Report from Dagstuhl Seminar 17262

Federated Semantic Data Management

Edited by

Olaf Hartig[1], Maria-Esther Vidal[2], and Johann-Christoph Freytag[3]

1 **Linköping University, SE,** olaf.hartig@liu.se
2 **Universidad S. Bolivar, Caracas, VE,** mvidal@usb.ve
3 **Humboldt-Universität zu Berlin, DE,** freytag@informatik.hu-berlin.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17262 "Federated Semantic Data Management" (FSDM). The purpose of the seminar was to gather experts from the Semantic Web and Database communities, together with experts from application areas, to discuss in-depth open issues that have impeded FSDM approaches to be used on a large scale. The discussions were centered around the following four themes, each of which was the focus of a separate working group: i) graph data models, ii) federated query processing, iii) access control and privacy, and iv) use cases and applications. The main outcome of the seminar is a deeper understanding of the state of the art and of the open challenges of FSDM.

Seminar June 25–30, 2017 – http://www.dagstuhl.de/17262
1998 ACM Subject Classification H.1 Models and Principles, H.2 Database Management, H.3 Information Storage and Retrieval
Keywords and phrases Linked Data, Query Processing, RDF, SPARQL
Digital Object Identifier 10.4230/DagRep.7.6.135

1 Executive Summary

Olaf Hartig
Maria-Esther Vidal
Johann-Christoph Freytag

The Semantic Web is an extension of the World Wide Web in which *structured data and its meaning* is represented in a form that can be readily accessed and exploited by machines. The foundation of this representation is a graph-based data model defined by the Resource Description Framework (RDF). This framework allows for data management approaches that focus on manipulating and using data in terms of its meaning. We refer to this type of data management as *semantic data management*.

In addition to centralized access to RDF datasets, Web-based protocols such as the SPARQL protocol enable software clients to access or to query RDF datasets made available by remote servers. By integrating such remote data sources as members of a *federated system*, software clients may answer cross-dataset queries without having to retrieve various datasets into a single repository. Given such a federation, the complexity of problems of query processing and semantic data management increases due to additional parameters such as variable data transfer delays, a changing availability of federation members, the size of the federation, and distribution criteria followed to place and semantically link data in different datasets of the federation. Moreover, whenever data is replicated across federations,

Federated Semantic Data Management, *Dagstuhl Reports*, Vol. 7, Issue 06, pp. 135–167
Editors: Olaf Hartig, Maria-Esther Vidal, and Johann-Christoph Freytag
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

synchronization is required to ensure that all changes are propagated and the semantics of data is preserved. Despite a large number of technologies developed by the Semantic Web and Database communities to address problems of semantic data management, we still observe a significant lack of efficient and effective solutions to the problems of federated semantic data management (FSDM), which prevents the development of real-world applications on top of Semantic Web technologies. Additionally, existing proposals to evaluate such solutions do not sufficiently cover the large number of parameters that affect FSDM and the complexity of tradeoffs. More specifically, variables and configurations that considerably affect the federated semantic data management problems are not sufficiently defined or even considered in state-of-the-art testbeds (e.g., network latency, data fragmentation and replication, query properties, or frequency of updates).

The aim of the Dagstuhl seminar was to gather experts from the Semantic Web and Database communities, together with experts from application areas, to discuss in-depth open issues that have impeded FSDM approaches to be used on a large scale.

The following crucial questions were posed as a basis for the discussions during the seminar:

Q1 *Can traditional techniques developed for federations of relational databases be enriched with RDF semantics, and thus provide effective and efficient solutions to problems of FSDM?*

Q2 *What problems of FSDM present new research challenges that require the definition of novel techniques?*

Q3 *What is the role of RDF semantics in the definition of the problems of FSDM?*

To discuss these questions the participants of the seminar were grouped according to their areas of expertise and interests. In particular, the seminar focused on four main topic areas (see below). The results of the group discussions were presented in plenary sessions and will be compiled into manuscripts with which the seminar outcomes will be disseminated. As a basis of the group work, and to establish a common understanding of key concepts and terminology, the seminar included a few short, survey-style talks on a number of related topics. In particular, these talks covered:

- "Graph data models and graph databases" (given by Olaf Hartig),
- "RDF and semantics" (by Claudio Gutierrez),
- "Policies and access control" (by Sabrina Kirrane and Piero Andrea Bonatti),
- "Database privacy" (by Johann-Christoph Freytag),
- "Distributed database systems" (by Katja Hose), and
- "Federated query processing" (by Maria-Esther Vidal).

In addition to these survey talks, every participant was given the chance to briefly highlight their research as relevant for the seminar. Moreover, in a demo session, some of the participants showcased their FSDM-related systems and tools, which gave interested attendees of this session an opportunity to play with and better understand these systems and tools. The systems and tools demonstrated in this session were the following:

- *Triple Pattern Fragments client* that runs in a browser and executes queries over a federation of Triple Pattern Fragment (TPF) interfaces (demonstrated by Joachim Van Herwegen),
- *Network of Linked Data Eddies* (*nLDE*), an efficient client-side SPARQL query engine for querying server-side data that can be accessed via a TPF interface (demonstrated by Maribel Acosta),
- *Ladda*, a framework for delegating TPF-based query executions among multiple browsers (demonstrated by Hala Skaf-Molli),

- *Quartz*, a system for querying replicated Triple Pattern Fragments (demonstrated by Hala Skaf-Molli),
- *Ontario*, a federated SPARQL query engine for heterogeneous sources represented in different raw formats (demonstrated by Maria-Esther Vidal),
- *UltraWrap*, a framework for integrating relational databases using SPARQL federation (demonstrated by Juan Sequeda),
- *Ephedra*, a SPARQL federation engine that combines SPARQL services with other services (demonstrated by Peter Haase),
- *JedAI*, an entity resolution toolkit (demonstrated by Themis Palpanas), and
- *Exemplar Queries*, a framework for query answering using knowledge graphs (demonstrated by Themis Palpanas).

As mentioned before, besides the short survey talks, the demos, and the participants' presentations, the major focus of the seminar was on discussions in four working groups, where each of these groups addressed a different topic area. The remainder of this section provides a brief overview of the four topic areas covered by the groups and the respective results. More detailed summaries provided by each of the four groups can be found in a separate section of this report.

Graph Data Models. Graph data models such as the RDF data model allow for a representation of both data and metadata using graphs of nodes that represent entities, and edges that model connections between entities. Graph data management encompasses techniques for managing, querying, and analyzing graph data by utilizing graph-oriented operations. SQL-like query languages have been defined for evaluating declarative queries over graph data; additionally, well-known algorithms are utilized for computing graph invariants (e.g., triangle counting or degree centrality) and for solving typical graph problems (e.g., finding shortest paths, traversals, or dense subgraphs). Furthermore, several real-world applications have been built on top of existing graph-based tools (e.g., community detection, centrality analysis, and link prediction). Graphs naturally represent a wide variety of domains (e.g., social networks, biological networks) in which data, interconnectivity, and data topology all are first-class citizens, with RDF data being one example of graph data.

During the Dagstuhl seminar, a working group was formed to discuss whether tools for graph data management are sufficient to model and to manage the semantics in RDF data, taking into account that characteristics of the RDF data model (e.g., blank nodes and SPARQL operators) may affect tractability of the graph-based tasks in a federation of RDF graphs. As a first result of this discussion, the working group made the following observation. In contrast to other graph data models and query languages, the RDF data model is a "universal" data model in the sense that it is designed for sharing data and knowledge in an unbounded space such as the Web. To continue the discussion, the group introduced a definition of the notion of FSDM and identified five principles that characterize FSDM: universality, unboundedness, dynamicity, network protocols, and semantics. Based on further discussion that took into account these principles, the group made two conjectures that they plan to elaborate on in a future publication and that can be summarized as follows. First, it is impossible to build a FSDM system that fully achieves universality, unboundedness, and dynamicity, all at the same time. Second, the concepts of federation and semantics are interdependent and must be tackled together to develop effective and efficient solutions for building FSDM systems.

Federated Query Processing. A vast number of approaches have been developed to provide a unified interface for querying federations of data sources. In the context of federations of

RDF datasets, existing approaches focus on two problems: the problem of selecting the RDF datasets required to execute a federated query, and the problem of executing the resulting sub-queries efficiently against the selected data sources. Although federated query processing has been studied extensively, a number of important problems are still open, and more challenges are likely to come up as the complexity of federations increases (e.g., by increasing numbers of federation members, by replication and fragmentation of RDF data, and by federation members that update their RDF data autonomously).

During the Dagstuhl seminar, a working group was formed to discuss the problem of federated query processing over RDF data sources. Challenges imposed by the semi-structured nature of RDF, unpredictable behavior and dynamicity of Web-accessible RDF sources, and the role of the entailment regimes guided the group discussions and allowed for enumerating the main differences with the problem of federated query processing against relational databases. The group focused on the formal definition of the problem, as well as on the formalization of the subproblems of source selection, query decomposition, and query execution. As a first result, the group identified that the entailment regimes to be performed over a federation of RDF sources, as well as data replication and dynamicity, access control policies, and SPARQL query capabilities, play a crucial role in source selection, query decomposition, and query execution. State-of-the-art techniques implemented by existing approaches (e.g., FedX, ANAPSID, or Linked Data Fragments) were discussed and compared based on this formalization; the group concluded that none of existing approaches takes into account all these characteristics of RDF data sources, being required further analysis and work to empower them to solve the formalized problems. Finally, the impact of these characteristics on the performance of SPARQL operators (e.g., join, union, or optional) was discussed. The group concluded that although physical operators implemented by existing approaches are capable of adjusting query execution schedulers to RDF source availability, they are unable to adapt their execution to other RDF source characteristics, e.g., supported entailment regimes or data evolution. These issues remain open as well, and require further study from the semantic data management community.

Access Control and Privacy. Solutions to the problem of modeling access control policies for Web resources have been benefited from Semantic Web technologies. Existing rule-based logic languages rely on ontology-based reasoning tasks to represent reactive policies for access control, and to enforce and to propagate trusted and policy-compliant interactions across resources in RDF datasets. For instance, the Open Digital Rights Language (ODRL) is a rule-based approach that allows for a description of policies to access and to exchange Web resources. Nevertheless, as per the Linked Data publishing principles, RDF properties associated with any resource can be accessed by de-referencing their corresponding URL. In applications of domains of FSDM such as personalized medicine or finances, only authorized and privacy-respecting access is allowed. Thus, novel approaches are required to bridge the gap between access-control models and unrestricted access to RDF resources.

A working group with a focus on access control and privacy discussed the following open issues: a) formalisms to specify access-control and privacy policies of federation resources and to reason over the meaning of these resources; and b) techniques that enable systems to enforce privacy-aware and security-aware policies whenever a resource is accessed. After concluding that there are too many open challenges to be all solved immediately, the group agreed to focus on access control. Next, the group discussed conceptual access control models and achieved a better understanding of requirements of a conceptual framework to analyze policy-aware federated Semantic Web architectures. Finally, the group defined such a framework and made plans for a publication about it.

Use Cases and Applications. In addition to the first three working groups that focused on various more technical aspects of building FSDM systems, a fourth working group looked into applications of FSDM and use cases in which adopting FSDM would be beneficial. Specifying such use cases, as well as documenting the usage of FSDM systems in existing applications, is important to better understand the requirements and the challenges of FSDM and to derive realistic testbeds for approaches to build FSDM systems.

A key observation of the work group was that approaches to apply FSDM can be categorized into two classes depending on whether they focus i) on explorative, open-domain querying or ii) on controlled, close-domain querying. Then, the working group identified a broad set of general use cases of FSDM. Thereafter, the group defined a framework for developing specific use cases. This framework introduces a set of requirements for the specification of a use case. Finally, the group applied their framework to develop a number of example use cases.

2 Table of Contents

3 Overview of Talks

This section contains brief summaries by all participants of their research related to the topic of the seminar. During the seminar, all participants gave a lightning talk to highlight their work. Additionally, some of this work has been presented at a demo session during the seminar, others has been the topic of discussions during the seminar.

3.1 Query Processing over Graph-structured Data on the Web

Maribel Acosta (KIT – Karlsruher Institut für Technologie, DE)

Linked Data initiatives have encouraged the publication of large datasets on the Web. As a result, a huge dataspace of graph data has emerged, where data is represented using the RDF data model and can be queried using the SPARQL language. Despite these developments, the Web-like characteristics of Linked Data sources pose fundamental challenges on the efficiency and effectiveness of query processing engines over autonomous Linked Data sources. To address these challenges, this thesis focuses on the definition of flexible query processing strategies over RDF graphs on the Web.

Regarding efficient query processing, the lack of statistics about the data and unpredictable data transfer delays can negatively impact the performance of engines that consume Linked Data. This problem is mainly generated because existing engines execute fixed plans following the traditional optimize-then-execute paradigm. To tackle this problem, this thesis presents an adaptive SPARQL engine tailored to execute queries against remote Linked Data sources. Our solution comprises query optimization techniques to devise effective plans. The plans are executed following an adaptive strategy to change execution schedulers according to current conditions and reduce query runtime. The results of our empirical studies indicate that our solution outperforms static query schedulers. Our results also provide novel insights about the tradeoffs of different adaptive strategies when evaluating selective and non-selective queries.

An orthogonal but equally important aspect of querying Linked Data is the quality of the retrieved data. Executing SPARQL queries against graphs with quality issues leads to low-quality results. To tackle this problem, we propose a novel hybrid engine that integrates humans into query processing to enhance the quality of SPARQL query answers. Our solution relies on the graph structure of RDF data to decide on-the-fly which parts of a query should be crowdsourced. Experimental results show that our engine is able to enhance the completeness of SPARQL queries.

3.2 SPARQL Query Processing with Apache Spark

Bernd Amann (University Pierre & Marie Curie – Paris, FR)

Joint work of Hubert Naacke, Bernd Amann, Olivier Curé
Main reference Hubert Naacke, Bernd Amann, Olivier Curé: "SPARQL Graph Pattern Processing with Apache Spark", in Proc. of the Fifth International Workshop on Graph Data-management Experiences & Systems, GRADES@SIGMOD/PODS 2017, Chicago, IL, USA, May 14–19, 2017, pp. 1:1–1:7, ACM, 2017.
URL http://dx.doi.org/10.1145/3078447.3078448

For guaranteeing scalability, high availability and fault tolerance, RDF store implementations are rarely built from scratch but rather designed on top of a existing data processing engines. Following this line of work, we propose and compare five SPARQL query processing approaches using standard hash-join and broadcast join implementations on top of Apache Spark. Our experimentations on real-world and synthetic data sets emphasize that hybrid join plans using both broadcast or hash-join join operators simultaneously are outperforming plans using only one kind of operator.

3.3 Linked Data Containers – Shipping Linked Data and Data Management Capabilities to Consumers

Sören Auer (Universität Bonn, DE)

The amount of Linked Data both open, made available on the Web, and private, exchanged across companies and organizations, have been increasing in recent years. Maintaining and making this data available is mainly in the responsibility of data providers. Moreover, building applications on top of Linked Data in order to provide, for instance, analytics, data access control, and privacy is left to the end user or data consumers. However, many resources in terms of development costs and equipment are required by both data providers and consumers, thus impeding the development of real-world applications over Linked Data. We propose to encapsulate Linked Data and data processing functionalities in a client-side system called Linked Data Container, intended to be used by data consumers. Linked Data Containers can be deployed on the data consumer environments, ranging from Big Data to light-weight platforms.

As we learned in numerous workshops with more than 50 partner companies of the Industrial Data Space Association, keeping some level of control over the data – called data sovereignty – is a key requirement in industrial data sharing scenarios and currently the main obstacle for establishing data value chains in the industry. In many cases, cooperation partners should only gain access to a well defined fragment or usage access regime of the data.

For example, a cooperation partner in a customer bonus program, should be enabled to access information about a specific customer (e.g., identified by name or member id), but not be allowed to retrieve all customers with email and mailing addresses. In physical value chains, containers play a key role in material, component, half product, and product exchange. Containers in most cases fulfill the function to secure, condition (e.g., cool/warm), observe, or provide access to their containment.

We propose the concept of Linked Data Containers, which is a key element in the Industrial Data Space reference architecture. In order to realize the concept, we advocate for bundling data, security, access, and data processing functionality in a single artifact – the Linked Data Container. The approach can based on the recently emerging light-weight virtualization techniques and load balancing for scalable, high-performance query execution. As a result, LDC represents a novel data sharing and access paradigm, which balances costs and efforts differently between data provider and consumer than prior solutions (such as dumps, SPARQL endpoints, or TPFs). It enables controlling data access even after data shipping, thus it contributes to increased data sovereignty and consequently, in summary, it better fulfills the requirements of industrial data value chains.

3.4 Decentralizing the Semantic Web: Who will pay to realize it?

Abraham Bernstein (Universität Zürich, CH)

Joint work of Tobias Grubenmann, Daniele Dell'Aglio, Dmitry Moor, Sven Seuken, Abraham Bernstein
Main reference Tobias Grubenmann, Daniele Dell'Aglio, Abraham Bernstein, Dmitry Moor, Sven Seuken: "Decentralizing the Semantic Web: Who Will Pay to Realize It?", in Proc. of the Workshop on Decentralizing the Semantic Web 2017 co-located with 16th International Semantic Web Conference (ISWC 2017), CEUR Workshop Proceedings, Vol. 1934, CEUR-WS.org, 2017.
URL http://ceur-ws.org/Vol-1934/contribution-01.pdf

Fueled by enthusiasm of volunteers, government subsidies, and open data legislation, the Web of Data (WoD) has enjoyed a phenomenal growth. Commercial data, however, has been stuck in proprietary silos, as the monetization strategy for sharing data in the WoD is unclear. This is in contrast to the traditional web where advertisement fueled a lot of the growth. This raises the question how the WoD can (i) maintain its success when government subsidies disappear and (ii) convince commercial entities to share their wealth of data.

In this talk based on a paper [1], we propose a marketplace for decentralized data following basic WoD principles. Our approach allows a customer to buy data from different, decentralized providers in a transparent way. As such, our marketplace presents a first step towards an economically viable WoD beyond subsidies.

References

1 Tobias Grubenmann, Daniele Dell' Aglio, Abraham Bernstein, Dmitry Moor, and Sven Seuken. Decentralizing the Semantic Web: Who will pay to realize it?. In *Proceedings of the ISWC2017 workshop on Decentralizing the Semantic Web*, 2017.

3.5 Security, Privacy, and Semantics: Challenges and Opportunities

Piero Andrea Bonatti (University of Naples, IT)

After the initial focus on fully open data, the research on semantic data management is now facing the lack of support to access control and privacy enforcement. The knowledge-based nature of semantic (meta)data and the size of policies and policy-related information introduce further difficulties in the enforcement mechanisms, including anonymization, inference control

etc. There is an urgent need of collecting requirements both from federated query processing and from security/privacy enforcement, and assembling a framework for secure and privacy-enhancing federated, semantic query processing. Some of the hard challenges are:

1. Finding an optimal tradeoff between the expressivity of policy languages and the complexity of reasoning about policies.
2. Choosing a suitable confidentiality criterion to protect knowledge from attacks based on inference and metaknowledge. Such criterion should take into account also the probabilistic inferences that can be made with the help of machine learning algorithms.
3. According to the forthcoming General Data Protection Regulation, none of the anonymization methods known today produces data that can be regarded as anonymous in a legal sense. Consequently, it is crucial to manage data-subjects' consent to data processing (which makes it legal when data are not ideally anonymous and the processing does not belong to a short list of special cases of public interest).
   Here semantic languages and technologies can solve a number of problems related to expressiveness, flexibility, and interoperability. This is the approach taken, for instance, in the H2020 project SPECIAL.

References

1 Piero A. Bonatti: Datalog for Security, Privacy and Trust. Datalog 2010: 21–36
2 Piero A. Bonatti, Luigi Sauro: A Confidentiality Model for Ontologies. International Semantic Web Conference (1) 2013: 17–32
3 Joachim Biskup, Piero A. Bonatti, Clemente Galdi, Luigi Sauro: Optimality and Complexity of Inference-Proof Data Filtering and CQE. ESORICS 2014: 165–181
4 Piero A. Bonatti, Sabrina Kirrane, Axel Polleres, Rigo Wenning: Transparent Personal Data Processing: The Road Ahead. SAFECOMP Workshops 2017: 337–349

3.6 Framework for Allowing Secure and Private Access over a Federation of SPARQL Endpoints

Carlos Buil-Aranda (TU Federico Santa María – Valparaíso, CL)

Joint work of Sabrina Kirrane, Johann-Christoph Freytag, Piero Andrea Bonatti, Katja Hose, Jorge Lobo, Stasinos Konstantopoulos, Carlos Buil-Aranda

Semantic Federated Query Processing has been focused so far in improving the access to a set of SPARQL endpoints (RDF databases) and in selecting to which databases send the SPARQL queries in the main federated query. However all these improvement assume that all data is distributed across open and free to access RDF databases and none of the existing systems assume that these data may have restricted access or security policies to access the exposed data. To solve this problem we envisaged an abstract model for enabling policies in a federated data environment, security management and enforce nodes enforce in the federation engine to use a security and access framework. This model presents a framework in which a Semantic Federated Query Processing System should accommodate for effectively implementing security and privacy over the data it is being federated.

3.7 ACQUA: Approximate Continuous QUery Answering over Streams and Dynamic Linked Data Sets

Emanuele Della Valle (Polytechnic University of Milan, IT)

Joint work of Abraham Bernstein, Soheila Dehghanzadeh, Daniele Dell'Aglio, Shen Gao, Alessandra Mileo, Shima Zahmatkesh, Emanuele Della Valle

Web applications that federate dynamic data stream with distributed background data are getting a growing attention in recent years. Answering in a timely fashion, i.e., reactiveness, is one of the most important performance indicators for those applications.

The Semantic Web community showed that RDF Stream Processing (RSP) [1] is an adequate framework to develop this type of applications. However, RSP engines may lose their reactiveness due to the time necessary to access the background data when it is distributed over the Web. State-of-the-art RSP engines remain reactive using a local replica of the background data, but it progressively becomes stale if not updated to reflect the changes in the remote background data. For this reason, in the last two years, we investigated maintenance policies of the local replica that guarantee reactiveness while maximizing the freshness of the replica. They are collectively named ACQUA: Approximate Continuous QUery Answering over streams and Dynamic Linked Data sets.

In the early work [2], we focused on a continuous join operator between background data (accessed using a SPARQL 1.1 service clause) and stream data (accessed using an RSPQL WINDOW clause) assuming a 1:1 correspondence between the mappings returned on the window clause and those returned by the service clause. Then, we extended it in three directions: 1) we allowed an N:M relationship in the join [3], 2) we showed it is possible to dynamically adjust the policy [3] and 3) we added a filter clause to the service clause [4]. More recently, we showed that the opinion of multiple policies can be combined using rank aggregation[5].

References

1 Daniele Dell'Aglio, Emanuele Della Valle, Frank van Harmelen and Abraham Bernstein. Stream Reasoning: a Survey and Outlook: A summary of ten years of research and a vision for the next decade. In *Data Science Journal*, 1, 2017.
2 Soheila Dehghanzadeh, Daniele Dell'Aglio, Shen Gao, Emanuele Della Valle, Alessandra Mileo, and Abraham Bernstein: Approximate Continuous Query Answering over Streams and Dynamic Linked Data Sets. In *ICWE* 2015: 307–325
3 Shen Gao, Daniele Dell'Aglio, Soheila Dehghanzadeh, Abraham Bernstein, Emanuele Della Valle, Alessandra Mileo: Stream-Driven Linked-Data Access Under Update-Budget Constraints. In *ISWC*(1) 2016: 252–270
4 Shima Zahmatkesh, Emanuele Della Valle, Daniele Dell'Aglio: When a FILTER Makes the Difference in Continuously Answering SPARQL Queries on Streaming and Quasi-Static Linked Data. In *ICWE* 2016: 299–316
5 Shima Zahmatkesh, Emanuele Della Valle, Daniele Dell'Aglio: Using Rank Aggregation in Continuously Answering SPARQL Queries on Streaming and Quasi-static Linked Data. In *DEBS* 2017: 170–179

3.8 Why Federated Semantic Data Management Must Be FAIR

Michel Dumontier (Maastricht University, NL)

Main reference Mark D. Wilkinson, Michel Dumontier et al.: "The FAIR Guiding Principles for scientific data management and stewardship", Scientific Data 3, Vol. 3, 2016.
URL http://dx.doi.org/10.1038/sdata.2016.18

New infrastructure is needed to make digital content findable, accessible, interoperable, and reusable, as defined by the FAIR (Findability, Accessibility, Interoperability, and Reuse) principles [1]. The FAIR principles articulate a new direction for the management of digital content: that the use of globally unique, persistent identifiers to denote and retrieve structured data and metadata that meet the expectations of their communities and are expressed using global standards for semantic knowledge representation. These are all crucial aspects of Federated Semantic Data Management (FSDM). However, for FSDM to be truly realized on a global scale, new efforts must be made to create a social-technical infrastructure. Critically, we believe that efforts must be made to build out capacity, in sustainable manner, to publish, find, and reuse FDSM components including identifiers, descriptions, mappings, queries, formats, procedures, analytics, visualizations, etc. Having such components available will usher a next generation of the semantic web that people can truly embrace.

References

1 Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data 3*, 2016.

3.9 Privacy in the Context of Federated Semantic Data Management (FSDM) Systems

Johann-Christoph Freytag (Humboldt-Universität zu Berlin, DE)

Although privacy and its protection is quite well understood in the context of tabular data there has been little to no work how to take the concepts based on k-anonymity and differential privacy into the federated semantic web world. Based on the existing concepts for FSDM systems this workshop should give an understanding what the requirements and the challenges are to introduce privacy into this world.

For this workshop I presented our work on how to detect breaches of privacy when executing a sequence of queries over a database table that stays unchanged. I show how to transform this problem into a (bipartite) graph problem and outline the challenges of how to perform inference on a set of graphs that represent the anonymized query results.

3.10 Semantics of RDF and SPARQL: Some Considerations

Claudio Gutierrez (University of Chile – Santiago de Chile, CL)

Main reference Renzo Angles, Claudio Gutierrez: "The Multiset Semantics of SPARQL Patterns", in Proc. of the The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 9981, pp. 20–36, 2016.
URL http://dx.doi.org/10.1007/978-3-319-46523-4_2

The semantics of RDF and SPARQL combine opportunities and challenges for dealing with federation in the open world of the Web. On one hand, it they allow (real) distributed creation and management of resources and vocabulary and distributed population and linking of distributed data. On the other hand, addressing incomplete information in these specifications is both complex and obscure and the logic of RDF combined with that of SPARQL is complex (even if we restrict to RDFS). In one sentence, RDF and SPARQL offer rich opportunities to deal with federation at Web level.

Under the above border conditions, my suggestion is to address, to start building, federation with basic RDF and the relational core of SPARQL (Select, Filter, And, Union, and Except, that is, the Select, Where, Natural Join, Union All and Except of SQL, see [1]), frameworks that offer all the securities and background of SQL.

Once this basic floor is firmly established, one could think of extending in the several possible directions that this core offers, namely, Bags, Incomplete information (blanks and unbound), Paths, Subqueries and Aggregation combined with some of the others, and most important, delegation features (the From Named, Graph, and Service features). Of course there are more possible extension, that I consider at this point –due to the state of the art in the previous levels– theoretical exercises: depth Logical reasoning, the interplay between CWA and OWA, several protocols (like update), etc.

Summarizing, SPARQL is an extremely complex language, and we do not know yet how the semantics of each extension interacts with other parts of the specification. This void is extremely dangerous when developing federation (that assumes each party will trust other pieces of the system). Thus my suggestion to start studying and developing federation using the simple and trusted core indicated.

References

1 Renzo Angles, Claudio Gutierrez: The Multiset Semantics of SPARQL Patterns. International Semantic Web Conference (1) 2016: 20–36

3.11 Ephedra: Extending SPARQL Federation for Efficient Combination of RDF Data and Services

Peter Haase (Metaphacts GmbH, DE)

Joint work of Andriy Nikolov, Peter Haase

Knowledge graph management use cases often require addressing hybrid information needs that involve multitude of data sources, multitude of data modalities (e.g., structured, keyword, geospatial search), and availability of computation services (e.g., machine learning and graph analytics algorithms). Although SPARQL queries provide a convenient way of expressing

information needs over RDF knowledge graphs, the level of support for hybrid information needs is limited: existing query engines usually focus on retrieving RDF data and only support a set of hard-coded built-in services. In this paper we describe representative use cases of metaphacts in the cultural heritage and pharmacy domains and the hybrid information needs arising in them. To address these needs, we present Ephedra: a SPARQL federation engine aimed at processing hybrid queries. Ephedra provides a flexible declarative mechanism for including hybrid services into a SPARQL federation and implements a number of static and runtime query optimization techniques for improving the hybrid SPARQL queries performance.

3.12 Integration and Interoperability of Graph-Data Systems

Olaf Hartig (Linköping University, SE)

My current research agenda focuses on establishing and on studying the notion of a federation of graph data systems. More precisely, I will investigate approaches (i) to integrate graph data across different systems that manage and process such data, and (ii) to integrate such systems as members of a federated system; this federated system will be able to perform workloads of queries and analysis algorithms transparently on the data that is distributed over the federation members. As an initial step towards such an integration I am investigating approaches to reconcile RDF and Property Graphs, which are the two prevalent graph data models used in many graph data systems. My current effort to achieve such a reconciliation [1] focuses on extending the RDF data model and its query language SPARQL to allow users to capture and query statement-level metadata [2].

References

1 Olaf Hartig. *Reconciliation of RDF* and Property Graphs*. In CoRR abs/1409.3288, 2014
2 Olaf Hartig. *Foundations of RDF* and SPARQL* – An Alternative Approach to Statement-Level Metadata in RDF*. In Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW), 2017

3.13 Federated Linked Data in Libraries

Jana Hentschke (Deutsche Nationalbibliothek – Frankfurt am Main, DE)

URL http://www.dnb.de/EN/lds

As a provider of RDF data the German National Library is tracking requirements to data modeling as well as infrastructure to support the intended usage of its data on the web: to be queried along with related and linked datasets. In order to be able to offer a sensible and reliable service we need to forecast user needs by monitoring the latest developments in the relevant research disciplines. Library metadata management is currently seeing a shift in data model – semantically as well as technically. This is a worldwide process which the German National Library actively involves in.

3.14 Linked Open Data, Federations, and beyond

Katja Hose (Aalborg University, DK)

In the past couple of years numerous approaches and techniques for query processing in federations of SPARQL endpoints and over Linked Open Data have been proposed by the research community. These techniques cover various subproblems, such as indexing, join processing, reasoning, query optimization, knowledge extraction, quality and completeness of knowledge bases as well as data integration, semantic data warehouses, and many more. As discussed in this talk, we are still far from having reached a point where we can conclude that we have found sufficiently good solutions for query processing in this setup. Apart from finding an overall solution that combines solutions to all these subproblems [1, 2], we are even still missing solutions for seemingly small problems, such as encoding metadata.

References

1 L. Galárraga, K. Hose, S. Razniewski. *Enabling Completeness-aware Querying in SPARQL*. In Proceedings of the 20th Int. Workshop on the Web and Databases (WebDB), 2017.
2 Jacobo Rouces, Gerard de Melo, Katja Hose. *Heuristics for Connecting Heterogeneous Knowledge via FrameBase*. In The Semantic Web. Latest Advances and New Domains. (ESWC), 2016.

3.15 The Next Generation Internet of Autonomy

Sabrina Kirrane (Wirtschaftsuniversität Wien, AT)

The Next Generation Internet should be about autonomy. Both organisations and individuals need to be able to publish both structured and unstructured data in a manner that allows them to control who can access, what data, under which constraints. Such a vision will require the adoption of existing and the development of new security and privacy mechanisms for control, transparency and compliance checking. Data consumers will naturally need to deal with diversity in the data and the query mechanism, this becomes much more complicated when there is a need to query distributed data sources. The focus of our work is to understand how existing federated query engines can be enhanced in such a way that both open and closed data can be queried in a manner that is capable of dealing with access and usage policies that are attached to the data, taking into consideration various robustness requirements.

3.16 Intelligent Data Management

Stasinos Konstantopoulos (Demokritos – Athens, GR)

In general, looking into the intersection of artificial intelligence with various subjects, including robot perception [1], computational linguistics [2], health data processing, and in particular when it comes to Dagstuhl, federated query execution planning and optimization [3, 4].

The access control and privacy group attracted my attention, as it puts forward challenging requirements for federated query processors: to offer the SPARQL programmer a seamless and transparent integrated view of a system of endpoints that comprises public endpoints and endpoints that impose complex and heterogeneous restrictions on data access and processing.

References

1 Giannakopoulos T., Konstantopoulos S., Siantikos G., and Karkaletsis V.: Design for a system of multimodal interconnected ADL recognition services. Components and Services for IoT Platforms. Springer, September 2016.
2 Konstantopoulos S.: Looking for meaning in names. From Semantics to Dialectometry. College Publications, January 2017.
3 Charalambidis, A., Troumpoukis, A., Konstantopoulos, S.: Semagrow: Optimizing federated SPARQL queries. In: Proceedings 11th Int'l Conference on Semantic Systems (SEMANTiCS 2015), Vienna, Austria, September 2015.
4 Zamani, K., Charalambidis, A., Konstantopoulos, S., Zoulis, N., and Mavroudi, E.: Workload-aware self-tuning histograms for the Semantic Web. Transactions on Large-Scale Data- and Knowledge-Centered Systems 28, Springer, 2016.

3.17 Privacy and Security in the Semantic Web

Jorge Lobo (UPF – Barcelona, ES)

Although much work has been done in the semantic web, issues of privacy and security have not been explored. Nevertheless, the field and the existing federation systems have reached the maturity to require a more systematic study of these issues. I have been working in policy-based management for distributed systems for more than a decade and I have come to the meeting to better understand the particularities of working on the open web that might need to be considered when developing a security and privacy policy management framework for the semantic web. It is foreseen that such a framework will touch upon several lines of work in policy management, from authoring to refinement, composition and analysis.

3.18 Semantic Web in the Fog of Browsers

Pascal Molli (University of Nantes, FR)

URL http://pagesperso.lina.univ-nantes.fr/ molli-p/pmwiki/pmwiki.php

Imagine connecting thousands of web browsers with browser-to-browser connections, sharing storage, bandwidth, and CPU. This builds a fog of browsers where end-user devices are ready to collaborate. Imagine semantic fog applications running in fogs of browsers, querying the linked data servers hosted in the cloud and data hosted in the fog. Fogs of browsers running semantic fog applications create a new massively decentralized infrastructure where RDF data and SPARQL query processing are available both on web servers and on browsers. I explore new opportunities and research challenges opened by a fog of browsers for the semantic web.

3.19 Query Optimization against Federations of SPARQL Endpoints

Gabriela Montoya (Aalborg University, DK)

Joint work of Katja Hose, Hala Skaf-Molli, Pascal Molli, Maria-Esther Vidal, Gabriela Montoya
Main reference Gabriela Montoya, Hala Skaf-Molli, Katja Hose: "The Odyssey Approach for Optimizing Federated SPARQL Queries", in CoRR, Vol. abs/1705.06135, 2017.
URL http://arxiv.org/abs/1705.06135

Optimization of SPARQL queries against federations of SPARQL endpoints includes: i) source selection: identifying relevant sources for each triple pattern; ii) query decomposition: combining triple patterns into subqueries to be evaluated at the endpoints; iii) join ordering: identifying the best order to evaluate the subqueries and the best ways to combine their results.

Query decomposition in the context of federations with replicated data can exploit knowledge about how data have been replicated to decompose the queries into subqueries that reduce the amount of data transfer by sending more selective subqueries to endpoints. These subqueries exploit data locality present at the endpoints to improve their availability.

If some knowledge about the data exposed by the endpoints is available, it could be exploited to obtain good estimations of cardinality that lead to generate good plans. These plans have less subqueries and during execution they require the transfer of less data and exhibit fast execution time.

References

1 Gabriela Montoya, Hala Skaf-Molli, Pascal Molli, Maria-Esther Vidal: "Decomposing federated queries in presence of replicated fragments". J. Web Sem. 42: 1-18 (2017)

3.20 Online Query Answering Using Knowledge Graphs, and Entity Resolution for Very Large and Highly Heterogeneous Data

Themis Palpanas (Paris Descartes University, FR)

Search engines are continuously employing advanced techniques that aim to capture user intentions and provide results that go beyond the data that simply satisfy the query conditions. Examples include the personalized results, related searches, similarity search, popular and relaxed queries. In this work we introduce a novel query paradigm that considers a user query as an example of the data in which the user is interested. We call these queries "exemplar queries". and claim that they can play an important role in dealing with the information deluge. We provide a formal specification of the semantics of such queries and show that they are fundamentally different from notions like queries by example, approximate and related queries. We provide an implementation of these semantics for graph-based data and present an exact solution with a number of optimizations that improve performance without compromising the quality of the answers. We study two different similarity functions, isomorphism and strong simulation, for retrieving the answers to an exemplar query, and we provide solutions for both. We also provide an approximate solution that prunes the search space and achieves considerably better time-performance with minimal or no impact on effectiveness. We experimentally evaluate the effectiveness and efficiency of these solutions with synthetic and real datasets, and illustrate the usefulness of exemplar queries in practice.

In addition, we present JedAI, a toolkit for Entity Resolution that can be used in three different ways: as an open-source Java library that implements numerous state-of-the-art, domain-independent methods, as a workbench that facilitates the evaluation of their relative performance and as a desktop application that offers out-of-the-box ER solutions. JedAI bridges the gap between the database and the Semantic Web communities, offering solutions that are applicable to both relational and RDF data. It also conveys a modular architecture that facilitates its extension with more methods and with more comprehensive workflows.

3.21 FOWLA: A Federated Architecture for Ontologies

Ana Maria Roxin (Universite de Bourgogne, FR)

Joint work of Ana Maria Roxin, Christophe Nicolle, Tarcisio Mendes de Farias
Main reference Tarcisio M. Farias, Ana Roxin, Christophe Nicolle: "FOWLA, A Federated Architecture for Ontologies", in Proc. of the Rule Technologies: Foundations, Tools, and Applications - 9th International Symposium, RuleML 2015, Berlin, Germany, August 2-5, 2015, Proceedings, Lecture Notes in Computer Science, Vol. 9202, pp. 97–111, Springer, 2015.
URL http://dx.doi.org/10.1007/978-3-319-21542-6_7

The progress of information and communication technologies has greatly increased the quantity of data to process. Thus managing data heterogeneity is a prob-lem nowadays. In the 1980s, the concept of a Federated Database Architecture (FDBA) was introduced as a collection of components that, by means of loosely coupled federation, share and exchange information. Semantic web technologies mitigate the data heterogeneity problem, however due to the data structure heterogeneity the integration of several ontologies is still a complex task. For tackling this problem, I have worked on the definition of a loosely coupled

federated ontology architecture (FOWLA). This approach allows the coexistence of various ontologies sharing common data dynamically at query execution through Horn-like rules and inference. It is also at query time that the data access policies for the federated ontologies are checked. The implementation of the FOWLA architecture comes with several advantages for interoperating several ontologies, as it allows: (1) inferring new ontology alignments; (2) avoiding data redundancy; (3) modularizing the maintainability, thought preserving the autonomy among the considered ontology-based systems, (4) addressing queries composed of vocabulary terms issued from different ontologies and (5) improving query execution time through a selection of the rules pertaining to a given query.

3.22 DREAM: Distributed RDF Engine with Adaptive Query Planner and Minimal Communication

Sherif Sakr (KSAU – Riyadh, SA)

Joint work of Mohammad Hammoud, Dania Abed Rabbou,Seyed Mohammad Reza Nouri, Seyed Mehdi Reza Beheshti, Sherif Sakr
Main reference Mohammad Hammoud, Dania Abed Rabbou, Reza Nouri, Seyed-Mehdi-Reza Beheshti, Sherif Sakr: "DREAM: Distributed RDF Engine with Adaptive Query Planner and Minimal Communication", in PVLDB, Vol. 8(6), pp. 654–665, 2015.
URL http://www.vldb.org/pvldb/vol8/p654-Hammoud.pdf

DREAM is a distributed and adaptive RDF system. To the contrary of all existing RDF systems, DREAM partitions SPARQL queries instead of partitioning RDF datasets. By not partitioning datasets, DREAM offers a general paradigm for all types of queries, and entirely averts intermediate data shutting (only meta-data are transferred). On the other hand, by partitioning queries, DREAM presents an adaptive scheme, which automatically runs queries on different numbers of machines depending on their complexities. This is achieved via employing a novel graph-based, rule-oriented query planner and a new cost model. As a result, DREAM combines the advantages of the state-of-the-art centralized and distributed RDF systems, where data communication is avoided and cluster resources are aggregated, and precludes their disadvantages, where system resources are limited and communication overhead is typically hindering.

3.23 Federated Semantic Data Management Systems in Practice

Juan F. Sequeda (Capsenta Inc. – Austin, US)

We are observing the rise and deployment of real world data integration systems based on federation and semantic technologies. The common setup we see in practice consists of the following elements: a set of source relational databases; a target ontology which provides a global semantic description of the domain, independent of the sources; and a set of mappings from the databases to the ontology. The goal is to answer queries in terms of the target ontology in a federated manner. From a practical point of view, this begs the question: where does the target ontology and the mappings come from?

We are investigating and developing methodologies and tools that can help non-experts to design the main components of a federated semantic data management system. For example in one project, we propose a pay-as-you-go methodology to design the target ontology and mapping driven by the expected questions that the semantic federated system should answer. The goal is to create the target ontology and mappings in an incremental manner, thus provide answers to questions as early as possible.

3.24 Data Availability and Efficient Query Processing for the Semantic Web

Hala Skaf-Molli (University of Nantes, FR)

Data availability and efficient query processing are challenging problems for the Semantic Web. My current research is to build decentralized and federated infrastructures to reach these objectives. More precisely, data replication improves data availability but degrades federated query processing performances, how to handle replicated data during federated query processing? Cache at client-side reduces the overhead on the server but clients do not share their caches, how to build a decentralized cooperative cache so clients can share caches? Parallel query processing and SPARQL query processing in the Fog could be improve query execution time. How to execute SPARQL queries in the Fog?

3.25 Adaptive Decentralized Control in Distributed Web Applications

Rudi Studer (KIT – Karlsruher Institut für Technologie, DE)

Joint work of Felix Keppmann, Andreas Harth

Currently, we are witnessing the rise of new technology-driven trends such as the Internet of Things, Web of Things, and Factories of the Future that are accompanied by an increasingly heterogeneous landscape of small, embedded, and highly modularized devices and applications, multitudes of manufactures and developers, and pervasion of network-accessible "things" within all areas of life. At the same time, we can observe increasing complexity of the task of integrating subsets of heterogeneous components into applications that fulfil certain needs by providing value-added functionality beyond the pure sum of their components. Enabling integration in these multi-stakeholder scenarios requires new architectural approaches for adapting components, while building on existing technologies and thus ensuring broader acceptance.

To this end, we discuss current integration-related challenges, present our approach for automated component adaptation, and describe our integration architecture that enables decentralized control.

References

1 Felix Leif Keppmann, Maria Maleshkova, Andreas Harth. *Semantic Technologies for Realising Decentralised Applications for the Web of Things*. 21st International Conference on Engineering of Complex Computer Systems, Nov 6-8, 2016, Dubai, UAE

3.26 Federated Querying on the Web

Joachim Van Herwegen (Ghent University, BE)

Joint work of Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, Pieter Colpaert
Main reference Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, Pieter Colpaert: "Triple Pattern Fragments: A low-cost knowledge graph interface for the Web", in J. Web Sem., Vol. 37-38, pp. 184–206, 2016.
URL http://dx.doi.org/10.1016/j.websem.2016.03.003

Triple Pattern Fragments (TPF) is a lightweight interface, allowing for SPARQL queries to be evaluated by moving some of the workload from the server to the client. TPF endpoints can be used to evaluate SPARQL queries by requesting individual pattern information through many HTTP requests and joining the results locally. Due to this querying process, these can easily be queried in a federated way by sending pattern requests to all endpoints at the same time and ignoring an endpoint if it can not answer a pattern. Interestingly, despite the substantially lighter server-side interface, the completeness and execution time of the FedBench benchmark of a TPF setup is comparable to that of a SPARQL endpoint setup [1].

To fully support querying on the Web, many other problems still have to be overcome. These include investigating:

- how to handle heterogeneous interfaces, all with their own restrictions, at the same time,
- the existing benchmarks and whether they are sufficient,
- multiple metrics besides response time, and
- how to discover sources on the web.

References

1 Verborgh, Ruben and Vander Sande, Miel and Hartig, Olaf and Van Herwegen, Joachim and De Vocht, Laurens and De Meester, Ben and Haesendonck, Gerald and Colpaert, Pieter. *Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web*. Journal of Web Semantics, 2016.

3.27 Federated Query Processing over RDF Data

Maria-Esther Vidal (Universidad S. Bolivar – Caracas, VE)

Main reference Maria-Esther Vidal, Simón Castillo, Maribel Acosta, Gabriela Montoya, Guillermo Palma: "On the Selection of SPARQL Endpoints to Efficiently Execute Federated SPARQL Queries", in Trans. Large-Scale Data- and Knowledge-Centered Systems, Vol. 25, pp. 109–149, 2016.
URL http://dx.doi.org/10.1007/978-3-662-49534-6_4

The increasing number of RDF data sources that allow for querying Linked Data via Web services form the basis for federated query processing over Web-accessible RDF data sources. Federated SPARQL query engines provide a unified view of a federation of RDF data sources and rely on different components to exploit the semantics encoded in RDF data during query execution. The problem of federation query processing has been extensively studied by Database and Semantic Web communities; however, these technologies have not been used in large scale yet. Additionally, there is no standard and formal definition of the problem of query processing over a federation of RDF data sources, impeding a formal evaluation of properties of the state-of-the-art approaches. During this seminar, we analyze different

scenarios of federations of RDF data sources, e.g., entailment regimes to be considered during query processing, source query capabilities, or access control, and propose a formal definition of the problem of query processing over a federation of RDF data sources. State-of-the-art approaches are evaluated in terms of the proposed formalization. We hypothesize that this characterization will allow for a better understanding of the state-of-the-art, as well as for uncovering the limitations of the current technologies that have impeded the use of existing approaches on a large scale.

References

1 Maribel Acosta, Maria-Esther Vidal, Tomas Lampo, Julio Castillo, and Edna Ruckhaus. ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints. In *The Semantic Web – ISWC 2011 – 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I.*

4 Working groups

In this section, each of the four working groups provides a summary of their discussions and the results of their work during the seminar.

4.1 Foundations of Federated Semantic Data Management on the Web

Bernd Amann (University Pierre & Marie Curie – Paris, FR), Emanuele Della Valle (Polytechnic University of Milan, IT), Claudio Gutierrez (University of Chile – Santiago de Chile, CL), Olaf Hartig (Linköping University, SE), Themis Palpanas (Paris Descartes University, FR), and Rudi Studer (KIT – Karlsruher Institut für Technologie, DE)

4.1.1 Introduction

The Semantic Web vision introduced by Tim Berners-Lee almost 20 years ago has attracted a considerable attention from various computer science domains such as databases (research tracks on RDF data management in the main database conferences like VLDB[1] and SIGMOD[2]), Artificial Intelligence (AI Magazine[3]), Web (WWW conference), and Information Retrieval [11]. The corresponding communities developed solutions for generating [10], analyzing [8], storing, and querying [4] large RDF / knowledge graphs [3, 7] which are used in many uses cases and applications[4]. SPARQL query federation engines and Linked Open Data infrastructures are initial steps towards building such applications at the Web level. However, the vision of a universal and open space for meaningfully sharing data on the Web is still not fully achieved[5].

[1] http://vldb2016.persistent.com/VLDB2016-FullProgram.html#TueF1115T1245R2
[2] http://www.sigmod2015.org/toc_sigmod.shtml
[3] https://www.aaai.org/ojs/index.php/aimagazine/article/view/2161
[4] https://www.w3.org/2001/sw/sweo/public/UseCases/
[5] https://cacm.acm.org/magazines/2016/9/206254-a-new-look-at-the-semantic-web/fulltext

During the week of the Dagstuhl seminar, the "RDF and Graph DB" working group spent several group sessions discussing the current and future issues in federated semantic data management. During the first session we mainly discussed how standard, RDF-based Semantic Web technologies compare to other data and knowledge graph systems and to existing data federation infrastructures. We made two main observations:

1. The RDF data model is a "universal" data model in the sense that it is designed for sharing data and knowledge in an unbounded space such as the Web [2]. The universality and the unbounded nature of this new scenario, for which RDF/SPARQL were originally thought of, presents challenges that are fundamentally different from other data/knowledge graph data models and query languages which are implemented in "closed" systems (regarding data and users) for complex graph queries and analytics.
2. There exist several federated data management frameworks (relational, P2P, Web services) including advanced declarative approaches for building overlay networks [6], and for distributed query processing and reasoning [1, 5]. These approaches address many fundamental concepts of federated semantic web management but they still require a further development and integration effort in order to move from the research prototype stage to standardized open federation frameworks for building applications.

Based on these observations, we decided to revisit the initial Semantic Web vision by taking into account the current scientific and technological state of the art related to Federated Semantic Data Management (FSDM). The goal was to define the fundamental characteristics of FSDM, where the purpose of this definition was to provide a basis for analyzing the opportunities and the limitations of existing semantic and/or federated data management solutions and for preparing a scientific roadmap in FSDM research.

4.1.2 Foundations and Characteristics

Our first step was to define the notion of federated semantic data management and the main abstract principles that distinguish FSDM from other frameworks like federated RDBMSs, P2P data management, graph databases, etc.

Definition (initial version; inspired by the definition of LOD[6]): Federated Semantic Data Management (FSDM) refers to universally and meaningfully publishing, connecting, and processing data in an unbounded space through a network of autonomous data sources exposed on the Web. The word "meaningfully" in this context refers to the transparent use of knowledge that is made available in an autonomous way by the data sources that participate in the federation.

Based on this definition, we have identified five principles that characterize FSDM: universality, unboundedness, dynamicity, network protocols, and semantics. Universality and unboundedness are two main principles related to the Web. Universality[7] denotes the possibility for any federation member (data source or client) to publish, connect, and consume data "anywhere on the Web." In the context of RDF, the universality principle is mainly represented by the notion of URI. Unboundedness reflects the possibility to build graphs of unbounded size where the notion of graph may refer to any of the following: raw and structured data, knowledge (vocabulary, schema, ontology), and data sources connected through a network. As a consequence of these two principles, universality and unboundedness, the complete set of all federation members cannot be assumed to be known in advance, and

[6] http://linkeddata.org/
[7] https://www.w3.org/1999/04/WebData#gloloc

neither can the exact content or the size of data and data sources. Dynamicity reflects the temporal evolution of these graphs; this evolution may be fast (e.g., RDF data streams) or slow (e.g., ontology evolution, network topology). The notion of federation in FSDM is mainly represented by the principle of network protocols. This principle stresses the application of a system of rules that allow two or more entities of a communication system to transmit information via any kind of variation of a physical quantity[8]. These rules are a fundamental part in the definition of distributed data and knowledge processing algorithms, cost models, and optimization techniques. In the RDF context, this principle is mainly represented by the SPARQL protocol recommendation, a means of conveying SPARQL queries and updates from clients to SPARQL processors[9]. Finally, the notion of semantics in FSDM represents the capacity to define "intensional" data and knowledge which can be made explicit through different inference mechanisms (RDF entailment regimes).

4.1.3 Systems

Based on the previous analysis, our second goal was to understand the impact that the identified characteristics and principles have on current and future FSDM systems. To measure this impact we started to study existing models and systems in the literature and in practice. This first study led us to the formulation of the following two hypotheses.

1. First, we argue that it is impossible to build a "perfect" FSDM system that fully achieves universality, unboundedness, and dynamicity, all at the same time [9]. As a consequence, we see the need to define new concepts and new metrics that will play, in this space of FSDM, the role played by soundness and completeness play in logic, or by precision and recall in information retrieval.
2. Second, we conjecture that the two principles of federation and semantics are interdependent, and must be tackled together. In particular, we believe that, for building effective and efficient solutions, it is not sufficient to “simply” extend a federated data management system with semantics or, vice versa, extend a semantic data management system with the notion of federation.

These two hypotheses raise a number of new challenges for current and future FSDM systems including:

- the formalization of notions of federated semantic queries,
- the definition of effective cost models and optimization techniques for federated query engines,
- the definition and the implementation of benchmarks for evaluating and comparing FSDM systems,
- the elaboration of guidelines for choosing solutions and building applications (which might have different levels of constraints for various characteristics).

4.1.4 Next Steps

The immediate next step for our working group is twofold: We aim to survey the state of the art of the aforementioned existing frameworks of federated data management and highlight their relationship (or the lack thereof) to the five principles of FSDM that we have identified, and we want to document in detail the discussions that we had in Dagstuhl. The purpose

[8] https://en.wikipedia.org/wiki/Communications_protocol
[9] https://www.w3.org/TR/sparql11-protocol/

of this work will be to provide a detailed justification and rationale for the aforementioned observations, hypotheses, and challenges. Thereafter, and based on this work, we aim to provide recommendations on research topics and problems that need to be addressed in order to build FSDM systems. We are planning to bring together the results of this work in a publication co-authored by all members of the working group.

References

1. Serge Abiteboul, Pierre Senellart, and Victor Vianu. The erc webdam on foundations of web data management. In *Proceedings of the 21st International Conference on World Wide Web*, pages 211–214. ACM, 2012.
2. Tim Berners-Lee. Www: Past, present, and future. *Computer*, 29(10):69–77, 1996.
3. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia – a crystallization point for the web of data. *Web Semant.*, 7(3):154–165, September 2009.
4. Zoi Kaoudi and Ioana Manolescu. Rdf in the clouds: A survey. *The VLDB Journal*, 24(1):67–91, February 2015.
5. Boon Thau Loo, Tyson Condie, Minos Garofalakis, David E Gay, Joseph M Hellerstein, Petros Maniatis, Raghu Ramakrishnan, Timothy Roscoe, and Ion Stoica. Declarative networking: language, execution and optimization. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 97–108. ACM, 2006.
6. Eng Keong Lua, Jon Crowcroft, Marcelo Pias, Ravi Sharma, and Steven Lim. A survey and comparison of peer-to-peer overlay network schemes. *IEEE Communications Surveys & Tutorials*, 7(2):72–93, 2005.
7. Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.
8. Radhika Sridhar, Padmashree Ravindra, and Kemafor Anyanwu. Rapid: Enabling scalable ad-hoc analytics on the semantic web. In *Proceedings of the 8th International Semantic Web Conference*, ISWC'09, pages 715–730, Berlin, Heidelberg, 2009. Springer-Verlag.
9. Jürgen Umbrich, Claudio Gutierrez, Aidan Hogan, Marcel Karnstedt, and Josiane Xavier Parreira. The ace theorem for querying the web of data. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 133–134. ACM, 2013.
10. Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In *Proceedings of the 8th International Semantic Web Conference*, ISWC'09, pages 650–665, Berlin, Heidelberg, 2009. Springer-Verlag.
11. Gerhard Weikum, Gjergji Kasneci, Maya Ramanath, and Fabian Suchanek. Database and information-retrieval methods for knowledge discovery. *Commun. ACM*, 52(4):56–64, 2009.

4.2 Summary of Federated Query Processing

Juan F. Sequeda (Capsenta Inc. – Austin, US), Maribel Acosta (KIT – Karlsruher Institut für Technologie, DE), Peter Haase (Metaphacts GmbH, DE), Katja Hose (Aalborg University, DK), Gabriela Montoya (Aalborg University, DK), Sherif Sakr (KSAU – Riyadh, SA), Hala Skaf-Molli (University of Nantes, FR), Joachim Van Herwegen (Ghent University, BE), and Maria-Esther Vidal (Universidad S. Bolivar – Caracas, VE)

Our group had extensive discussions on the state of the art in Federated Query Processing from the traditional Relational Databases and Semantic Web perspectives. The goal was to understand the limitations of current approaches in considering ontological knowledge during federated query processing.

We started off by discussing what we understood by the term "semantics" within federated query processing. For some members of the group, it was assumed that federated query processing over RDF already implied "semantics" because it was considering the simple entailment[10] of RDF for reasoning. For others, federated query processing over RDF was rather a change of representation from relational to graph data model for the federated query processing problem. Thus, the differences between traditional federated query processing and federated semantic query processing was not clear from the beginning.

In order to overcome the discrepancies of our assumptions, our goal was to come up with a formal definition of federated semantic query processing where different ontological entailments of reasoning (RDFS, OWL 2 QL, etc) were explicit in the definition. This formal definition was one of the main result of our group.

During our presentation of our formal definition to the entire Dagstuhl group, we received comments that exposed possible research problems in the area of federated semantic data management. These are listed in the answer to Question 2 below.

Our next step is to write a survey paper with the goal of highlighting

1. What is Federated Semantic Data Management (FSDM)
2. How FSDM differs from traditional Federated Data Management (FDM)
3. What are the research problems and open challenges in FSDM.

4.2.1 Results

The main results of our group are the following:

- Formal definition of the Federated Semantic Query Processing problem which includes entailment regimes for reasoning (i.e., RDFS, OWL 2 QL, etc).
- Analysis of state-of-the-art Federated Semantic Data Management tools (FedX, ANAPSIS, Triple Pattern Fragment) with respect to the definition of the Federated Semantic Query Processing problem.
- Definition of the Source Selection and Query Decomposition Problem for Federated Semantic Query Processing based on our previous definition.

[10] https://www.w3.org/TR/rdf11-mt/#simpleentailment

4.2.2 Answers to Seminar Questions

Q1 *Can traditional techniques developed for federations of relational databases be enriched with RDF semantics, and thus provide effective and efficient solutions to problems of federated semantic data management?*

The characteristics of the RDF model impose restrictions on query processing over RDF data sources that impede Relational Databases technologies from providing efficient and effective solutions in general. For example, because datasets are described using binary predicates, for the source selection problem, we start out in the worst possible case scenario. Furthermore, in the Semantic Web, datasets also have "general predicates" e.g., from ontologies such as RDF/S and OWL), which cannot be used efficiently for source selection.

SPARQL operators have been implemented following the computational models of the relational algebra operators, e.g., block nested loop, dependent join, XJoin. Existing SPARQL physical operators implement the simple entailment regime. The question is how/if these operators need to be further extended to support entailments with higher expressivity (i.e., RDFS, OWL 2 QL, etc.)

Q2 *What problems of federated semantic data management present new research challenges that require the definition of novel techniques?*

The following are new research challenges within FSDM:

1. Unboundness: In the traditional federated data management problem, we have as input the set of known sources that we want to federate against. On the Web, this may not be the case: The set of sources may be unknown.
2. Correctness: In the traditional federated data management problem, we have a strict definition of correctness (a federated query is equal to a query over the entire universe of graphs). If the sources are not all known, then we may need a relaxed version of correctness. The tradeoff between soundness and completeness vs. precision and recall needs to be studied.
3. Dynamicity: How to deal with data that may change in different sources during the time of execution?

The following are extensions to existing problems of federated data management.

4. Access Control: Check access control before a query is executed and rewrite the query in order to make sure policies are enforced OR check access control during the execution of the query.
5. Source Selection: There are many more new constraints to consider within federated semantic data management, which makes the problem harder. For example,
   - Timeouts of the sources
   - Max k results
   - Different versions of SPARQL (1.0 vs 1.1)
   - Hardware capability
   - Different semantics of replicated sources (mirrors of sources)
   - Pay for accessing a source (public vs private)
   - Robots.txt
   - Query expressivity of different RDF sources: SPARQL, Linked Data Fragments, Triple Pattern Fragments
6. Heterogeneity: Although we are assuming a common data model (RDF), federated sources that access semantic data may be heterogeneous at the level of:
   - the schema (different ontologies used in different sources),

- different versions of SPARQL having an impact on the type of queries that sources are able to answer (e.g., SPARQL endpoint vs. TPF),
- computational/physical resources of the source,
- type of supported entailment regimes (RDFS, OWL 2 QL, etc)

7. Query Results: Adding provenance to the results to explain where the answer came from.

Q3 *What is the role of RDF semantics in the definition of the problems of federated semantic data management?*

Our simple answer: being able to do 1) reasoning/inferencing over 2) unbounded/unknown sources.

4.3 Privacy and Security Group Summary

Sabrina Kirrane (Wirtschaftsuniversität Wien, AT), Abraham Bernstein (Universität Zürich, CH), Piero Andrea Bonatti (University of Naples, IT), Carlos Buil-Aranda (TU Federico Santa María – Valparaíso, CL), Johann-Christoph Freytag (Humboldt-Universität zu Berlin, DE), Katja Hose (Aalborg University, DK), Stasinos Konstantopoulos (Demokritos – Athens, GR), and Jorge Lobo (UPF – Barcelona, ES)

4.3.1 Day 1 – Knowledge Sharing & Planning

The aim of day one was to discuss the status quo, to identify gaps that need to be addressed and to come up with a working plan for the rest of the week. After discussing existing work by the semantic web community on anonymisation, encryption and access control the group concluded that there are many open research challenges and it is clear that all problems cannot be solved immediately. As such the group agreed to focus on enhancing federated querying with access control. The output of the discussion was a plan for the remaining days with a view to working towards the following objectives:

- Identify a set of requirements that need to be considered (e.g. secure (compliant), soundness, maximal, computational complexity, bandwidth, robust against loss, leakage of meta policies, availability)
- Derive a conceptual framework that can be used to examine the trade-offs between different architectures and implementations
- Propose query and policy evaluation strategies taking into consideration the fact that there will be a tight coupling between access control and query planning
- Define an execution strategy towards optimisation for the identified requirements

4.3.2 Day 2 – Brainstorming

Building on existing work from the database and security communities, the group started by discussing conceptual access control models. This was followed by the mutual sharing of background information in relation to federated querying and policy enforcement. Here the term policy is used in the broader sense, for example constraints, recommendations, access policies, privacy policies, agreements etc. This naturally let to a discussion on the tight coupling between federated query planning and policies. The output of the discussion was a better understanding of:

- The set of requirements that need to be considered
- Initial thoughts on what the conceptual framework might look like

4.3.3 Day 3 – The conceptual framework

Day 3 focused exclusively on formally defining the conceptual framework that can be used to analyse policy aware federated semantic web architectures and their implementations. Key discussion points included:

- Who evaluates the policy, when and where? federation engine, endpoints, both?
- Do we need authentication, and if so who is responsible for authentication?
- What does the user/client send to the federation engine?
- What happens at the federation engine?
- What information is sent to the endpoints?
- What happens at the endpoints?
- What information is sent back from the endpoints?

4.3.4 Day 4 – Bringing it all together in the form of a paper

The final day was dedicated to discussing the shape of the paper, brainstorming about suitable publishing outlets, creating an initial structure for the paper, identifying who will be responsible for what, and deciding on next steps.

4.3.5 Results

The primary output of the group is a conceptual framework that can be used to analyse policy aware federated semantic web architectures and their implementations. In follow up work the framework will be used to examine the design space of possible solutions and discuss the tradeoffs of various architectural choices.

4.3.6 Answers to Seminar Questions

Q1 *Can traditional techniques developed for federations of relational databases be enriched with RDF semantics, and thus provide effective and efficient solutions to problems of federated semantic data management?*

Although it is possible to draw inspiration from databases to a certain extent, these techniques are not always directly applicable. One instance of this is how policies bring additional semantics that can be very naturally captured using Semantic Web technologies. Other topics such as the open nature of the Web which were mentioned by T1 also need to be considered.

Q2 *What problems of federated semantic data management present new research challenges that require the definition of novel techniques?*

Inclusion of policies in the overall architecture has not been addressed. We need to look into semantic specification, modelling, enforcement and inference, and general implications for federation engines e.g. when planning query execution.

Q3 *What is the role of RDF semantics in the definition of the problems of federated semantic data management?*

Policies always have semantics, and to leverage this in the semantic federation engine, RDF specific entailment needs to be explored (hinting at the topic of the second working group). The Semantic Web is also an opportunity for policy interoperability and smooth integration into the query planner (here we mean policy in a broad sense as in it can also be used for planning).

4.4 Federated Semantic Data Management: Use Cases and Applications

Michel Dumontier (Maastricht University, NL), Sören Auer (Universität Bonn, DE), Jana Hentschke (Deutsche Nationalbibliothek – Frankfurt am Main, DE), Pascal Molli (University of Nantes, FR), and Ana Maria Roxin (Universite de Bourgogne, FR)

The main objective of the group was to articulate a vision, benefits, use cases, and current limitations facing federated semantic data management (FDSM). The group consisted of 5 members: Sören Auer (University of Hannover), Ana Maria Roxin (University of Burgundy), Jana Hentschke (Deutsche Nationalbibliothek), Pascal Molli (Nantes University), and Michel Dumontier (Maastricht University).

Our group envisioned that Federated Semantic Data Management would enable access to structured information across heterogeneous, distributed knowledge sources in an accurate, reliable, and performant manner that will usher a new era of research and innovation. FSDM enables two main approaches to querying: (A) Explorative, open domain querying, where users are able to query the accessible web and adapt to its continuous evolution by including new relevant sources, enriching queries with relevant attributes, and suggesting improved queries based on similar federated queries. (B) Controlled, closed domain querying, in which specific data sources are used, queries are optimized for performance, and quality assessment is performed through constraint satisfaction, and that accuracy and recall query results are used in workflows and have real world applications. Our group established a framework for developing use cases. FSDM should enable a broad set of use cases including: i) the automatic discovery and querying of newly published knowledge sources, ii) the ability to answer previously unanswered questions, iii) the automatic, but parameterized gathering of more relevant data to strengthen statistical analyses, iv) the ability to perform real time fact checking, vi) the catalyst for marketplace of queries and their answers, and vii) the discovery of subtle, but important findings obtained through the analysis of massively distributed knowledge sources.

Our group established a framework for further developing specific use cases. This framework requires that use cases addresses aspects of query formulation, query execution, result generation, stakeholders, social, legal, ethical aspects, performance and availability, change management, and quality considerations. We used this framework to develop use cases to illustrate explorative open domain querying as well as controlled, closed domain querying.

We also addressed the 3 main questions of the seminar. We argued that while there may be research in relational database federation that FSDM can learn from, it is open world reasoning and inconsistency management that offer a tantalizing opportunity to move beyond the relational model, although this has yet to be fully explored. We indicate that the main problems of FSDM that require novel techniques include combinations of expressive logics,

synchronicity, and optimizations. Finally, we believe that the role of RDF semantics is to use the open world assumption to reason over an unbounded knowledge graph, and perhaps that this could help develop more advanced artificial intelligence systems.

Participants

- Maribel Acosta
KIT – Karlsruher Institut für Technologie, DE
- Bernd Amann
University Pierre & Marie Curie – Paris, FR
- Sören Auer
Universität Bonn, DE
- Abraham Bernstein
Universität Zürich, CH
- Piero Andrea Bonatti
University of Naples, IT
- Carlos Buil-Aranda
TU Federico Santa María – Valparaíso, CL
- Emanuele Della Valle
Polytechnic University of Milan, IT
- Michel Dumontier
Maastricht University, NL
- Johann-Christoph Freytag
Humboldt-Universität zu Berlin, DE
- Claudio Gutierrez
University of Chile – Santiago de Chile, CL
- Peter Haase
Metaphacts GmbH – Walldorf, DE
- Olaf Hartig
Linköping University, SE
- Jana Hentschke
Deutsche Nationalbibliothek – Frankfurt am Main, DE
- Katja Hose
Aalborg University, DK
- Sabrina Kirrane
Wirtschaftsuniversität Wien, AT
- Stasinos Konstantopoulos
Demokritos – Athens, GR
- Jorge Lobo
UPF – Barcelona, ES
- Pascal Molli
University of Nantes, FR
- Gabriela Montoya
Aalborg University, DK
- Themis Palpanas
Paris Descartes University, FR
- Ana Maria Roxin
Universite de Bourgogne, FR
- Sherif Sakr
KSAU – Riyadh, SA
- Juan F. Sequeda
Capsenta Inc. – Austin, US
- Hala Skaf-Molli
University of Nantes, FR
- Rudi Studer
KIT – Karlsruher Institut für Technologie, DE
- Joachim Van Herwegen
Ghent University, BE
- Maria-Esther Vidal
Universidad S. Bolivar – Caracas, VE