PlanetData

Network of Excellence

FP7 – 257641

D7.2 PlanetData Roadmap

Coordinator: Lyndon Nixon and Simeona Cruz Pellkvist (STI2)
With contributions from: Thomas Bauereiß (UIBK), Graham Hench (STI2)

1st Quality reviewer: Oscar Corcho (UPM)
2nd Quality reviewer: Giorgos Flouris (FORTH)

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	31-03-2011
Actual delivery date:	31-03-2011
Version:	1.0
Total number of pages:	27
Keywords:	Roadmap, Trends Semantic Technology, Trends Linked Data Management, Trends Databases, Trends Data Markets, Trends Data Stream Management

Abstract

The Planetdata Network of Excellence presents the PlanetData roadmap to the consortium. This roadmap will be used as input in the Planetdata programs and will use this roadmap as the consortium initial motivation.

[End of abstract]

Executive summary

The goal of this deliverable is to deliver the roadmap of Planetdata Network of Excellence. This deliverable describes the intended PlanetData roadmapping activity, which will also be used as input for the PlanetData programs. The PlanetData research activities within the community will be supported through a roadmapping activity. This roadmap is created through consultations with experts from the main research communities, and representatives of technology vendors.

The first section contains the introduction of the PlanetData roadmap that explains why the PlanetData consortium needs to have this roadmap. The second part of this deliverable introduces what methodology has been used to create the PlanetData roadmap. There are 4 steps of methodology used to create the PlanetData roadmap. The 4 steps are described more in this section. On the next section current situation of the large-scale data management is described according to the situation when this deliverable is written. In this section we try to describe what is the current situation technology of large-scale data management. Before we go into the main point of our roadmap, which is on the fourth section of this deliverable, challenges of new scales of data being produced, consumed and processed by IT systems are described in some paragraphs.

Fourth section is the main point of this deliverable. The relevant technology that related to the large-scale data management is explained. The relevant IT domains are described in different key trends of Planet Data topics, future developments, research challenges and addressable gaps. The key trends are semantic technologies, linked data management, databases, data markets and data stream management.

On the last section we summarise the results in the conclusion. We present how critical each of the presented technology for short, medium or long term as the conclusion. The PlanetData Network of Excellence will use the result of this roadmap as an initial motivation for further research in the identified areas particularly and for the large-scale data management generally.

Document Information

IST Project Number	FP7 - 257641	Acronym	PlanetData
Full Title	PlanetData		
Project URL	http://www.planet-data.eu/		
Document URL			
EU Project Officer	Leonhard Maqua		

Deliverable	Number	D7.2	Title	PlanetData Roadmap
Work Package	Number	WP7	Title	Dissemination and community building

Date of Delivery	Contractual	M6	Actual	M6
Status	version 0.1		Final □	
Nature	prototype □ report X dissemination □			
Dissemination level	public X consortium □			

Authors (Partner)	STI2			
Responsible Author	Name	Simeona Cruz Pellkvist	E-mail	simeona.pellkvist@sti2.org
	Partner	STI2	Phone	+43- 1 23 64 002

Abstract (for dissemination)	The Planetdata Network of Excellence presents the PlanetData roadmap to the consortium. This roadmap will be used as input in the Planetdata programs and will use this roadmap as the consortium initial motivation.
Keywords	Roadmap, Trends Semantic Technology, Trends Linked Data Management, Trends Databases, Trends Data Markets, Trends Data Stream Management

Version Log			
Issue Date	Rev. No.	Author	Change
24-01-2011	0.1	Simeona Cruz Pellkvist	Initial Draft ToC
26-01-2011	0.15	Lyndon Nixon	Initial Draft ToC revised
07.02.2011	0.3	Thomas Bauereiß	Current situation, Link Data Management, Databases, Data Markets
14.02.2011	0.6	Simeona Cruz Pellkvist	Introduction, Roadmap Methodology, Trends in key Planet Data topics, Semantic Technologies, Service Technologies
24.02.2011	0.8	Lyndon Nixon	Revised introduction and trends, wrote conclusion
03.03.2011	0.9	Lyndon Nixon	Refinement and completion of sections following expert round table and F2F discussions in Innsbruck meeting
07.03.2011	1.0	Simeona Cruz Pellkvist	Final editing prior to QA
25.03.2011	1.3	Lyndon Nixon	Further refinement after QAs
30.03.2011	1.5	Simeona Cruz Pellkvist	Another refinement prior to final EC submission

Table of Contents

List of figures and/or list of tables

Abbreviations

IT – Information Technology

KDD – Knowledge Discovery and Data Mining

OWLIM – n instance of a Semantic Repository

OWL-DL – Web Ontology Language-Description Logics

SKOS – Simple Knowledge Organization System

SOA – Service of Agreement

SPARQL – SPARQL Protocol and RDF Query Language

RDF – Resource Description Framework

RFID – Radio-Frequency Identification

1 Introduction

The PlanetData Network of Excellence brings together expert research organisations in Europe to address the future needs arising from the increasingly large scales of data being produced and shared on the Internet. In reaction to this new technological challenge, it is vital to identify the key IT domains which can provide new solutions to producing, publishing, distributing, consuming, manipulating and processing previously unimaginable scales of data. However, these domains today are only able to represent a partial solution, since they have their own limits and gaps.

The goal of this PlanetData roadmap is to analyse the current trends in those key IT domains and postulate future developments in each area. We then turn to the question whether the currently foreseeable developments will also adequately address the challenge of large-scale data. As a network, PlanetData will identify research challenges and addressable gaps which form a call to action for the next years, so that future IT will truly be ready for the scales of data future IT systems will be called upon to produce, consume and process.

In the following chapter 2 we introduce the **Methodology** followed by this Roadmap.

Then, in chapter 3, we ground this work in the **Current Situation** which is the looming challenge of new scales of data being produced, consumed and processed by IT systems.

Chapter 4 turns to each of the relevant IT domains in turn, identifying therein the **Trends in key Planet Data topics**, future developments, research challenges and addressable gaps.

We summarise the results in the **Conclusion** in Chapter 5.

2 Roadmap methodology

There are many different methodologies for roadmapping. In this section, the PlanetData project presents the methods that are used in creating the roadmap. We choose a pragmatic approach within the constraints of the time and resources available, and making use of the access to experts within the Network of Excellence.

The large-scale data management roadmap for PlanetData was created in a multi-step process as follows:

Identify problem areas and propose realistic solutions

Our project aims to enable maintainable large-scale access to structured data that has been exposed by organisations within the European community. To support the future production and consumption of large scales of data, the methodology behind this roadmap focused on compiling a collective perspective on the most prominent problems, and proposed solutions, of large-scale data management.

Identify potential trends

Following the structure of the PlanetData Network of Excellence it can be seen that the network experts have already focused the relevant IT domains for providing solutions to future large scale data management into four main areas: semantic technology, linked data, databases, data markets and data streams.

Evaluation, refinement

A round table discussion held during the Innsbruck meeting of the PlanetData consortium (February 28, 2011) involved all of the present experts in evaluating and refining the initial inputs to the roadmap. This was followed up by several face-to-face discussions with individual experts to clarify last points and complete the contribution.

Publication of technical roadmap

The roadmap will be published as a PlanetData deliverable and will be distributed through all dissemination channels used in the project.

3 Current Situation

Several major information and communication technology trends underlie the explosive growth expected in data provision in the coming years:

- Transition of closed enterprise IT systems (intranets, data silos) to open Web-based models such as SOA and cloud, where business data is stored and shared across the Internet, promoting data availability, cross-enterprise data exchange, and integration with other data sources[1];
- Uptake of 'open data' principles for large data sets whose Web-based publication using structured vocabularies can be for the public or social good[2];
- User-generated content is becoming the form of content on the ("user") Web[3];
- An Internet of Services, meaning increasingly more social and commercial transactions will take place over the Internet, both presuming increased data provision to inform those services and increased data generation from the use of those services[4];
- An Internet of Things, meaning an explosive growth in devices connected to the Internet, leading to a parallel explosive growth in the amount of data being created and processed by those devices (e.g. sensor networks providing streams of measurement data, RFID tags tracking object location or human-connected devices reporting medical condition)[5].

Hence the scale of data being produced and consumed is growing continually. Large-scale data management is becoming a requirement for more and more stakeholders. Businesses and enterprises increasingly rely on the ability to make sense of large amounts of data, in order to make informed business decisions and in order to deliver innovative services and applications to their customers. This applies not only to data generated inside the enterprise, but also to data from external, heterogeneous sources that need to be properly integrated, as enterprises transition from closed IT systems to more open models in order to leverage the data available on the Web. Complementary, the growth of the Linking Open Data cloud shows a trend among organisations and institutions, particularly in the public sector, to make large amounts of data openly available. Examples for this are the datasets available from government portals like data.gov.uk or data.gov, or the scientific dataset from the biomedical domain at LinkedLifeData.com. Another trend that will play an important role in the coming years is the Internet of Services, where a growing number of social and commercial services will take place over the Internet, acting as both consumers and producers of data.

Another situation with regards to large-scale data, the current development of Cloud Computing is, that everything is becoming a „service". There is already, i.e. „Software-as-a-Service", „Platform-as-a-Service", "Human-as-a-Service", etc. Another trend into this direction is „Data as a Service". In general means that big amounts of datasets are provided for commercial or free. Data-as-a-service is becoming more popular now and use by more people as the centre of information of their businesses. Some are commercialized and sold for different purposes. As an example NASDAQ has created a service, which allows their customers to

[1] Forrester predicts data warehousing will evolve into a "virtualized, cloud-based, supremely scalable distributed platform". As an example of data scale, eBay has 6.5 petabytes of data in its "Enterprise Data Cloud". (source: http://cloudcomputing.sys-con.com/node/992716, June 2009)

[2] The Open Linked Data initiative (http://linkeddata.org) has already led to the publication on the Web of billions of statements of public data; the US government is releasing public data in structured form at the site http://www.data.gov; European governments, starting with the UK, are following suit

[3] Every week Facebook users share 1 billion information items, and upload 2.5 million videos and 225 million photos. Twitter sees over 1 million tweets a day. YouTube receives 20 hours of video every minute. The IDC estimates that nearly individuals will generate 70% of the digital universe in 2010. By 2011 there will be 2 billion users of the Web. By 2020 the average individual's information footprint will grow from 1 terabyte today to 16 terabytes.

[4] A major driver of services is the mobile/telecommunication market. In 2011, there is forecast to be 4.39 billion mobile subscribers. Mobile data revenues are growing at 16% annually. The smart phone market grew 15% between 2008 and 2009. The iPhone application store has more than 85 000 apps, which were downloaded over 2 billion times in the first year, as a major indicator for the growth of mobile social and commercial services.

[5] A Harbor Research report on Pervasive Internet & Smart Services sees a potential 3.5 billion mobile Internet devices and 1.75 billion controllers & sensors by 2013. There may be as many as one trillion devices in total connected to the Internet (cars, cameras, home appliances etc.)

access their data by Web API instead of downloading massive data or look at it online. This service called NASDAQ Data-on-Demand.

Data-as-a-Service provider will increase their numbers in the next few years. At the moment there are already few providers available online, e.g. Google Squared, Factual, InfoChimps, etc. Each company use different architecture, API, tools, etc. The data provided by these companies are not specialized. Customer who wants specific data will need a specific service. This specialized service will become important for ease of use, search and processing.

With these developments, the current state of the art in data systems is being pushed to its limits. To handle data at large scale, suitable technologies and know-how are required. Database systems need to provide high-performance and highly scalable storage and querying of data. Datasets from heterogeneous sources, including the Linked Open Data cloud, need to be integrated and processed, taking into account different data models, provenance and quality assessments of datasets. Web services need to be provided in a way that allows their easy discovery, federation and execution. Stream data, e.g. sensor readings or newsfeeds, is becoming increasingly important, and requires appropriate modelling and representation, real-time processing, filtering and analysis.

The combination of these trends will lead not only to previously unimaginable scales of data being stored, processed, managed, shared, and exploited over the Internet, but also that the data will be being provisioned in various formats, at various levels of quality, in different contexts, for different usage purposes, and rapidly changing over time. In the following section, we describe important trends in the area of large-scale data management addressing these issues, identified by experts inside and outside the PlanetData Network of Excellence.

4 Trends in key Planet Data topics

In order to pursue the goals of the roadmap in the PlanetData Network of Excellence, we have identified trends in several relevant technology domains for addressing large-scale data management, in order to anticipate future developments in each area parallel to the increases in data scale being handled.

From this, the PlanetData experts have identified sets of research challenges and addressable gaps.

4.1 Trends in semantic technologies

Trends in semantics were first addressed in the KnowledgeWeb technology roadmap and, subsequently, the ServiceWeb 3.0 roadmap on semantics [1]. The latter has also inspired a book chapter on Future Trends relating to semantic technology [2].

Semantics, referring to the use of structured data models which are based on formal domain models known as ontologies, have grown in relevance to IT solutions as data has become more heterogeneous and complex, while the importance of data integration and validation has risen. The use of ontologies, which are based on logics, make semantic data suitable for processing on the basis of logical inference, hence permitting new inferences to be made from the provided data, or more semantically-correct validation of content to be pursued. The main stumbling block for the update of semantic technology has been its maturity, and resulting lack of enterprise mature tools for creating, publishing, consuming and processing semantic data. For the World Wide Web Consortium (W3C), semantic technology is on the cusp of maturity and widespread uptake, at least among early adopters, please refer to Figure 1.

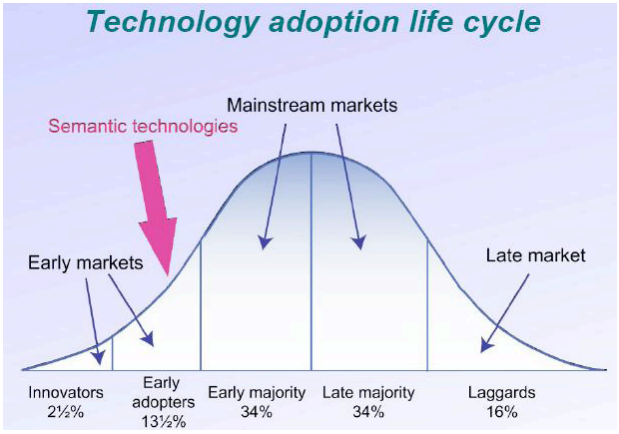

Figure 1. Semantic technologies adoption life cycle (October 2009)[6]

The Gartner group, on the other hand, has also predicted the mainstream adoption of the Semantic Web, albeit after 2018, as seen in the "Priority Matrix for Web and User Interaction Technologies 2008".

It seems at least clear that this decade (2010-19) will be a significant one for semantic technology. However, as semantic technology reaches the mainstream market will it be ready for the large scales of data it will encounter there?

Scalability has been repeatedly mentioned as a limitation for semantic technology, based as it is on computationally complex logical formalisms. There is an unavoidable trade-off between (logical) expressiveness and (computational) performance. Classical solutions to this trade-off such as data distribution solve less when dealing with semantic data due to the need to replicate ontologies across nodes if reasoning is still to be possible. Finding the most expressive logic while minimizing the effect on computational efficiency has become a key area of research, together with techniques for making semantic data storage, retrieval and reasoning as efficient as possible. We note some of the key trends in semantic technology relevant to large-scale data management:

[6] From W3C / Ivan Herman cf. http://www.w3.org/People/Ivan/CorePresentations/Applications/

4.1.1 Annotation

In the context of the Semantic Web *annotation* is understood as the process of assigning terms from an ontology to a resource (e.g., a text document, Web page, image, graphics, audio or video) or a part of a resource (e.g., a sentence, term, spatial/temporal region of a media item) in order to describe the resource in a machine-understandable way. End-user applications, which benefit from the availability of semantic annotations are numerous and range from information and knowledge management to personalization or data integration.

Despite promising advances, approaches that can be generically and efficiently applied to automate annotation across media still remain to be defined. In contrast to textual resources, which are annotated automatically to a large extent, non-textual media semantic annotation heavily relies on human input. Other challenges in this context include solutions for cross-media annotation, as well as the scalability and portability of existing solutions to other domains.

Semantic annotation of texts relies on existing technology for information extraction, named entity recognition or document classification, which can be adapted to the purpose of semantic applications. What has still not happened on a large scale is the automated interconnection of information from different sources or even different media. In an enterprise setting, for instance, information is distributed across a high number of different repositories and has to be consolidated to solve problems. This is not only true for knowledge management but also in other business intelligence scenarios. The support of individual knowledge workers still lacks contextualization, which results in systems that still to some extent suffer from information overload. This is an issue especially when information should be delivered as targeted or as timely as possible.

In the area of legacy data integration, which aims at bootstrapping annotation by exploiting richly structured data sources to make them part of the annotated Web, solutions have been proposed to convert relational databases to RDF or to provide wrapper-based approaches to query databases using SPARQL. Furthermore, tools to semantically enrich office documents like emails, text documents or Excel sheets and to provide rich semantic annotations of these resources will become mainstream.

The area of computer vision provides methods to make visual resources eligible for machines. Recent years have seen considerable advancement in the range of things, which are detectable in still and moving images. This includes object detection scaling up to a considerable amount of different objects for some tools, to object tracking in e.g. surveillance videos. All of these approaches try to derive meaning from low-level features (like colour histogram, motion vectors, etc.) automatically. Despite constant advances these tools are still not capable to exploit the full meaning of visual resources as not all meaning is localized in the visual features and needs human interpretation. Two ways are currently followed in current research: The first one is to provide rich human annotations as training data for future automated analysis. The second one relies purely on analysis of raw content, which only performs well for specialized domains and settings in which relevant concepts can be easily recognized. Richer semantics, capturing implicit features and meaning derived from humans cannot be extracted in this manner. Present trends put therefore the human more and more into the loop by lowering the entry barrier for his participation. This is done by adopting Web2.0 or game based approaches to engage users in annotation of visual resources. Recent approaches e.g. try to support automatic analysis with tagging or vice versa. What is still missing are approaches that are capable of exploiting more high level features also in visual resources of lower quality and which can be adapted across domains.

Due to aforementioned issues, multimedia analysis has still to be supported by end users to a great extent. What is still needed are tools that allow to capture subjective views of visual resources and combine these views to deliver a consolidated objective view that can represent a view which holds across users. Also scalability is an issue here. While tagging based approaches are proven to ease large-scale uptake, motivating users to provide more meaningful annotations is still an issue.

4.1.2 Knowledge extraction

This refers to the act of extracting useful knowledge from data making use of analytic, statistical and mining techniques. The persistent and rapid growth of data to be processed in IT systems due to the Internet, increased digitalization of content and processes, and trends such as sensor networks have created an immense need for KDD methodologies. With respect to semantic technology, KDD can aid in extracting

useful knowledge for ontology modelling (cf. the next point) as well as ontology population (cf. i) Annotation above).

The awareness of large collections of data that can be used for extracting knowledge has been raised with the advent of Web 2.0. However, Web 2.0 is not the only place where large amounts of data accumulate. The automatic extraction of knowledge (and lightweight semantics) is extremely relevant not only on a Web scale, but even more in closed world environments such as enterprises. Intelligent methods for the automatic extraction of lightweight semantics can be another step in overcoming the knowledge acquisition bottleneck.

There is a great need for interfaces for knowledge capture that comply with high usability requirements. Semantic wikis and games for semantic content creation are examples for interfaces that capture knowledge and try to address a large community. A related challenge is addressing closed world environments. Moreover, it is likely that there will not be huge tools for ontology building and annotation but many small plug-ins that allow work-integrated creation and maintenance of semantics plus immediate re-use of those semantics (comparable to tags).

4.1.3 Ontology engineering

Developing ontologies requires domain expertise and the ability to capture this knowledge in a clean conceptual model. The complexity of creating good ontologies has limited the amount of ontologies being generated and proves problematic as the application of semantic technology across a wide spread of data sources becomes more commonplace, if the advantages of ontological reasoning are to be made available in this growing number of cases.

In the past years, it has become clear that lightweight semantics are more feasible and realistic to be produced on a large scale. This does not only apply to ontology engineering but also other areas of ontology management, such as ontology matching and ontology re-use. A number of research challenges can be mentioned here:

Despite bits and pieces have been already done in the direction of **community-driven conceptual modelling**, a generic methodology is still missing. This challenge especially involves determining the degree of expressivity that is possible and how the collective intelligence of a community can be most efficiently channelled plus motivations for users to contribute. Another aspect is that many current approaches focus on the development but neglect the maintenance of ontologies.

Maintenance includes several different tasks that are related to ontology change [8]. The most important one is **ontology evolution**, which amounts to changing an ontology in response to a change in the modeled domain or its conceptualization; this may be necessary whenever new or previously unknown or classified information becomes available, or when modeling errors are discovered [8]. The growing diversity of people that are using ontologies is expected to increase both the need for keeping ontologies up-to-date, and the need to adapt them to the changing modeling needs of the users of an ontology. In addition, the increasing usage of ontologies in social contexts will require the development of easy-to-use tools for evolving ontologies by non-experts.

Evolving ontologies raise the need for several supporting tasks like **ontology versioning**, which is used to handle an evolving ontology by creating and managing different variants (versions) of it [9] and **change detection** which refers to the detection, identification and management of the differences (deltas) between subsequent versions of ontologies [10]. Furthermore, models that were designed by a large group are likely to contain modelling errors. Another challenge is to support the users by automatically detecting and repairing such modelling errors. One can identify patterns of common modelling mistakes that allow identifying and solving them. **Refactoring of semantic models** is different with lightweight semantics as logical inconsistencies are easier to track than, e.g., a wrong subClassOf relationship (flour and egg are no sub-classes of cake). The problem of refactoring semantic models is commonly referred to as the research field of **ontology debugging** and consists of **ontology diagnosis** (identifying inconsistencies and other modeling errors) and **ontology repair** (repairing such modeling errors) [8].

Discovering design patterns in ontology building can complement existing work on ontology engineering methodologies. Many methodologies do not go into much detail about the actual modelling part. **Design patterns** can be then re-used to ease the task of conceptual modelling.

Work in ontology matching focuses on the most relevant set-theoretic relations, such as disjointness, equivalence, more general or more specific. However, often denoted as equivalence, a subsumption relationship among two concepts lies often somewhere between "same as" and "not related". SKOS takes a first step in the direction of **"softer" mapping relations**. In order to make meaningful statements about the relation of two ontologies, more detail in mapping is required as well as sufficient automation support.

Ontology visualization is central for understanding conceptual models: so far, mostly graph-based approaches have been investigated. However, for a higher usability and openness to a large community, more sophisticated approaches to ontology visualization are required. Work-integrated lightweight semantics as mentioned previously, the creation of semantics should not be completely de-coupled from (1) work and (2) later re-use. Tagging gives a good example for being both work-integrated and providing immediate benefit.

It still remains to investigate how much expressivity is required and how much is possible (on a large scale) when creating ontologies. How lightweight can, in fact, semantics get in order to be still useful? These findings must be captured in underlying models that should build on the foundations provided by OWL, RDF, or SKOS.

4.1.4 Reasoning

Reasoning refers to the means to infer new knowledge from the basis of instance semantic data, ontologies and a formal logic model on which those ontologies are based (i.e. base axioms). Reasoners form a critical part of the semantic application architecture, whenever inference or semantic validation is required (as opposed to simply using RDF as a flexible data model). There are a number of benchmarks used to measure the limits of semantic stores to reason completely and correctly over a data set (i.e. determine all inferable knowledge on the basis of the contained knowledge, while not determining falsely inferable knowledge which should not be inferable on the basis of the contained knowledge).

Reasoning over finitely large data sets in controlled environments (e.g. on an enterprise network) is mature enough for commercial application today (using tractable logic fragments such as OWL-DL). The current proven maximal capacity of semantic data stores to allow querying over data with inference is measured and reported at the LargeTripleStores page of the W3C wiki[7]. At the time of writing, stores were reporting having achieved capacity of 12-15 billion triples, while OWLIM was able to materialise (infer) an additional 8 billion triples in the store, thus resulting in an effective total of 20 billion queryable triples. Several efforts are in progress to store 100 billion triples. Much progress is based on increasing the hardware capability on which the semantic data store runs. It is clear that the rate of progress in semantic data store capacity still needs to keep up with the potential amounts of semantic data to be published in the future, which may reach into the trillions (to start with).

Large scale distributed reasoning tasks (the ultimate example of which being reasoning over the Web) currently face much stronger scalability limits, since one can not simply increase memory or processor power on a hosting machine, while offering the possibility, if solved, to overcome this physical limitation of reasoning performance by shifting reasoning into the "cloud" of computing resources. The limitations of distributed reasoning can be gradually overcome by relaxing the inference guarantees of completeness and correctness, as well as permitting greater approximation and fuzziness in the calculation [3].

As such, it can be expected that current scalability limits for reasoning will continue to improve, however in view of the scales of data expected to be produced in the future, it seems that reasoning-based handling of structured, semantic data may possibly be necessarily always needing to be applied to subsets of entire data sets, so that the correct (and potentially on the fly) division of data into smaller, self-contained, datasets may prove to be the most important aspect of semantic data processing.

Conclusion

PlanetData includes Semantic Technology in different work packages, which depend on different key trends mentioned above. Starting from the first work package, PlanetData will develop novel ontology-based techniques for data integration. In the second work package, reasoning will be introduced to further improve the usefulness of the data published on the Web. The further use of reasoning will continue in the third and fifth work packages, Annotation is addressed in work package four to reference vocabularies.

[7] http://www.w3.org/wiki/LargeTripleStores (last seen February 28, 2011)

Table 1. Overview of the Vision & Solution of Semantic Technology

Vision	Solution	Short, medium or long term	Low, medium or high criticality
Large scale annotation of non-text	**Improved media processing Crowd sourced annotation**	**Medium to long**	**High**
Large scale knowledge extraction towards real time	**More efficient semantic data processing**	**Medium to long**	**High**
Large scale ontology creation	**Ontology tools and best practises More automation (extraction, modularisation,**	**Short to medium**	**Medium**
Large scale reasoning	**Increasing use of distributed components Loosening completeness and correctness guarantees**	**Short to medium**	**High**

4.2 Trends in linked data management

In recent years, a set of best practices for publishing structured data on the Web has emerged under the name "Linked Data". The basic principles that this term refers to are summarized in [8] as:

- "Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs, so that they can discover more things."

The aim of these guidelines is to facilitate the creation of a Web of Data, by encouraging the publication of information in a flexible structured representation such as RDF, and the use of established Web standards such as URIs and HTTP to allow the easy interlinking of this data. While using specifications from the Semantic Web effort such as RDF and SPARQL, Linked Data deliberately avoids the focus on ontologies and use of reasoning (while this can be built over Linked Data, as a RDF model). Rather, it focuses through its simplicity on encouraging the eased publication of data in an accessible, structured form with basic aspects of semantic models (typing of resources i.e. class-instance relation, resources have properties whose values are other resources, interoperability increased by the use of shared vocabularies).

The Linked Data principles and the vision of a global Web of Data are being adopted by a growing number of projects and organisations. The DBpedia project[9], for example, extracts structured information from Wikipedia and publishes it as Linked Open Data, thus feeding a vast amount of information compiled by the Wikipedia community into the Web of Data. Other information sources include governments (e.g. data.gov.uk or data.gov), researchers (e.g. LinkedLifeData.com), enterprises (e.g. LinkedOpenCommerce.com), and individuals (e.g. Revyu.com). To structure this information and to create links inside and between datasets, a number of popular vocabularies have emerged, for example SKOS to capture taxonomic relations between resources, FOAF to describe and link persons and their social networks,

[8] Tim Berners-Lee (2006-07-27). "Linked Data - Design Issues". W3C. http://www.w3.org/DesignIssues/LinkedData.html
[9] http://dbpedia.org/

or Good Relations to represent commerce information. The number of datasets and links between them is growing, as visualised in the different versions of the Linking Open Data cloud diagram[10], and it will continue to grow.

There are still open issues and research challenges related to Linked Data[11], which are detailed here:

Data integration: as Linked Data is published in a decentralised manner, anyone can publish Linked Data about anything and different people can publish Linked Data about the same thing. However, in the concept space, humans tend to associate together information about the same thing, regardless of its separate publication. Data integration is the process of merging multiple data items representing the same subject into a single, consistent, and clean representation. This process is usually conducted in three main tasks: schema mapping, identity resolution and data fusion. Schema mapping aims at normalizing vocabularies in order to increase homogeneity of data description. Identity resolution detects multiple descriptions of the same real world objects. Data fusion focuses on the resolution of data conflicts, where different sources provide different values for the same property of a resource. Conflict resolution may rely on trust and quality-related meta-information to allow machines to determine how to resolve inconsistencies. First, there is currently a scarceness of such trust and quality descriptors. Work on capturing and propagating provenance information about Linked Data is needed. Second, it is unclear if data should be cleaned and republished[11], or if integration tasks should generate qualifiers that would be stored so that quality can be assessed at query time. This becomes specially challenging in Web scale, requiring distributed solutions (e.g. based on cloud computing).

User interfaces and interaction paradigms: Linked Data browsers are currently still very based on a RDF-centric view of the Linked Data space, i.e. pretty printing RDF data. A more resource-centric view must also visualise and guide users in moving along different facets of the resource and across different data sets, which are interlinked. As the scale of Linked Data increases, and browsers can find and integrate relevant data across the Web of Data, even for a single resource as a focus of a browser, the amount of statements found about that resource would need solutions in terms of filtering and visualisation. When the focus of the browser is more than a single resource, i.e. a set of resources fulfilling a certain condition, the scale of answers to be visualised may number into the thousands, even the millions. Another issue that Linked Data should better address in the user interface would be the multilinguality of associated content.

Trust and quality assessment: A significant consideration for Linked Data applications is how to ensure the data most relevant or appropriate to the user's needs is identified and made available. For example, in scenarios where data quality and trustworthiness are paramount, how can this be determined (or heuristically assessed), particularly where the data set may not have been encountered previously? A sort of PageRank for the world of Linked Data is necessary, however such algorithms will need to be adapted to the linkage patterns that emerge on the Web of Data.

Access control: Applications that consume data from the Web must be able to access explicit specifications of the terms under which data can be reused and republished. Availability of appropriate frameworks for publishing such specifications is an essential requirement in encouraging data owners to participate in the Web of Data, and in providing assurances to data consumers that they are not infringing the rights of others by using data in a certain way. For example, to ensure the publication of public data by organisations, it is key to follow PSI directives from the EU and individual countries.

Privacy: The ultimate goal of Linked Data is to be able to use the Web like a single global database. The realization of this vision would provide benefits in many areas but will also aggravate dangers in others. One problematic area is the opportunities to violate privacy that arise from integrating data from distinct sources. Protecting privacy in the Linked Data context is likely to require a combination of technical and legal means together with a higher awareness of the users about what data to provide in which context.

Nevertheless, Linked Data can be and is already being used to build applications. There are specialized search engines for the Web of Data, e.g. Sig.ma[12] or Falcons[13], and classical search engines like Google begin to use structured data to enhance their search results[14]. There are also domain-specific applications and

[10] Richard Cyganiak and Anja Jentzsch (September 2010). The Linking Open Data cloud diagram. http://lod-cloud.net
[11] http://lod2.eu/Deliverable/D4.3.1.html
[12] http://sig.ma/
[13] http://iws.seu.edu.cn/services/falcons/documentsearch/
[14] http://googlewebmastercentral.blogspot.com/2009/10/help-us-make-web-better-update-on-rich.html

mash-ups that make use of Linked Data, for example using government data from data.gov.uk[15]. With the amount of data growing and the technology behind it maturing, provided the above research challenges are adequately addressed, Linked Data will be used increasingly to build innovative applications for customers and to improve agility inside enterprises[12].

Conclusion

Linked data management is included in the PlanetData work package four. PlanetData aims that the data will be offered in a self-descriptive manner easing the automated discovery and usage by machines. The data itself will be made available according to the Linked Data principles wherever possible, and made accessible via the Web site.

Table 2. Overview of the Vision & Solution of Linked Data Management

Vision	Solution	Short, medium or long term	Low, medium or high criticality
On the fly homogeneous view on a resource	**Dynamic data integration** **Contextualisation**	**Medium to long**	**Medium**
Exploring easily Web of Linked Data	**UI/visualisation** **Ranking/weighting**	**Short to medium**	**High**
Trust and quality embedded in Web of Data	**Provenance tracking and explanations**	**Medium**	**Medium**
Control mechanisms for (re-)use of data and privacy protection	**Access control models for Linked Data** **Data licensing**	**Medium**	**High**
Processing large scales of Linked Data efficiently	**Linked Data scalable reasoning / querying**	**Medium**	**High**

4.3 Trends in databases

As discussed in the previous section, semantics in general and Linked Data in particular will play an increasingly important role in the Web of Data. Therefore, an important trend in database research is the development of repositories optimised for the storage of semantic data. This data is represented as triples of subject, predicate and object, sometimes extended by a fourth component to store context (e.g. to implement named graphs[13]). The SemData initiative is bringing together researchers and developers from the semantic and database communities in workshops and round tables to discuss "issues such as semantic repositories, their virtualization and distribution, and interoperability with relational solutions, XML and others"[16]. In its mission statement[17], the initiative describes the benefits of using semantics in data management as:

- "more efficient, faster and cheaper data integration
- data access, regardless the physical location and representation
- more comprehensive data analysis and interlinking

[15] http://data.gov.uk/apps
[16] http://semdata.org/
[17] http://semdata.org/about#mission

- flexible querying against multi-schema datasets
- interlinking of text and Web content with structured data
- hybrid information retrieval techniques
- easier data exploration and understanding."

To enable the use of semantic technologies in large scale data management, a lot of the discussions in the SemData workshops and round tables revolve around the challenge of getting the performance of semantic repositories on par with established database approaches like relational databases. Work is being done to **adapt techniques from the database community to semantic repositories**, like the use of column-stores in MonetDB[14] or Virtuoso[15]. Another approach is the combination of vertical partitioning and multiple indexing for RDF data in Hexastore[16].

Distributed architectures can improve the scalability of semantic repositories regarding data storage or query processing, e.g. OWLIM's replication clusters for load balancing of concurrent requests[17]. Leveraging different kinds of hardware can also boost performance, like using main memory instead of persistent storage for fast random access to data, or using hardware for parallel execution of data processing like multi-core CPUs or even GPUs.

Regarding **querying**, there is work being done on optimizing the handling of specific characteristics of Linked Data, e.g. the large number of identity links using the owl:sameAs property[18]. The query language commonly used for semantic repositories is SPARQL[19], but there are also efforts to extend this language with features for efficiently querying special kinds of data, like spatio-temporal[18][19] and stream data[20]. Also, for distributed databases, solutions to distributed (semantic) querying are sought.

In order to handle the dynamic changes in the stored data and the load and structure of queries against that data, databases need to become more **self-adaptive**. To achieve consistently good performance over both small and large data, work is performed on **elastic databases**, which already in the relational database world represents a new approach, and hence is further away from realisation within semantic stores.

Semantic databases will keep maturing in terms of performance and features, and can be expected to enter the mainstream in the coming years. To help users integrate existing data from relational databases in semantic repositories, tools like R2RML, which is ODEMapster[20] or D2R Server[21] exist. Combined, these technologies will allow users to benefit from rich, semantic data models and the possibility of integrating data from heterogeneous sources on the Internet and on intranets.

Conclusion

Databases will be addressed in the work packages one and two of PlanetData. In PlanetData, we envision using the open-source database management system MonetDB, which provides a software architecture stack, including libraries for event stream processing, where solutions can be experimented with in a controlled lab setting.

Table 3. Overview of the Vision & Solution of Databases

Vision	Solution	Short, medium or long term	Low, medium or high criticality
Semantic stores on par with RDBMS	**Triple storage optimisations**	**Short**	**High**

[18] http://www.ontotext.com/factforge/inference.html
[19] SPARQL Query Language for RDF. W3C Recommendation 15 January 2008. http://www.w3.org/TR/rdf-sparql-query/
[20] http://neon-toolkit.org/wiki/ODEMapster
[21] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

Semantic storage able to deal with Web data (scale and heterogeneity)	**Federation or distribution of the store Data wrappers and**	**Short**	**Medium**
Semantic storage on par with large scale query load	**SPARQL optimisations**	**Medium**	**High**
Real time adaptation to query and data loads	**Self adaptation in databases**	**Medium**	**High**
Achieve good performance over large and small data	**Elasticity of cloud storage**	**Long**	**Medium**

4.4 Trends in data markets

The growing amount of data on the Web can be a valuable tool for application developers, businesses, policy makers, researchers, journalists or other interested parties for building services or gaining new insights. For this to work, on the one hand, consumers must be able to find and use the datasets they need. On the other hand, datasets must be made available by potential suppliers, which might be prevented if suppliers don't know how or where to publish, or if they don't see benefit in doing it.

Data market places are data exchange facilities that aim to bring consumers and suppliers of data closer together. By collecting datasets and advertising them on their website, they can improve the visibility of datasets and make them easier to find for consumers. For suppliers, they can give support in the publishing process. Some data market places also allow suppliers to sell their datasets in exchange for money, which can provide incentives for publishing datasets that might otherwise not have been made available.

Infochimps.com, for example, is a website where users can share datasets from any domain freely or in exchange for money. DataMarket.com offers datasets specifically about "Icelandic economy, society, nature, and industry". Timetric focuses on economic data "from sources including The World Bank, Eurostat and the Office for National Statistics". Microsoft's Windows Azure MarketPlace DataMarket aims to provide a marketplace for "trusted commercial and premium public domain data"[22].

These data market places differ in several properties[23].

Firstly, not all data markets are domain-neutral, but some focus on specific domains. Timetric focuses on economic data, and DataMarket.com originally only offered data about Iceland, and still has a dedicated sub-site for datasets on Iceland[24]. It is difficult at the moment to know which data market offers which types of data to make it easier to identify where to go and what can be found there. Data markets could learn from the principles of (Web) services, in which the publication of the service description on the Web is an important aspect of allowing agents to find out what functionality is available and how to access it. Data markets, as a form of data service, could use a **data market description model** to advertise their data offers.

Second, there are different pricing and business models. Azure DataMarket allows data publishers to offer flatrate subscription models or payment according to the number of transactions used by customers. DataMarket.com and Infochimps offer APIs for programmatic data access and bill their users by usage volume. Infochimps additionally offers dataset publishers to sell their datasets for fixed amounts of money. Timetric doesn't have a pricing model in effect at the time of this writing, but plans to offer premium datasets to paying customers. Again, as the number of data markets online increase and a manual overview of the different offers becomes increasingly impossible, it would be desirable to have data market descriptions, which identify the used pricing, and business model. Since the sustainability of data markets are predicated on the value of their data, monetarized through their business model, being higher than the costs

[22] https://datamarket.azure.com/
[23] http://www.slideshare.net/marin_dimitrov/linked-data-marketplaces
[24] http://iceland.datamarket.com/

of acquiring and maintaining that data. So both **surveys on the economics of data and research into deriving greater value from data** are very important to drive the nascent data markets market.

A third difference between currently available data markets lies in the contribution models. Microsoft's Azure DataMarket and DataMarket.com are not open for community contributions, while the other sites allow all users to upload datasets. On the one hand, selection and quality control performed by a trusted third party can add value to datasets. On the other hand, the success of projects like Wikipedia has shown that websites with an open contribution model have the potential to attract large user bases. Despite this trade-off between openness and trustworthiness, there can be reasonable use cases for both kinds of data markets. While community-built data markets might appeal to end users with an immediate information need, consumers requiring high-accuracy information for business or policy decisions might prefer selected datasets from trusted data brokers. Data markets will need to ensure **trust and quality of shared data** in either case, whether crowd-sourced control (a la Wikipedia) or via expert quality control.

A fourth issue with the range of existing data markets, however, can make it harder for consumers to make use of the available datasets. Today's data markets often employ custom data models and access APIs that demand development effort from consumers to integrate datasets into their applications. **Common data models and access interfaces for data markets**, ideally based on open standards like the Linked Data principles described in Section 4.2, can improve interoperability between data market places and make them even easier to use.

Although there will inevitably be some consolidation among data market places, the fundraising success of start-ups like Infochimps shows their economic potential. By making it easier for data suppliers to publish their datasets, and by offering "data-as-a-service" to consumers, data market places have the potential to form the basis of a growing data economy.

Conclusion

In PlanetData we are creating a directory of open datasets, which will form part of the PlanetData catalogue. The catalogue will contain different types of data sets from diverse domains. The trust and quality of the shared data will be ensured as well in the work package four together with the data catalogue provisioning. Work package five includes the common data model and access interface.

Table 4. Overview of the Vision & Solution of Data Markets

Vision	Solution	Short, medium or long term	Low, medium or high criticality
Finding data markets easily on the Web	**Data market descriptions**	**Medium**	**High**
A „free data market" with data licensing and pricing models	**Data market economic theories**	**Long**	**High**
Consensus around ‚good' data sets (no data market spam)	**Data quality via analysis and crowdsourcing**	**Medium**	**High**
Fully interoperable data markets	**Agreed APIs and data models**	**Long**	**Medium**

4.5 Trends in data stream management

The combination of sensor networks with the Web, web services and database technologies, was named some years ago as the Sensor Web or the Sensor Internet.

Urban computing is one of the technologies that are producing and using large-scale data streams [4]. It refers to the increasingly ubiquity of Internet and devices connected to the Internet in the urban environment

leading to smart cities. Current technology lacks the capability to effectively solve urban computing problems, as it requires combining a huge amount of static knowledge about the city with larger, real time data being generated by sensors and devices in a heterogeneous and noisy manner. To act upon the data acquired, its combination is insufficient; rather there must be intelligent reasoning to draw in-time inferences.

Current projects provide test solutions in more controlled and controllable environments, but the technological challenge remains to satisfactorily deal with large scale, heterogeneous and "dirty" data. In order to tackle challenges of processing the resulting streams of data, some research groups focus on the combination of semantic technology with the Sensor Web, which becomes known as the Semantic Sensor Web [5], yet – in line with our expectations for large scale semantics outlined in section 4.1 - mature solutions for true, open urban environments may be expected first in the middle to long term.

Current challenges for data stream processing include the need to handle dynamic data (e.g. from sensors), extracting actionable knowledge (data mining), processing in real time, and handling data, which is, not clean, consistent or correct (probabilistic approaches). The most relevant challenges in the Sensor Web have been reviewed and, in this case, semantic technology has been proposed as the most promising solution [6].

Capturing the right **abstraction level** of the captured data is important for how the data can subsequently be processed. Typically, data is captured in streams at a low level, being very specific to the device and network. However, if this data is to be re-usable in other contexts, a higher-level abstraction needs to be provided over the data, which abstracts from syntactic and infrastructural heterogeneities. Some sensor network ontologies are being specified, also within PlanetData, to enable such high-level data abstractions.

Dynamic data handling refers to coping with time-dependant data. Noisy, uncertain and inconsistent data is an unavoidable aspect of future data streams, for example traffic data. Different sensors observing the same road may give apparently inconsistent information. Moreover, a single data coming from a sensor in a given moment may have no certain meaning. In the urban computing case, for example, the number of cars passing a traffic detector can vary very much over time. This means that the system must have a notion of "observation period", meaning an interval for which the system can be subject to querying, and within a given observation period, the system must be able to capture invariable knowledge, invariable data, periodically changing data or event driven changing data.

As a result, means for determining and expressing the **quality and provenance of data in a stream** should be provided at the stage of data capture ideally, since subsequent analysis of the data to determine quality and provenance may no longer be able to determine important characteristics for the correct determination. The accuracy of data captured at any instant can be affected by various internal and external conditions on the sensor network. Hence, quality criteria must not only be identified but those criteria measured at the point of data capture. Likewise, **data stream quality of service** is an important factor for systems seeking to make use of data from that stream and a correct QoS measurement must take into account the various factors. For example, the unavailability of a piece of data for a period of time can have different meanings, such as the sensor was not available or that there was no event to trigger data capture at that moment, which are difficult or impossible to know separately from the data capture moment.

Stream reasoning[25] or **complex event processing** are emerging research areas to extract knowledge from data streams and infer further knowledge from it while handling the obvious issues of temporality of and variance in the data being used. Its ultimate goal is to be able to provide the results of such inference in real time, since many data stream reasoning use cases require rapid action on the knowledge that can be derived from a real time stream (e.g. medical cases).

An aspect of this is **integration and fusion of data.** Sensor networks are autonomously deployed, heterogeneous in data structure and semantics, and exhibit different quality characteristics. Their data may be integrated not only with data from other networks but also with persisted data from other sources such as static data and archived data. A recent research trend is focused on the generation of Linked Data from sensor network data streams [7] by means of transforming sensor-based data into RDF and making it available using HTTP by means of sensor-related URIs. This will allow the seamless navigation across sensor-based (and other types of) data. To execute such integration and fusion tasks, the semantics of

[25] http://streamreasoning.org/

continuous query languages such as C-SPARQL and for joins between continuous and stored data from different sources require still more research work.

The **identification and location of relevant sensor-based data sources** becomes a challenge as the number of sensor data streams over the Web increase, and is very important for automating the data fusion and integration tasks. Sensor data registries are required which ideally use open Web standards.

Finally, **the rapid development of data stream applications** becomes key for future uptake of the usage of data streams in IT systems, taking into account the mentioned challenges and finding research solutions. There is a need for common interfaces and formats between applications, sensor networks and other components, and that between the different architectural parts of a data stream application, the different levels of data abstraction and varying qualities of services can be handled.

Conclusion

Data streams are one of the main research challenges of PlanetData project. Starting in work package one, PlanetData aims to improve the storage and data management infrastructure in the context of processing data streams. Furthermore, PlanetData researches the mining of streaming sources for publishing such data in more structured manners. Finally, work will be conducted on annotating streaming data sources in order to facilitate the processing, mining, and fusion of such resources.

Table 5. Overview of the Vision & Solution of Data Stream Management

Vision	Solution	Short, medium or long term	Low, medium or high criticality
Identification and location of streams	**Stream directories**	**Medium**	**High**
Abstraction level	**Semantic query languages, SSN ontology**	**Short**	**High**
Handling time dependent data	**Continuous query languages**	**Short**	**High**
Stream reasoning/complex event processing	**cSPARQL**	**Medium**	**Medium**
Data integration and fusion	**cSPARQL**	**Medium to Long**	**Medium**
Data quality and provenance	**Stream analysis and metadata**	**Medium**	**High**
Data stream QoS	**GSN**	**Medium to Long**	**High**
Rapid application development	**Stream data APIs**	**Short**	**Medium**

5 Conclusion

The purpose of this roadmap has been to consider key trends in different domains of IT, which the PlanetData Network of Excellence has identified as key areas for overcoming future challenges in large-scale data management. From these trends, we have identified the research challenges which are open for being solved so that these technologies will fulfil our vision of enabling future data management at the scales of data that can already be foreseen, but barely imagined. Such research solutions are vital if future IT will be able to provide systems that meet the needs of both the public and private sector in producing, publishing, consuming and processing huge scales of data in the (near) future.

We have interviewed experts in each area and drawn up a list of visions for enabling large-scale data management via this research topic in the future. Where possible, we propose possible or emerging solutions (or their sources) – it is also interesting to note the presence of research challenges where a clear activity towards finding a solution could not be identified by the expert. We also suggested how soon we could imagine a solution being found and implemented, and how critical the implementation of the solution would be for enabling large-scale data management in the context of this IT topic. The results have been given in the respective conclusions in chapter 4, and also plotted below onto a single diagram of 'time to solution' against 'criticality of solution', please refer to Figure 2

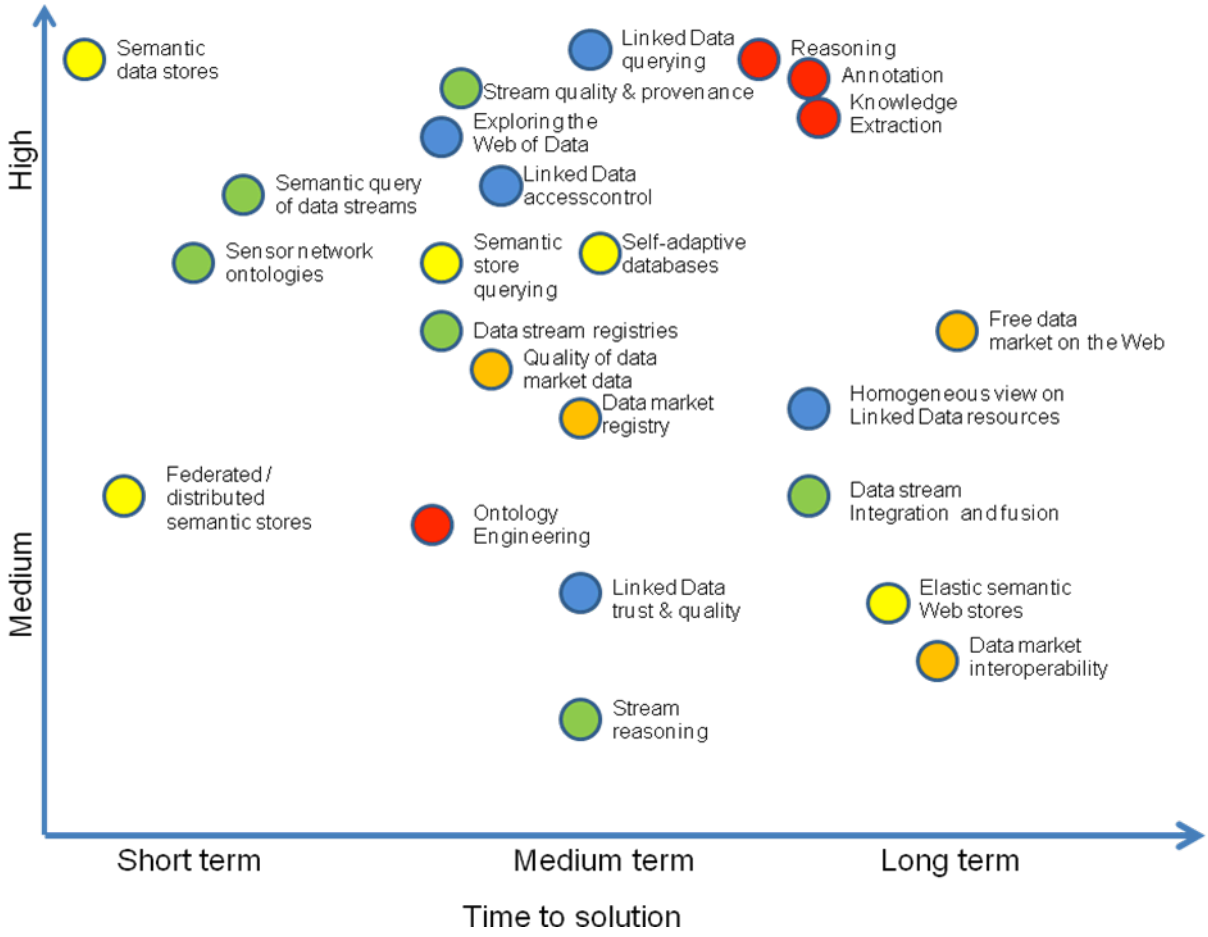

Figure 2. Conclusion of the Vision and Solution in a Diagram

From this, we can note that semantic data storage is close to scalable solutions both in single and federated/distributed stores, so the challenge is truly shifting to the Web as "global store of semantic data" to be found and processed. Data streams are also becoming more (re-) usable in terms of their modelling at a higher abstraction level via ontologies and being queried semantically to extract actionable knowledge. Hence it can be expected that such "single" source solutions will be semanticised and scale in the next years, but when data is distributed across the Internet, or the Web, in various heterogeneous forms, we face new challenges in working on such scales of data, as is anticipated by PlanetData. Possibly, earlier solutions in

those database and data stream fields can help to drive the discovery of final solutions for the more challenging domains of (Web scale) semantics, Linked Data and data markets. In the medium term, the Web of Data – the mass of Linked Data published to the Web – will become more easily explorable, with aspects like access control and querying making publication of Linked Data and subsequent re-use more technologically feasible. Since Linked Data benefits from less complexity in processing the data, avoiding the trade offs of large scale reasoning, solutions for handling Linked Data at Web scale will emerge before those for the Semantic Web, where ontologies are combined with RDF data to permit additionally inference over the data. Reaching this critical moment where semantic data and ontologies can be combined at the scale of the Web will also be driven by advances in semantic stores, where querying with reasoning will be more scalable and approaches like self-adaptation will ensure that Web based semantic stores can better react to the dynamicism of activity on the Web, such as rapidly shifting query loads. In the longer term, the most critical research challenges will be the fulfilment of the Semantic Web vision – while ontology engineering has progressed to a mature science, Web scale reasoning, annotation of media and knowledge extraction may still be the last challenges to face in this decade.

Within PlanetData, we will use this Roadmap as an initial motivation for further research in the identified areas. This will not only be a focus for the consortium members and also the associate members, who may provide new competencies towards solving open research challenges, but also a focus for the future Planet Data Programs.

References

[1] "The Service Web 3.0 Roadmap on Semantics in Services", Lyndon Nixon, Graham Hench and Elena Simperl. Available from www.serviceweb30.eu

[2] "Future Trends", Lyndon Nixon, Raphael Volz, Fabio Ciravegna and Rudi Studer. In "Handbook of Semantic Web Technologies", Springer, 2011.

[3] Atanas Kirkyakov, "Measurable Targets for Scalable Reasoning", LarKC deliverable D5.5.1, June 2008.

[4] "Challenging the Internet of the Future with Urban Computing", Emanuele Della Valle, The OneSpace Workshop at Future Internet Symposium (FIS) 2008, September 2008.

http://emanueledellavalle.org/blog/2008/09/28/challenging-the-internet-of-the-future-with-urban-computing

[5] Amit Sheth, Cory Henson, and Satya Sahoo, "Semantic Sensor Web," IEEE Internet Computing, July/August 2008, p.78-83.

[6] Corcho, Óscar and Garcia Castro, Raul (2010) Five Challenges for the Semantic Sensor Web. Semantic Web - Interoperability, Usability, Applicability, 1 (1-2). pp. 121-125. ISSN 1570-0844

[7] J. Sequeda, O. Corcho, Linked Stream Data: A Position Paper. Proceedings of the 2nd Int. Workshop on Semantic Sensor Networks (SSN09), Washington DC, USA, October 26, 2009

[8] Giorgos Flouris, Dimitris Manakanatas, Haridimos Kondylakis, Dimitris Plexousakis, Grigoris Antoniou. Ontology Change: Classification and Survey. Knowledge Engineering Review (KER), 23(2), pages 117-152, 2008.

[9] M. Klein and D. Fensel. Ontology versioning on the semantic web. In Proceedings of the International Semantic Web Working Symposium (SWWS), pages 75-91, 2001.

[10] Vicky Papavassiliou, Giorgos Flouris, Irini Fundulaki, Dimitris Kotzinos, Vassilis Christophides. On Detecting High-Level Changes in RDF/S KBs. In Proceedings of the 8th International Semantic Web Conference (ISWC-09), 2009.

[11] Tom Heath and Christian Bizer (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

[12] Dean Allemang. Semantic Web and the Linked Data Enterprise. In: David Wood (ed.), Linking Enterprise Data, Springer, 2010.

[13] Andreas Harth, Stefan Decker. Optimized Index Structures for Querying RDF from the Web. 3rd Latin American Web Congress, Buenos Aires, Argentina, October 31 to November 2, 2005, pp. 71-80.

[14] Stefan Manegold, Martin L. Kersten, Peter Boncz. Database Architecture Evolution: Mammals Flourished long before Dinosaurs became Extinct. VLDB '09, August 24-28, 2009, Lyon, France.

[15] Orri Erling. Directions and Challenges for Semdata. Workshop on Semantic Data Management (SemData@VLDB) 2010, September 17, 2010, Singapore.

[16] Weiss, C., Karras, P., and Bernstein, A. Hexastore: Sextuple Indexing for Semantic Web Data Management. VLDB '08, August 23-28, 2008, Auckland, New Zealand.

[17] Atanas Kiryakov, Barry Bishop, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, Ruslan Velkov. The Features of BigOWLIM that Enabled the BBC's World Cup Website. Workshop on Semantic Data Management SemData@VLDB 2010 September 17, 2010, Singapore.

[18] Jonas Tappolet, Abraham Bernstein. Applied Temporal RDF: Efficient Temporal Querying of RDF Data with SPARQL. In: L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, pp. 308–322, 2009.

[19] Manolis Koubarakis, Kostis Kyzirakos. Modeling and Querying Metadata in the Semantic Sensor Web: the Model stRDF and the Query Language stSPARQL. The Semantic Web: Research and

Applications. Lecture Notes in Computer Science, 2010, Volume 6088/2010, 425-439, DOI: 10.1007/978-3-642-13486-9_29

[20] Davide Francesco Barbieri, Daniele Braga, Stefano Ceri and Emanuele Della Valle and Michael Grossniklaus, Continuous Queries and Real-time Analysis of Social Semantic Data with C-SPARQL, in SDoW 2009 Colocated with ISWC 2009.