Dagstuhl Perspectives Workshop Report: Semantic Web Reflections and Future Directions

Organisers: John Domingue (Open University, United Kingdom), Dieter Fensel (University of Innsbruck, Austria), Jim Hendler (Rensselaer Polytechnic Institute, United States), Rudi Studer (Karlsruhe Institute of Technology, Germany)

Editors: Andreas Harth, Rudi Studer (Karlsruhe Institute of Technology, Germany)

March 29, 2010

Abstract

With an ever increasing amount of data being stored and processed on computers, and the ubiquitous use of the Web for communication and dissemination of content, the world contains a vast amount of digital data that is growing ever faster. The available data is increasingly used to gain insights for science and research, to create commercial value, and to hold governments accountable. Semantic Web technologies for supporting machine-readable web content aim at facilitating the processing and integration of data from the open web environment where large portions of the publicly available data is being published. Since the first Dagstuhl seminar "Semantics on the Web" in 2000 the amount of machine-readable data on the web has exploded, and Semantic Web technologies have matured and made their way from research labs and universities into commercial applications. This report identifies lessons learned and future directions for the field as discussed at a Perspectives Workshop on Semantic Web, which took place in Dagstuhl, Germany, in June/July 2009.

Dagstuhl Seminar Proceedings 09271
Perspectives Workshop: Semantic Web Reflections and Future Directions
http://drops.dagstuhl.de/opus/volltexte/2010/2533

Contents

1 Introduction

In March 2000 Dagstuhl hosted the "Semantics for the Web" seminar [6, 7] which brought together a variety researchers interested in creating technologies for supporting machine-readable web content sustaining autonomous interactive agents. The seminar was a huge success with significant impact. At the time of the first workshop, the term "Semantic Web" was not yet well-known, and the Dagstuhl meeting is widely acknowledged as the first major meeting for the emerging field. Now, there is a thriving research community including a number of conferences, three active journals, a wide-range of funding activities, and a number of research laboratories at major universities and companies.

The objectives for the workshop on semantic technologies in 2009 was to reflect on the original vision for the Semantic Web and to create a roadmap for future research, taking into account new web technologies that have emerged in the last few years as well as the needs of the business community. Accordingly, the seminar brought together distinct groups such as:

- the original inventors and drivers of the Semantic Web,
- the leaders of emerging web technologies such as Web 2.0, web services, Service Oriented Architectures (SOA), and the Grid,
- industrialists who have an interest in using Semantic Web technology,
- decision makers in public funding activities.

In this document we report on the group discussions taking place at the workshop as follows: first, we discuss issues and future trends in the area of vocabularies on the Semantic Web, ie. how data can be modelled and marked up for online publication, integration, and reasoning. Second, given the scale of the web, we discuss issues pertaining to scalability. Third, as the topic of provenance of data emerges in environments where data is being integrated from large numbers of sources, we provide an overview of the field and discuss possible approaches for provenance on the Semantic Web. Fourth, we discuss the concept of infospheres, coveringing issues such as privacy and souvereignity over one's personal data. Fifth, we consider the question of how users can interact with huge amounts of automatically integrated Semantic Web data. Finally, we discuss e-science and mobile and real-world networks as possible applications of the foundational Semantic Web technologies.

2 Vocabularies and the Semantic Web

Tom Heath, Talis, United Kingdom

2.1 Introduction

This discussion group was formed to explore, understand, and begin to address a range of issues related to vocabularies/ontologies and the Semantic Web. In this context we define vocabularies as lightweight ontologies oriented as much toward publishing of data on the Web for integration and reuse as for reasoning through description logics. The role of vocabularies is critical to the success of the Semantic Web from both a research and a deployment perspective. Without resolution of a number of issues related to deployment and usage of vocabularies, the continued growth of the Semantic Web is questionable. If this growth does continue it will inevitably bring further to the fore a number of ongoing research challenges. The report that follows summarises the progress made in the discussion group.

2.2 Vocabulary Development

The group discussion made it clear that vocabulary development was still seen very much as a niche activity undertaken by those with some degree of prior specialist experience. In order to ease the vocabulary development bottleneck it was agreed that further work was required in producing best practice guidance for vocabulary developers, incorporating increased material on design patterns and more in-depth guidance about how to choose a modelling formalism, and the implications of this choice. The first of these goals may be achieved through extension of the work carried out by the ontologydesignpatterns.org activity, which has since sponsored a VoCamp in order to help move this issue forward.

2.3 Vocabulary Management

Assuming the vocabulary development bottleneck can be eased, there is a pressing need to provide more stable, robust and persistent hosting for vocabularies. For example, if novel business ventures are to be created that make extensive use of certain vocabularies, those engaged in these ventures will need to understand and limit the risks associated with a vocabulary disappearing from the Web or being subjected to domain hijacking, for example. One mitigating strategy for these scenarios is to introduce more robust hosting arrangements for vocabularies, including greater redundancy through e.g. mirroring. Integral to this issue is the question of persistent identifiers for vocabularies. Current namespaces such as purl.org and sharednames.org are popular, however many others are in use that may not have long-term support arrangements in place.

Tightly coupled to the issue of hosting, and persistence through archiving, is the question of versioning of vocabularies. As a vocabulary is deployed in the wild it may adopt novel forms of usage not originally anticipated by the authors, but that may ultimately lead to a change in the vocabulary specification itself. Clearer guidance is needed about how to address these versioning issues. For example, the W3C approach is to create a double set of URIs for vocabularies, one dated and one not. The undated URI always points at the latest version, meaning it is always possible to go to previous versions via undated URIs, but can make it difficult to trace the change reason for one specific term.

Because even minor changes to vocabularies can have significant impacts on deployed applications, an essential addition to more robust hosting arrangements is the implementation of clear governance procedures for publicly available vocabularies. Just as in open source software projects, a vocabulary may be dependent on a large and active community, or on the stewardship of a "benevolent dictator". In either case, the process by which changes are made needs to be clear to potential users of the vocabulary.

Just as the importance of having clear rights statements for data is being increasingly recognised, so must the terms under which a vocabulary can be reused be made clear. At present very few published vocabularies have such statements, although the Creative Commons licenses and RDF vocabulary likely provide a suitable framework. It was concluded that a greater number of published examples of this good practice are required.

2.4 Vocabulary Usage

There was strong agreement in the group that one of the key requirements for driving vocabulary usage is the availability of vocabulary discovery services. While several have been created over the years there is a precedent for these to disappear. Various approaches for building discovery services are appropriate, such as curation or crawling, however the key factor is that these services receive similar support, to ensure their persistence, as vocabularies themselves. One outstanding question related to whether generic or domain-specific services (e.g. for health care) were more appropriate.

The second major point of agreement was that significantly more guidance for users is needed to demonstrate how terms from different vocabularies may be used in combination to publish a particular data set. For example, even relatively simple RDF documents frequently contain terms from a number of different vocabularies. Potential data publishers often need guidance regarding which combinations of terms are suitable and common for describing particular types of entities. These example graph patterns must encompass appropriate schema combinations (i.e. compatible domains and ranges) and also reflect patterns of usage in the wild. Profiling usage in this way may also help in the tracing and documentation of dependencies between vocabularies.

2.5 Vocabulary Research

While the the preceding topics raised many applied research questions, a number of more theoretical questions emerged from the discussions.

For example, a significant effort has already been expended in defining potential quality metrics for ontologies and vocabularies. Not only do these need to be disseminated more widely, further research is also required to understand the scenarios in which particular metrics have validity. For example, in a lightweight data sharing scenario, the depth of expressivity of a vocabulary may be of relatively little interest or consequence to the data publisher.

Two related issues concern vocabulary alignments, and the consequences of alignments. Firstly, it was felt that the semantics of vocabulary alignments was not always sufficiently clear, and this was underpinned by a lack of adequate languages and models for publishing and defining alignments. Similarly there is a lack of an adequate means to express transformation rules between terms in different vocabularies. Secondly, it was felt that further research is required into the potential for and effects of reasoning based on partial knowledge of a vocabulary. Lastly, it was questioned whether there were issues of robustness affecting vocabularies and class hierarchies, or potential security issues in general. For example, is it possible to create a vocabulary that would infer some sort of a malware?

3 Scalability

Frank van Harmelen, Vrije Universiteit Amsterdam, The Netherlands

3.1 Status of the Field

Not much time was spent on analysing the status of the field, as most participants were closely familiar with it, and since the field is moving at such a high pace that any analysis would be rapidly outdated.

3.2 Open Issues

A number of important open issues were identified in the opening presentation:

Does scale matter? Often in engineering, when quantitative scale grows beyond a certain limit, the trade-offs change in a qualitatively different way, necessitating an entirely different approach to the problem, instead of just maintaining the same engineering directions but at larger scale.

Centralisation or distribution? Distribution is often claimed to be a promising road to scalability, and also one that meshes very well with the distributed nature of the Web. But surprisingly, centralisation seems to have worked very well on today's Web: the large search engines all work by locally caching the entire web and running indexes etc on that local, centralised cache. Will distribution be needed for the Semantic Web?

Logic or IR? The world of logic is dictated by the discrete criteria of soundness and completeness: a calculus or an inference engine is either sound and complete, or it isn't. Information retrieval on the other hand works with the measures of recall and precision, which can take on any value between 0 and 1. Until now the Semantic Web has been dominated by the logical approach, but scalability might well be served by adopting the IR approach.

Forward or backward inference? Most available scalable RDF stores implement inference as forward reasoning, deriving all consequences from an RDF/OWL graphs materialising the full closure. Very few scalable tools have explored the more traditional query-driven backward reasoning processes.

Which parameters matter? Every time a new benchmarking study of semantic web stores or reasoners is published, a debate emerges about what the benchmark should have been varying, and which parameters they should have been measuring. In short, there is very little agreement on which parameters should be used to define the problem space of "scalability" on the semantic web.

3.3 Perspectives of How to Address the Issues

A first crucial insight into how to address the issue of scalability is that this issue cannot be considered in isolation. We will have to take into account what types of queries are to be supported, what quality of answers will be required, what kinds of ontologies must be supported, etc.

Of course a basic approach to scalability is to "simply" improve raw performance. A wide variety of strategies was identified to achieve this, including

- satisficing strategies,
- heuristics for caching,
- trading quality, soundness or completeness for time,
- ranking strategies,
- anytime behaviour,
- modularisation and selection,
- knowledge compilation,
- forgetting, and

- statistical analysis of data and query patterns.

There was wide agreement among the participants that some combination of ranking and reasoning would be required if reasoners were to escape from the logical "fully complete" paradigm (and even with that paradigm, ranking of answers would be important). It would also require to recognise that statements which are logically equivalent do not always have to be treated in equivalent ways. It would be possible to do this either in a loose coupling (ranking applied to the results of a classical reasoner) or in a tight coupling (ranking built in to the reasoning algorithm). Sources of such ranking could be such extra-logical sources as the origin of the statements or the strings used for the identifiers occurring in the statements. One would also have to decide which items would be the subject of such rankings: query answers, triples, nodes, molecules, documents, or entire graphs?

A complicating factor in scalability is the potential dynamics of the datasets: the rate at which data is changing compared to the rate at which queries have to be answered, distinguishing between the frequency of data changes and the volumes of data that are involved in the changes. Dynamic data-sets will become stale, hampering reuse of previous results, and leading to incorrectness for such approaches as off-line materialisation and result-caching. An important special case is when the data is only changing monotonically, hence leading to incompleteness on stale datasets, but not to incorrectness. The severity of the results of staleness clearly depend on the use-case for which the data and queries are deployed.

Current benchmarking was widely seen as inadequate, with the often used LUBM and BSBM datasets as unrealistic and too regular (and hence too easy), but with the supposedly realistic Billion Triple Challenge datasets to be too incoherent. The LDSR dataset[1] promises to be a good alternative. Although the community has been fairly good at providing datasets, few if any of these datasets come with sets of queries to run in the benchmarks. Such benchmark queries were widely seen as badly needed. (Comment: since the Dagstuhl meeting, both the LDSR and the LLD datasets have been extended with a set of benchmark queries). Finally, it was felt that the community should look at the experience in other communities as regards benchmarking, in particular communities such as databases, theorem proving and information retrieval.

Finally, any work on scalability must of course take into account the rapid changes in computing environments: the possibility of multi-threading on multi-core machines, the increasing availability of compute clusters, either locally or "in the cloud", the increasing bandwidth available on wide-area networks, etc.

4 Provenance

Yolanda Gil, University of Southern California, United States
Harry Halpin, University of Edinburgh, United Kingdom

4.1 Introduction

The Semantic Web is finally taking off, but provenance is still a key critical missing component necessary for usage in most scenarios. In almost every discussion of deploying the

[1] `http://ldsr.ontotext.com/`

Semantic Web in real-world deployment, the issue of provenance is brought up. However, provenance has barely been a topic of research in the context of the Semantic Web. This is unsurprising, as even within the world of relational databases, the topic of tracking provenance has just begun in the last few years. This discussion at Dagstuhl helped overview current and future perspectives on the Semantic Web (including perspectives on relational data and scientific workflows in particular) and involved Paolo Bouquet, Carlos Pedrinaci, Oscar Corcho, Markus Krötzsch, Yolanda Gil, Harry Halpin, Frank van Harmelen, Ivan Herman, James Hendler, Deborah McGuiness, and Jeff Pan among other seminar participants.

4.2 Background

Provenance typically refers to the entities and processes involved in creating, managing, or storing an artifact. Provenance is important in a variety of contexts, such as business sourcing and manufacturing, scientific hypothesis tracking, origins of government policy and data, authenticity of cultural artifacts.

Recently, provenance has become of great importance to the Semantic Web. The advent of Linked Open Data has made it a primary concern of any data consumers to consider whether the data is usable based on its provenance. Reasoners in the Semantic Web need explicit representations of provenance of the information they use in order to decide what assertions and axioms to use. Provenance is also important in determining trust on agents and web resources.

Despite its importance, provenance is itself not a well-defined first-class subject and is interpreted differently in various fields. In particular, provenance is often conflated with issues of *trust*. However, this is a mistake, as trust metrics in general are more about whether or not a particular user "trusts" some data. Provenance is a more basic issue, concerned more about "where" the data came from without making any sort of judgement about whether or not the data is to be trusted. However, provenance information would under most circumstances be critical to establishing any reliable trust judgement. Furthermore, provenance can be used for applications far outside of trust as well.

4.3 Existing Work on Provenance

Research on provenance has been conducted in many diverse areas of computer science. The Semantic Web and agents communities have developed algorithms for reasoning about unknown information sources in a distributed network. Logic reasoners can produce justifications of how an answer was derived, and explanations that help find and fix errors in ontologies. The information retrieval and argumentation communities have investigated how to amalgamate alternative views and sources of contradictory and complementary information taking into account its origins. The database and distributed systems communities have looked into the issue of provenance in their respective areas. Provenance has also been studied for workflow systems in e-Science to represent the processes that generate new scientific results. Licensing standards bodies take into account the attribution of information as it is reused in new contexts.

4.3.1 Surveys on Provenance

There are many surveys of existing work on provenance from both a background in workflows [1] and database backgrounds [5]. However, the surveys and the research to date are not well known to Semantic Web researchers in general. In addition, it is not clear what aspects of prior work are relevant to Semantic Web. Given these issues, a better understanding of the state of the art in this area would be very beneficial.

4.3.2 Relational Databases and Provenance

Recent work in provenance on relational databases has aimed at creating a minimal vocabulary covering provenance, albeit for traditional relational data rather than RDF. In particular, the foundational work on provenance distinguishes between two kinds of provenance, the *where* provenance, which is the "locations in the source databases from which the data was extracted," and the *why* provenance, which is "the source data that had some influence on the existence of the data" [3]. Buneman et al. [3] present a Datalog-based model-theoretic semantics for calculating this kind of *where* and *why* provenance over queries. However, the traditional database community has been confronted with the same issues at the Semantic Web community with regards to Datalog, as it would be better for most databases to keep the provenance in pure relational data. Therefore, current theoretical database work attempts to create more realistic models of provenance whose formal semantics do not rely on Datalog yet can still trace both the *where* and *why* provenance and can be implemented on top of run-of-the-mill SQL databases [2].

4.3.3 Scientific Workflows and Provenance

Workflows capture the process and origins of datasets. Workflow systems can easily capture provenance of new data products. Provenance has been of great interest to scientific workflow researchers. There are biannual meetings to compare and contrast approaches in what is known as the *Provenance Challenge*[2]. The Fourth Provenance Challenge is planned for 2011. There is also a community effort to develop a standard representation for provenance, named the Open Provenance Model (OPM), which is still being extended by the community. More information on OPM is available online [3]. OPM is well known outside the scientific workflow community, and it would be useful to understand its applicability to provenance tracking in the Semantic Web.

4.3.4 Semantic Web and Provenance

Trust has always been an important layer in Web architecture. Berners-Lee's famous "Oh Yeah button" was meant to challenge the origins (provenance) of what is being asserted and request proof of those origins. Therefore, an important issue here is identity and verification. Do we trust *http://www.example.org*? The domain-name registration process allows a way for these URIs to be traced to individuals, and projects such as *OKKAM* can provide another layer of reliability, particularly for entities that do not already have a domain-name.[4] Furthermore, domain names as URIs may only the beginning. A popular technique

[2] `http://twiki.ipaw.info/bin/view/Challenge/`

[3] `http://twiki.ipaw.info/bin/view/Challenge/OPM`

[4] See `http://www.okkam.org` for more information.

to identify individuals over the web now involves *OpenID*. In addition to identifying entities, authentication of such identities is an issue. Furthermore, verification of what provenance information is asserted by whom and whether it is the true provenance of an object remains an open issue. These issues are being investigated in the context of security working groups for the Web.

Berners-Lee's *N3* toolset has also started supporting "proof-based" provenance, where the derivation of a fact is recorded. The advantage of the proof-based techniques is they allow a clear separation of provenance representation (i.e. the proof) and the calculations that serve as a strategy for determining trust.

A widely-implemented and simple model that is closely relevant, while not officially a standard, is *named graphs* [4]. Named graphs allow entire groups of graphs to be given a URI. Then further provenance information can either be stated in additional RDF statements using that URI. Although not yet officially a standard by itself, the named graph construct is built into SPARQL using the *GRAPH* construct.

Other approaches to provenance include the "Proof Markup Language" (PML)[5], and the "Provenance Vocabulary"[6].

There has been the beginnings of work connecting work from relational databases to the Semantic Web. Two of the most fruitful contributions have been a formal analysis in terms of "where-and-why" provenance for RDF graphs [8]. Also, practical software has been developed by Talis that helps deal with these issues, the *Changeset* vocabulary.[7]

4.4 Future Work

4.4.1 Research

It is clear over the next few years that provenance will be a major research area for the Semantic Web, vital to its successful deployment. Furthermore, as a "open-world" distributed system, the notion of provenance is even *more* important on the Semantic Web than it is in traditional "closed-world" databases. Thus, it is possible that the Semantic Web may be an ideal place for work on provenance *qua* provenance. Some major research issues are:

- What are the basic levels of provenance? What is the minimal units on the Web to attach provenance to?
- What aspects of provenance are directly of concern for the (Semantic) Web architecture? Are any of them separable?
- How to present provenance to both Semantic Web developers and end-users (especially for large scale problems)? Should this presentation differ based on different contexts? In particular, can we separate the presentation of provenance from its representation and storage?
- What are the issues in the control of provenance information?
- Can we explain provenance to users and developers, in particular what is happening in a larger context of an application or part of the Semantic Web?
- Can this work be integrated with work from cryptography and its current (lack of use) on the Semantic Web?

[5] `http://inference-web.org/wiki/Publications`

[6] `http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Provenance\_Vocabulary`

[7] Available at `http://vocab.org/changeset/schema.html`

4.4.2 Towards Standardisation

Also, at the Dagstuhl seminar it was decided to start a W3C Incubator Group to explore the possibilities of standardisation of provenance and the Semantic Web. Yolanda Gil agreed to chair and with the help of Ivan Herman, to write a charter for the Incubator Group and submit a proposal to W3C.

The W3C Provenance Incubator Group[8]) was approved in September 2009 with the goal of developing a state-of-the-art understanding and a roadmap for development and possible standardisation in the area of provenance. In particular, the group will:

- Develop use cases to illustrate the need and challenges of provenance recording, management, and use
- Articulate requirements for accessing and reasoning about provenance information
- Identify issues in provenance that are of direct concern to the Semantic Web
- Articulate relationships with other aspects of Web architecture
- Report on state-of-the-art work on provenance
- Report on a roadmap for provenance in the Semantic Web
- Identify starting points for provenance representations
- Identifying elements of a provenance architecture that would benefit from standardisation

As of January 2010, the group has 38 participants, including many attendees to the Dagstuhl seminar. It has produced more than 30 use cases, and organised them according to important research challenges in provenance. It has also collected more than 200 requirements motivated by the use cases. Like all W3C incubator groups, the group is chartered for one year so it will finish its deliverables by October 2010.

5 Personal Infospheres

Jérôme Euzenat, INRIA Grenoble Rhône-Alpes, France
Philipp Cimiano, University of Bielefeld, Germany
John Domingue, Open University, United Kingdom
Siegfried Handschuh, National University of Ireland, Galway
Hannes Werthner, Vienna University of Technology, Austria

5.1 Data Sharing and Semantic Web

Semantic web technologies are spreading to numerous applications: semantic desktop, semantic sensor networks, semantic web services, linked data, etc. The purpose of many of these applications is to collect data and to interpret them through interlinking.

On the side of users, the more bits of information are given away, the better the services they can have and the more the Semantic Web improves. Users want that the information provided to them is relevant to what they are doing or what they have to do. Hence, the bit of additional information that they give away (GPS coordinates, account information, etc.) should be used for providing contextual information.

[8] `http://www.w3.org/2005/Incubator/prov/wiki/Main_Page`

However, users also want their data to be protected: they will accept to give more information, if they can control how and by whom this information can be accessed. This is a question of balance between services and control.

This, of course, is related to the raising concern of "privacy" which applies to the Semantic Web as well as the general web. In a system like the Semantic Web, allowing for connecting all the information, the lack of control is an obstacle to the adoption of the technology. However, Semantic Web technologies can also be used to tackle the issue.

The standpoint of this contribution is to consider how people could be encouraged to give away part of the information they hold so that it can be used to the benefit of a wider group of people. Only by providing more control to users, it will be possible to have a more positive and reasoned data sharing.

5.2 The Continuum

The "privacy" problem is stated as if there were private data and public data. It is in fact ill-formulated. In reality, there are various kind of data with various degrees of privacy.

These degrees are related to:

- *what* data;
- *who* can access it;
- in what occasion (*when*): context;
- with what degree of detail (*how*).

Hence this problem is:

- multidimensional: it depends on the various dimensions mentioned above (what/who/when/how)
- gradual: instead of a simple private/public, the degree to which the data can be exposed, e.g., oneself, family, family-and-friends, public.

5.3 Control to the People

Examples here are centred around individuals. However, agents in this model can be people as well as services (in the web service sense). Indeed, it can be understood that information can be communicated to such a service for providing its benefit, e.g., disclosing where someone is for delivery. Similarly, services may hold information that other users may want to be disclosed.

Hence the important principle is that control over data must be decided by those who have data to disclose. This means that control over data is decided by individuals on their personal basis instead of by general policies, like social network software policies. Of course such policies may and certainly should play well with other policies such as corporate policies, but this is not the concern here. This may seem like access control. However, instead of rigid access control schemes, it is necessary to define access control in function of flexible concepts closer to the individuals (instead of dictated by operating systems and administrators).

The use of semantic technologies should provide more flexibility (because they can be extended) and more precision (because they can go to deep levels of details). So they are an ideal tool for giving people the control over their information.

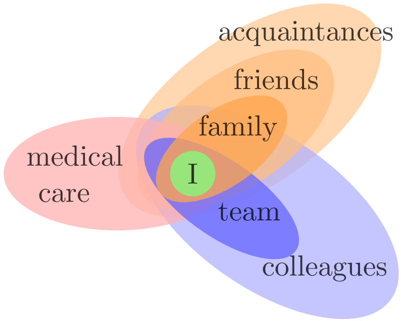

Figure 1: Typical spheres of relationships (who). Spheres are always centred on the individual.

5.4 The Sphere Model: Approximating the Continuum

There is a, from the user experience point of view, a continuum from data stored on one's disk, phone, social network accounts, the "cloud" and the web. There is also a seamless continuum between personal sensors, those in phones and computers, and mass sensors, CCTV cameras. It is very difficult to control this data: indeed, who has access to the MAC address of the WiFi card in a telephone when it is connected to a network?

It is difficult to prohibit or grant access to information in general (like distinguishing in the absolute that something is public or private), so we propose a general model for expressing the multidimensional space in which this data evolve. It is called the "sphere model" and is illustrated by Figure 5.4.

A sphere is a set of elements (person, event, context) identified by a name. It defines a compact area in a space centred at one point, e.g., the user. A system of spheres is a set of such spheres centred on the same point and partially ordered. So, basically these spheres are organised in a direct acyclic graph.

The spheres defined here are rather "personal spheres". They are in fact related to what is called "personal information". Of course, one could also consider corporate spheres or national spheres. Personal information on the Semantic Web is modelled by project such as Nepomuk[9]. This means that starting from her semantic desktop, a user has already a lot of information available for defining spheres. Our goal is to use the information that matters to people (PIM data) to control the access of information that matters to people.

5.5 Who: Spheres of Relations

One can have various spheres: personal, colleague, people sharing a train compartment, people with whom one is currently writing a paper, etc. They are dynamic and can be subdivided into close family, work team, etc.

These kind of spheres are basically a group of agents which have the same rights with regard to (part of) one's data. Having these categories is convenient for granting or prohibiting access, based not on individual identities but on their belonging to a sphere.

Fortunately, these sphere descriptions are everywhere in address books and social networks. Hence they can easily be expressed in RDF and other semantic technologies. General

[9] `http://nepomuk.semanticdesktop.org/`

groups can be defined through classes either in extension (the list of people in one's team) or in intension (all the people in one's address book, all those working in the same company as oneself). One particular individual may be identified directly (through a URI) or indirectly through her role (like "my boss").

Identifying people can be achieved by specific technologies like foaf+ssl[10] (but this is out of focus here).

5.6 What: Characterising Data

However, this is not all: data is also organised in such spheres. A calendar contains various types of events: personal, work, sport events. Not only access may be granted depending on the sphere someone belongs to, but it may depend on the sphere in which this information belongs to.

Of course, one's music band members will have access to the information about gigs, locations, etc. One's family may have it as well, but colleagues can only know that the person is unavailable. On the opposite, the band members will only know from business schedule that one is not available for practising at some periods, while colleagues will have more precise information.

In terms of semantic technologies, these items can be characterised by classes and properties (sport events, family gatherings, work meetings) and filters about what is accessible can use designed in SPARQL or rule languages.

5.7 How: Granularity

Not only access may be granted depending on the sphere someone belongs to, but granularity of what is disclosed may be granted depending on the spheres people are in.

Granularity offers a gradual access to information over time and over spheres. Consider the following pieces of information describing the same event:

- "I am not available"
- "I am in Karlsruhe from October 26 to October 30"
- "I am working on a paper about S-match and algebra of relations in room 452 of Novotel with Paul, on Monday, October 28th, from 17h to 19h."

They are three representations of the same thing at three different levels of granularity. Moreover, the granularity changes apply in three different dimensions:

- spatial,
- temporal,
- thematic.

The point with granularity is to be able to present information at different levels of detail so that only what is useful for the reader is available. This adaptation aims at efficiency by reasoning at the required level, and, may also be aimed at protecting privacy.

This can be defined through rules and/or views which alter information rather than simply filtering it, like: for sports events, colleagues can know that I am not available but in town.

[10] `http://esw.w3.org/Foaf+ssl`

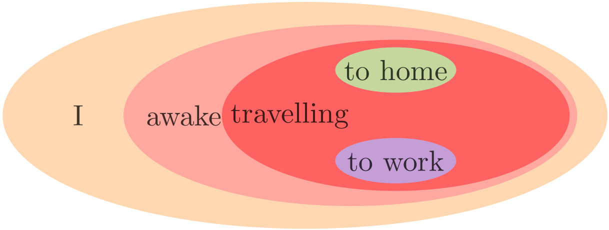

Figure 2: Context and granularity as spheres.

5.8 When: Characterising Context

We call context information that characterises the situation, e.g., current task, location, time, purpose of disclosure, etc.

One may not want to disclose precise information about her localisation at anytime, but for a colleague in the surroundings, e.g., at the same conference, looking for her, this information may be useful to deliver. However, in particular contexts, e.g., an extraordinary event has happened, very important information (like medical data) may also need to be disclosed.

This information may of course be expressed with Semantic Web technologies. But it can also be expressed as a sphere: a travelling context is a narrower context than being simply awake and a wider context than driving to work.

In terms of Semantic Web technology, the disclosure can again be processed by filters which will consider if the context falls within a sphere in which the information can be communicated.

However, for characterising the context, one needs to ask the requester for information about his or her context, that he may not be willing to disclose. Hence, disclosure of information will be made on the basis of negotiation between two parties when one has to disclose the reason why some information is wanted, in order to have this information delivered. This negotiation will use exactly the same techniques as above. However, specific protocols may be needed for guaranteeing that the process converges to an acceptable result.

5.9 Conclusions

Semantic Web technologies provide a very good basis for a more flexible and precise control over disclosed information. For this purpose, we introduced the abstract notion of spheres which symbolise the multiple and contextual aspects of situations.

We have identified five challenges for implementing this sphere approach:

- How to specify spheres and access policy;
- How to evolve spheres and create ad hoc spheres (like the train compartment);
- How to navigate between multiple context embeddings;
- Interaction between spheres and context.
- How to negotiate information access (on technical, social, or legal grounds);

This section focusses on data access. However, spheres can be used for other purposes. For instance, users can use the same spheres or other ones for ascribing trust to information once they know their provenance (which was another important topic at the seminar).

5.10 Related Projects

There are already several project that can be considered as providing some ground or early experimentation for the principles presented here:

Nepomuk by developing the concept of Semantic Desktop and providing a semantic version of PIM categories has paved the way to the introduction of spheres.

Iyouit is a project of DoCoMo labs experimenting with the exploitation of automatically extracted context information from mobile phones. It does already use semantic PIM information in order to implement access control to the data[11].

Persist has developed the notion of personal smart spaces which tries to provide services to users based on their contexts and preferences. When two users encounter, their smart spaces can interact and exchange information in a controlled way[12].

6 Interaction

Lora Aroyo, Vrije Universiteit Amsterdam, The Netherlands
Valentina Presutti, National Research Council (CNR), Italy

The rising amount of data on the Semantic Web allows for developing applications to search, query, and analyse data described in arbitrary vocabularies.

The group identified a number of challenges for interacting with such data that has been integrated from thousands of sources:

- What general, domain-independent interaction methods beyond keyword search are possible?
- How can results of a sequence of interaction steps be visualised appropriately?
- In lieu of a predefined schema which can be used to display data, which other mechanisms are viable to prioritise and arrange data in an effective manner?

7 Mobile and Real World Networks

Anna Fensel, FTW Forschungszentrum Telekommunikation Wien, Austria
Oscar Corcho, Universidad Politécnica de Madrid, Spain

7.1 Introduction

In this report we summarise the starting points and the discussion outcomes of the “Mobile and Real World Networks” Dagstuhl working group with respect to the status of the field, open issues and perspectives on how to address the issues.

7.2 Status of the Field

The increasing availability of mobile devices, which can be connected almost anytime anywhere to the Internet, together with cheaper, more robust, deployable, mostly wireless sensor

[11] http://www.iyouit.eu/
[12] http://www.ict-persist.eu/

networks, and the possibility of using mobile devices as a large-scale set of mobile sensors that are carried by persons, has increased the variety and heterogeneity of Internet applications and data sources, increasing the relevance of research in the areas of mobile and real world networks.

However, none of these fields, nor their combination, have yet reached the success of the Web. This is mostly because the Web protocols, APIs, applications, etc., were originally developed with the idea of using traditional databases as their backends, and with server to server or server to browser interactions in mind, hence neither considering mobile applications and devices, nor stream data sources. In both cases, a good number of challenges can be associated to the restrictions that devices impose in their ability to run high resource-consuming applications, and the needs to handle different connection qualities of service and to consider unreliable and noisy data.

Some of the most relevant challenges for mobile platforms are related to their hardware and connectivity restrictions: bandwidth, memory and CPU availability, storage capacity, connectivity options and issues, security and user interaction and display. For the eventual end-user relevant impact, mobile services are to repeat or even exceed the success of the Web if they in particular become simple to use, find, trust, create and set up. Research on mobile platforms and mobile services are also closely related to research on telecommunications networks, i.e. efficient information exchange protocols, quality of service.

In the context of real world networks, the terms Sensor Web and Sensor Internet have been coined to refer to the combination of sensor networks with Web, Web service and database technologies. These efforts focus on providing extensible platforms that contain fine-grained modular services and can be used as building blocks for developing sensor-based applications. These platforms include libraries for common domain-independent tasks, such as data cleaning, storage, aggregation, query processing, etc., and can be used to provide domain-specific aggregated services (e.g., coastal imaging, patient care, etc.). Sensor Web efforts are characterised by: variability in data, devices and networks (including unreliable nodes and links, noise, uncertainty, etc.) and in application requirements; the use of rich data sources (sensors, images, GIS, etc.) in different settings (live, streaming, historical, processed); multiple administrative domains; and multiple, concurrent, uncoordinated queries to sensors.

7.3 Open Issues

A number of open issues can be identified in these areas, and classified according to whether they are related to the role of communication and content/service providers in the provision of new types of services across platforms, to the computational complexity derived from the emerging needs of these services, and to the complexity of the application domains where they are used. In more detail, some of these challenges are:

- Communication and content/service provider challenges The role of telecommunication companies have to change from their main focus in the provision of access services to the provision of added-value services on top of their infrastructure, enabling new business models that consider users not only as consumers, but also as “prosumers” (providers and consumers of services and content at the same time). Together with this, users have to be better assisted with respect to how to use these new services, making them easier to employ, and hiding their complexity and heterogeneity, especially in terms of providing multi-vendor and multi-technology platform support. Other set of challenges in this area can be related to the quality of service, seen in many occasions as the need to provide these services timely, accelerating the creation and delivery of services, hence reducing the time-to-market for new services.

- Computational complexity challenges for the new service needs An emerging set of representational and computational needs can be derived from this range of services, ranging from the need to manage spatio-temporal information adequately and with different degrees of scale and uncertainty, to the need to handle reasoning on the existing information at different levels of abstraction, and the need to handle large amounts of data coming in the form of streams. Other challenges related to the management of this type of information are related to the distribution of computation and data, which generates new needs in terms of query processing and data integration.
- Application domain complexity challenges Some of the complexity challenges come also from the complexity of the application domains that are being covered by mobile services and sensor networks, what requires consideration of the need to provide better tools to end users to generate on-the-fly combinations of different data sources, mainly in the form of mash-ups.

7.4 Perspectives of How to Address the Issues

Semantic technologies are only since recently entering the mobile and sensor networks fields, in Europe often in projects going under the umbrella of the “Future Internet”. At present semantic technologies promise a large facilitation and resolution of the indicated challenges, and their common usages here are as follows:

- Semantic data modelling - the “easy” part (or low-hanging fruit) comprises: i) agree on or develop a network of sensor or service network ontologies and ii) use these ontologies to annotate data, e.g. SensorML readings, or user profiles, where (semantic) vCard is already a standard for mobile devices. Though such modelling is common for Semantic Web researchers, numerous explanations and adaptations for other communities, such as geographers, sociologists, are often needed.
- Integration with the (Semantic) Web is inevitable for having a common large information pool, Linked Open Data cloud-like. Ontology technology will facilitate handling heterogeneity and variety of information sources. Creation, discovery, composition of enablers and services is to be accelerated on the basis of shared ontologies and semantic techniques. Data stream management and reasoning techniques for semantic data would need to de adopted.
- Semantically enabled smart user interfaces will link humans to the mobile and real world.

8 e-Science

Yolanda Gil, University of Southern California, United States

This discussion group focused on whether there is something fundamentally different between science and other activities. Is there something to e-Science that poses fundamentally different requirements from other Semantic Web communities and applications? What kinds of symbiosis could be established between what is happening in science and what is happening in the Semantic Web?

The lines that have traditionally separated science from regular life are blurring thanks to the web. Traditionally, science has been done in experimental laboratories, guided by the scientific method, and published in strictly reviewed journals. Today, scientific research and publication have been forever changed with the web. Consider citizen science. Regular people can now log into sites set up by scientists to make contributions to scientific questions. After a bit of training in the analysis of stellar objects galaxyzoo.org allows people to contribute assessments of images of the sky regarding whether they contain particular kinds of objects. The site received 1M hits in its first day, a testament to the thirst of private citizens to contribute. A teacher in Belgium made unique observations to this site that have resulted in scientific paper submissions. Another notable citizen science effort is ebird.org, where people can contribute sightings of birds in their neighborhoods to help scientists study migration patterns. There are many such sites where regular citizens can contribute to scientific efforts that would otherwise not have the resources to collect and analyze the large amounts of data needed. Another way in which the lines are being blurred between science and regular life is citizen involvement in medical treatment and personal medical records. Many sites are now active where information about specific diseases is being shared by patients, questioning the benefits of specific drugs and sharing the results of alternative medicine and understudied treatments. Science is being transformed by the sheer size of the contributions that private citizens can share through the web.

There are aspects of science that can benefit the Semantic Web at large that were discussed in the breakout group. We mention here two topics: 1) the new approach to scientific publications that could be adopted by all web documents and publications, and 2) the process of creating ontologies that is used in scientific communities and could be used to populate the Semantic Web.

Today, scientific publications are being reinvented with the web as a substrate. Indeed the web was originally conceived to support scientific publications and collaboration. Many scientific publications now include the data that supports the experiments described in the paper, method details, and all manner of supplemental materials. The concept of provenance is paramount, where each new result is backed up with details about the data, information, and processes that were used to arrive to that result. Cross-references among papers, linking of papers to ontologies by manual or automatic methods, and the creation of thematic resources that compile related results (eg, the Rat Genome Database), all contribute to the extensive hyperlinking of scientific products. The web could benefit from better provenance and justification mechanisms that are so important in science but also in all facets of information analysis that is commonly done by regular web users.

Scientists have embraced the use of ontologies and Semantic Web technologies as useful tools to organize and integrate knowledge, particularly in the life sciences. There are many examples of processes established to create ontologies, typically involving a combination of a curation team, an editorial board, and a well defined process for community testing and feedback. In contrast, the Semantic Web is shifting de facto its emphasis from ontologies to linked data. This might be because web users at large have not figured out how to organize themselves into communities of interest associated with a creation processes for ontologies. It might be beneficial to collect best practices that have worked in scientific circles, and share them through a standards body such as W3C. If it were as easy to stand up and run a community ontology as it is to stand up a web site, it is easy to imagine the Semantic Web being populated with large amounts of grass-roots ontologies that would be linked to usage and therefore data.

9 Conclusion

In this workshop participants from academia, industry, and government presented and discussed further directions for Semantic Web research. In general, the field of Semantic Web research has matured in the last decade (cf. [9]). One indication for that, among others, is the discussion of advanced issues such as scalablity. More data is becoming available and vocabulary managment is being operationalised, and research on provenance tracking and technologies and methods addressing privacy concerns has commenced. How users can appropriately interact with the flood of data is an open question. Some prominent areas for applications of Semantic Web technologies we have covered in the report are e-Science and mobile and sensor networks.

One result of this meeting was that a number of related and actionable conversations happened. In addition to discussions leading to workshops co-located with WWW and ESWC other conversations resulted in further actions. For example, four attendees (Brickley, Chaudhri, Halpin, McGuinness) of the Dagstuhl meeting had discussions while at the meeting and decided to run a related meeting called Linked Data Meets AI[13]. Efforts are under way to standardise some of the topics covered at the workshop, for example the W3C Incubator Group on provenance initiated by Gil and Herman. Moreover, a number of funding proposals have been submitted based on the discussions at the workshop.

References

[1] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Journal of Web Semantics*, 5(2), 2007.

[2] P. Buneman, A. Chapman, and J. Cheney. Provenance management in curated databases. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pages 539–550. ACM Press, 2006.

[3] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. In *ICDT '01: Proceedings of the 8th International Conference on Database Theory*, pages 316–330. Springer, 2001.

[4] J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs. *Journal of Web Semantics*, 4(3):247–267, 2005.

[5] J. Cheney, L. Chiticariu, and W. C. Tan. Provenance in databases: Why, where and how. *Foundations and Trends in Databases*, 4(1):379–474, 2009.

[6] D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster. Dagstuhl-seminar: Semantics for the WWW. Technical report, 2000.

[7] D. Fensel, J. A. Hendler, H. Lieberman, and W. Wahlster, editors. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential.* MIT Press, 2003.

[13] `http://www.foaf-project.org/events/linkedai`

[8] P. Pediaditis, G. Flouris, I. Fundulaki, and V. Christophides. On explicit provenance management in RDF/S graphs. In *Workshop on the Theory and Practice of Provenance, in conjunction with the 7th USENIX Conference on File and Storage Technologies*, 2009.

[9] S. Staab and R. Studer. *Handbook on Ontologies.* Springer, 2nd edition, 2009.