans generated from them:

- Leigh University Benchmark (LUBM), which would be the equivalent to TPC-C in relational databases.
- Berlin SPARQL Benchmark (BSBM), and all of its variations: BSMBi, BSBM Update, BSBM-IR.
- *SP<sup>2</sup>Bench*
- Social Network Intelligence Benchmark (SIB)

Other benchmarks that were mentioned as well were DBpedia benchmark (which includes extracts of query loads and operates over a real dataset), FedBench for federated SPARQL queries, a streaming data benchmark being developed jointly by UPM and CWI, and some RDB2RDF and geospatial benchmarks that were also being developed.

All of these benchmarks present some limitations. Some of those were identified as follows:

- Some constructs are not too well-represented in the benchmarks, either on the query side or on the data side, or in both. For example, named graphs, aggregates (although BSMBi/BIBM has them, extending BSBM), sameAs links, rdfs:label, or SPARQL Update constructs
- Query logs are not very useful for characterising typical usage. For instance, it may be the case that a SPARQL endpoint is never contacted since all data has been downloaded and used locally in an application; in a federated query, a SPARQL endpoint only receives part of the complete query; query logs will normally contain those queries that can be done (while normally one should put higher the barrier of what can currently be done).
- There are some difficulties in replicating experiments, especially in the case of federated query processing and Linked Data processing. For instance, data and the underlying schemas of that data change frequently, what makes these queries break easily, and it is difficult to know the characteristics of the underlying data sources / capabilities of the endpoints / latency
- Metrics like cost per metric-throughput or Watts per transaction has not been applied yet in RDF benchmarking

In summary, there was an agreement that the current benchmarks are mainly for the semantic data management community and not for outsiders, and they show RDF very light when compared to relational databases. In this context, there was a mention to the upcoming Linked Data Benchmarking Council project<sup>1</sup> that would start by the end of 2012 and would focus on many of these limitations. This was also related to some of the topics for future work that were identified as next steps:

- Create better RDF benchmarks, with a special focus on convincing also outsiders. For example, combining Business Intelligence analytics with strong data integration plus some transitivity/reasoning, among others.

---

<sup>1</sup> <http://ldbc.sti2.at/>



- Generate a map of current SPARQL benchmarks according to a set of dimensions
- Identify benchmarks by typical usages: Linked Data browsers use typical exploratory queries (get concepts, get properties); Faceted browsers; In graphs, find the most connected person in the social graph; Detect properties that are very scarcely used (probably mistypings), Loading data (e.g., User Generated Content). It is not a multi-step transaction process
- Define clearly the measurement protocols, processes and rules (caches, updates, etc.)

### Provenance and scalability

*Reported by Frank van Harmelen.* This group discussed about whether there were major scalability challenges related to provenance storage and exploitation taking into account the current technological infrastructure. In this context, provenance was roughly classified in two types: coarse-grained (that is, provenance at the graph level), and fine-grained (that is, provenance focused on individual triples). The former tends to be rather small while the latter may get really large and applicable to problems like access control to individual triples.

The group included some representatives from companies that were already working on realistic scenarios on the representation and use of provenance in both types of cases, and in general the working group considered that there were no major constraints or challenges in terms of scalability to handle this type of metadata, except for the case of reification.

### Provenance-specific benchmarks and corpora

*Reported by Paolo Missier.* This group discussed about the need of having provenance-specific benchmarks, since these were not well covered in the current state of the art of semantic data management benchmarking. Such benchmark should include a good corpus of provenance data and queries, and would be useful to test the performance of query processors of provenance, including, for instance, pattern detection problems in provenance traces.

The group concluded that to provide support to such benchmarking activity, there is a need to build some infrastructure to allow benchmarking, including allowing people to submit example, there is also a need to classify provenance queries, and there is a need to focus on interoperability between different provenance models.

### Novel usages of provenance information

*Reported by José Manuel Gómez-Pérez.* This group identified a set of scenarios where provenance information could be useful, and reported them as potential scenarios that should be considered to showcase the usefulness and need for provenance information recording and exploitation, as well as to drive some additional applied research on this area. This included.

- Data integration (assisted analysis and allow exploration along different dimensions of quality, in the context of SmartCities and OpenStreetMap)
- Analytics in social networks (detect cool members in social networks)
- How to extract differences between provenance logs
- Billing purposes / Privacy aspects (you bill for dataset usage and have to bill people for the answers provided to queries)
- Credit, attribution, citation and licensing
- Result reproducibility (e.g., Executable Paper Challenge)
- Determining quality in the report that has been generated by 3rd parties for an organisation (e.g., Government report)

## Provenance and uncertainty

*Reported by Norbert Fuhr.* Sources may vary in their reliability. This group listed a few use cases where uncertainty plays an important role:

- Bias in judgment or prejudice: A popular example is the question "Is Barack Obama a Muslim": When going to the Web, there is a substantial fraction of pages answering this question with yes. Customer reviews of products may be biased in one direction or the other
- Sensors may suffer from malfunction or degrade over time. In "Smart Cities", sensor inputs and tweets from people are analysed for detecting important events, like e.g. a traffic jam or an accident.
- Large knowledge bases (like e.g. Freebase) are built from different sources. If some of them are based on information extraction, it is important to consider the imperfect precision of these methods.
- Also, facts in a knowledge base might be modified maliciously (like e.g. in Wikis), thus identification of contaminated data is crucial.

So the major causes for uncertainty in provenance are human error, bias or fraud, sensor errors and the limited precision of information extraction methods.

Before applying further methods, it is essential that the authenticity of provenance can be validated. One interesting method for tracking provenance is the "information cascades" approach. When considering the uncertainty due to provenance, the correlation of statements is necessary (e.g., a sensor typically produces a series of faulty readings, so assuming the independence of the probabilities of correctness is inappropriate. For further reasoning, the uncertainty values can either be propagated, or a cutoff can be applied at some point, after which a statement is considered either true or false with certainty. Hard integrity constraints (in contrast to soft rules) are also helpful for eliminating potentially incorrect statements. However, exceptions to these rules might occur sometimes.

For future research, the following challenges were identified:

- There is the need for a data model for uncertainty and provenance. Following the W3C model for provenance, it seems reasonable to assign the uncertainty to the agent, so that all entities produced by the agent (e.g. a sensor) are reliable with the same probability.
- We also need a (probabilistic) dependence model. Markov logic might be a candidate here.
- When inferences from uncertain statements are drawn, the system should be able to explain the derivation of the uncertainty values of the results. For that, a summarization of the lineage trees involved in deriving the result should be presented: Instead of a detailed presentation, most users might prefer to see only the most important factors influencing the result. Thus, appropriate methods for generating such explanations are to be developed.
- Methods for quantifying the uncertainty/reliability of a source are required. Typically, these methods would be based on the further inference process, like comparing the output of nearby sensors, or detecting overwhelming evidence for contrary of a statement.

## Participants

- Karl Aberer  
EPFL – Lausanne, CH
- Grigoris Antoniou  
University of Huddersfield, GB
- Marcelo Arenas  
Univ. Catolica de Chile, CL
- Wolf-Tilo Balke  
TU Braunschweig, DE
- James Cheney  
University of Edinburgh, GB
- Oscar Corcho  
Univ. Politec. de Madrid, ES
- Philippe Cudré-Mauroux  
University of Fribourg, CH
- Gianluca Demartini  
University of Fribourg, CH
- Orri Erling  
Openlink Software, NL
- Dieter Fensel  
Universität Innsbruck, AT
- Norbert Fuhr  
Universität Duisburg-Essen, DE
- Avigdor Gal  
Technion – Haifa, IL
- José Manuel Gomez-Perez  
ISOCO – Madrid, ES
- Alasdair J G Gray  
University of Manchester, GB
- Marko Grobelnik  
Jozef Stefan Inst. – Ljubljana, SI
- Paul Groth  
VU – Amsterdam, NL
- Andrey Gubichev  
TU München, DE
- Peter Haase  
fluid Operations AG –  
Walldorf, DE
- Stephen Harris  
Garlik Ltd. – London, GB
- Olaf Hartig  
HU Berlin, DE
- Manfred Hauswirth  
National University of Ireland –  
Galway, IE
- Jeff Hefflin  
Lehigh Univ. – Bethlehem, US
- Spyros Kotoulas  
IBM Research – Dublin, IE
- Paolo Missier  
Newcastle University, GB
- Luc Moreau  
University of Southampton, GB
- Charalampos Nikolaou  
National and Kapodistrian  
University of Athens, GR
- Ivana Podnar Zarko  
University of Zagreb, HR
- Edna Ruckhaus  
Univ. S. Bolivar – Caracas, VE
- Satya S. Sahoo  
Case Western Reserve Univ., US
- Manuel Salvadores  
Stanford University, US
- Ralf Schenkel  
Universität des Saarlandes, DE
- Juan F. Sequeda  
University of Texas at Austin, US
- Wolf Siberski  
Leibniz Univ. Hannover, DE
- Elena Simperl  
KIT – Karlsruhe Institute of  
Technology, DE
- Kavitha Srinivas  
IBM TJ Watson Research Center  
– Hawthorne, US
- Rudi Studer  
KIT – Karlsruhe Institute of  
Technology, DE
- Kerry Taylor  
CSIRO, AU
- Martin Theobald  
MPI für Informatik –  
Saarbrücken, DE
- Bryan Thompson  
SYSTAP – Greensboro, US
- Duc Thanh Tran  
KIT – Karlsruhe Institute of  
Technology, DE
- Frank van Harmelen  
VU – Amsterdam, NL
- Maria-Esther Vidal  
Univ. S. Bolivar – Caracas, VE
- Valentin Zacharias  
FZI Karlsruhe, DE

