# Semantic Data Management in Practice

## Tutorial Description

Olaf Hartig
Linköping University
Linköping, Sweden
olaf.hartig@liu.se

Olivier Curé
Université Paris-Est Marne la Vallée
Paris, France
olivier.cure@u-pem.fr

## ABSTRACT

After years of research and development, standards and technologies for semantic data are sufficiently mature to be used as the foundation of novel data science projects that employ semantic technologies in various application domains such as bio-informatics, materials science, criminal intelligence, and social science. Typically, such projects are carried out by domain experts who have a conceptual understanding of semantic technologies but lack the expertise to choose and to employ existing data management solutions for the semantic data in their project. For such experts, including domain-focused data scientists, project coordinators, and project engineers, our tutorial delivers a *practitioner's guide to semantic data management*. We discuss the following important aspects of semantic data management and demonstrate how to address these aspects in practice by using mature, production-ready tools: i) storing and querying semantic data; ii) understanding, iii) searching, and iv) visualizing the data; v) automated reasoning; vi) integrating external data and knowledge; and vii) cleaning the data.

[500]General and reference Surveys and overviews

## Keywords

Semantic Technologies; RDF; Storage; Querying; Search; Cleaning; Visualization; Reasoning

## 1. INTRODUCTION

The term *semantic data* refers to data whose meaning has been made explicit in the form of meta-data. Such meta-data may then be used in semantics-based approaches to manage the data. The perhaps most prevalent approach to represent semantic data and its meta-data is based on the Resource Description Framework (RDF) [7] and a family of related standards proposed by the World Wide Web Consortium (W3C), e.g., SPARQL, RDFS and OWL. Today, these standards and various software implementations that support them can be considered sufficiently mature to be used as a foundation of projects that aim to apply semantic technologies in a broad variety of domains. Examples of such projects are ValCri[1] (visual analytics for sense-making in criminal intelligence), Waves[2] (management of potable water networks), and Graphe Culture[3](management of knowledge graphs related to activities of the French ministry of culture and communication).

Practitioners who aim to conduct such an application project typically are experts in the application domain, and they may have a conceptual understanding of semantic technologies and how these technologies should be put to use to achieve the goals of the project. However, these experts may not have the knowledge and experience to address the various aspects of data management that typically have to be addressed in such projects. Based on our experience with such projects and on interviews with other practitioners, we have identified seven aspects that present the most prominent stumbling blocks in many application projects. In the tutorial we discuss these aspects and provide practical guidance on how these aspects can be addressed by using mature, production-ready tools and systems. To deepen the practical nature of the tutorial we use the aforementioned Waves project as a running example based on which we demonstrate the application of concepts and tools. This project aims to support the analysis of semantic data streams (typically coming from sensors of the Internet of Things) in an application domain focused on the management of potable water networks.

## 2. CONTENT AND OUTLINE

In this section we describe the seven aspects of semantic data management that the tutorial covers and, for each of them, outline the discussion and guidance that we deliver in the tutorial.

### 2.1 Storing and Querying the Data

Persistently storing data and executing declarative queries over it are among the most important aspects of managing data. Systems that provide such functionality for RDF-based semantic data either use an existing database management system (DBMS), for instance based on the relational model, e.g. PostgreSQL, or are designed from scratch, usually as a graph store. For data sets that can be handled on


*WWW'17 Companion,* April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3051096

[1]http://valcri.org/
[2]http://www.waves-rsp.org/
[3]http://cblog.culture.fr/projet/2013/11/07/groupe-de-travail-metadonnees-culturelles/

a single machine, a centralized architecture is usually preferred (e.g., RDF-3X [25], Hexastore [35], SW-Store [1]), but a distributed architecture can be adopted as well (e.g., [15], TriAD [12]).

Such systems are usually called *triple stores*, and the standard declarative query language for RDF that they support is SPARQL [13]. In the case of a relational database management system storage back-end these queries are automatically translated into SQL queries. Otherwise, they are compiled and optimized using a dedicated system.

Prominent, mature triple stores include Virtuoso[4], MarkLogic[5], Blazegraph[6], GraphDB[7] (formerly OWLIM) [4], Oracle[8], AllegroGraph[9], and Stardog[10]. In the tutorial we provide an overview of these mature triple stores, discuss their specific features, and, for 1–2 of them, demonstrate how they can be used (e.g., how to set them up, how to load data and how to run queries). In this context, we look not only at terminal-based and programming-language interfaces, but also at system-specific administration tools.

## 2.2 Understanding the Data

A typical problem for many practitioners who want to use a given set of semantic data is to obtain an initial understanding of the data set (e.g., what types of entities does the data set describes, what vocabularies are used to represent properties of entities and relationships among them?). We introduce the tutorial attendees to RDF-focused data summarization and data profiling tools such as ExpLOD [16], LODSight[11] [11], Loupe[12] [22], and ProLOD++[13] [2], which can be used to get such an initial understanding. Additionally, we introduce ontology visualization tools such as WebVOWL[14] [19] and Protégé[15] [31] based on which it is possible to explore the ontologies as used by the data set.

## 2.3 Searching the Data

In addition (or, as an alternative) to declarative queries, many semantic data projects adopt keyword search as a way to explore and to query the data set(s) involved in the project. To support such use cases most production-ready triple stores come with a built-in full-text search engine. In addition to this feature, some triple stores provide built-in functionality to integrate an external search engine such as Solr[16] and Elasticsearch[17]. In both cases, the typical approach to enable users to issue keyword (and perhaps more expressive information retrieval) queries is via special, vendor-specific predicates used in SPARQL queries. The tutorial provides an overview of these features. Additionally, the tutorial discusses options for how a dedicated search engine such as Solr or Elasticsearch can be employed for semantic data use cases *separately* from a triple store.

## 2.4 Visualizing the Data

Many semantic data projects involve the development of software applications (often, Web applications) in which the visualization of data is a key feature. While such applications typically target users that are not part of the project, data visualizations may also be used as a powerful tool within projects, where it may help data analysts to derive new insights by visually exploring data sets. We note that there exists a wealth of data visualization software that does not specifically focus on semantic data but that may be of great help for achieving the goals of semantic data projects. In the tutorial we showcase how some of this software has been employed in the aforementioned Waves project, and we provide pointers to how semantic data can be dealt with when implementing a software application. Additionally, based on recent literature surveys [9, 3, 8], we give a brief overview of data visualization techniques and tools that have been developed specifically for visualizing semantic data.

## 2.5 Automated Reasoning

A distinguishing feature of semantic data is its accompanying meta-data that describes the meaning of the data. This meta-data enables automated reasoning processes to derive data that is given implicitly by a semantic data set and its meaning, but that has not been expressed directly. In order to obtain a complete answer set to a given query, the reasoning processing can be performed either at the data loading or query run-times. In the former, all logical consequences are materialized in the data set. This impacts negatively the loading time and the size of the persisted database but ensures fast query processing. In the latter, all the reasoning machinery is performed at query run-time to produce a rewriting of the original query. Compared to the materialization approach, the query rewriting solution is thus characterized by a slower query processing by a faster data set loading time and a smaller persisted database. The tutorial provides an overview of these features and how they can be used in production-ready triple stores as well as other systems such as WaterFowl [6], RDFox [23], Inferray [30]. Additionally, we introduce tools that can be employed to materialize derived data, including tools such as WebPIE [32], that scale to very large data sets since they are built on Big Data processing frameworks such as Apache Hadoop[18] or Apache Spark[19].

## 2.6 Integrating Data from Multiple Sources

Many application projects require to combine data and knowledge from multiple sources. Such an integration process is one of the major use cases of semantic technologies. This is largely due to the availability of a large repository of data sets, knowledge bases, and ontologies via Websites and initiatives such as the Datahub[20] and Linkeddata.org[21].

The peculiarity of this integration process is the presence of ontologies. As presented in [28], several dedicated approaches have been proposed. They can be distinguished

[4] http://virtuoso.openlinksw.com
[5] http://www.marklogic.com/
[6] http://www.blazegraph.com/
[7] http://ontotext.com/products/graphdb/
[8] http://www.oracle.com/technetwork/database-options/spatialandgraph/overview/rdfsemantic-graph-1902016.html
[9] http://franz.com/agraph/allegrograph/
[10] http://stardog.com/
[11] http://lod2-dev.vse.cz/lodsight-v2/about.html
[12] http://loupe.linkeddata.es/loupe/
[13] https://hpi.de/naumann/projects/data-profiling-and-analytics/prolod.html
[14] http://vowl.visualdataweb.org/webvowl.html
[15] http://protege.stanford.edu/products.php
[16] http://lucene.apache.org/solr/
[17] https://www.elastic.co/products/elasticsearch
[18] http://hadoop.apache.org/
[19] http://spark.apache.org/
[20] https://datahub.io
[21] http://linkeddata.org/

on the availability or absence of a shared ontology. If such a general ontology ( e.g., SUMO (Suggested Upper Merged Ontology)[27]), exists, it is extended to relate external ontologies via some mappings. In its absence, heuristics-based or machine learning techniques are generally used, e.g., GLUE [10]. We briefly recall the main concepts of data integration in this context. Thereafter, we focus on demonstrating how to integrate semantic data by using a number of tools such as the Silk framework[22] [34], Karma[23] [17], LIMES[24] [26] and RDF Refine[25] [20] that have been developed specifically for semantic data.

## 2.7 Cleaning the Data

When starting to work with data, analysts often observe various quality issues. Some of these issues may be specific to the form in which the data is represented and accessed (e.g., encoding problems, syntax errors, wrongly used vocabularies, unavailable servers); other issues may be inherent in the data such as inaccuracies, inconsistencies, and undesired duplicates. Detecting such issues and removing them—a process called data cleaning (or data cleansing) [29, 24]—is crucial for the success of many data-related projects. A recent survey discusses research approaches for detecting quality issues in the context of Semantic Web data [36]. We describe the most prominent of these approaches in the tutorial, and demonstrate related tools such as RDFUnit[26] [18] and Sieve[27] [21]. Additionally, we demonstrate how quality issues cannot only be detected but also resolved by using OpenRefine[28][33] and Trifacta Wrangler[29], which are powerful tools for exploring data sets, discovering outliers, clustering and reconciling data records, transforming data, etc.

# 3. PRESENTERS

**Olaf Hartig** is an Assistant Professor at the Department of Computer and Information Science (IDA) of Linköping University. Olaf holds a Ph.D. in Computer Science from the Humboldt-Universität zu Berlin, Germany. His research interests are related to various areas of data management with a particular focus on Web data, graph data, and semantic data management. He has published 1 book [14], 2 book chapters, 5 journal articles, and 15 research papers in top international conferences in the fields of the Semantic Web and Databases. Moreover, Olaf presented 7 tutorials at such conferences including WWW 2010, WWW 2013, and ICDE 2014; and he was lecturer at the 2011 Indian-Summer School on Linked Data.

**Olivier Curé** is a tenured associate professor in Computer Science at the University of Paris-Est Marne la Vallée (UPEM) in France. He obtained his Ph.D. in Artificial Intelligence at the Université Paris V, France. His research interests are data and knowledge base management systems, semantic information and reasoning. He has published 1 book [5], 4 book chapters, 12 journal papers, and over 60 research papers in international, peer-reviewed conferences on data and knowledge bases, Semantic Web, and Big Data.

[22] http://silkframework.org/
[23] http://usc-isi-i2.github.io/karma/
[24] http://aksw.org/Projects/LIMES.html
[25] http://refine.deri.ie/
[26] http://rdfunit.aksw.org/
[27] http://sieve.wbsg.de/
[28] http://openrefine.org/
[29] https://www.trifacta.com/products/wrangler/

# 4. ACKNOWLEDGMENTS

We would like to thank a number of people with whom we discussed various aspects of the topics covered in the tutorial. These discussions have been tremendously helpful for designing the tutorial. Our thanks go to: Eva Blomqvist, Robin Keskisärkkä, Valentina Ivanova, Jeremy Lhez, Badre Belabbess and Xiangnan Ren.

Olaf Hartig's work on this tutorial has been funded partially by the CENIIT program at Linköping University (Sweden), project no. 17.04.

Olivier Curé's work on this tutorial has been funded partially by the FUI (Fonds Unique Interministériel) 17 Waves project.

# 5. REFERENCES

[1] D. J. Abadi, A. Marcus, S. Madden, and K. Hollenbach. SW-Store: A Vertically Partitioned DBMS for Semantic Web Data Management. *VLDB Journal*, 18(2):385–406, 2009.

[2] Z. Abedjan, T. Grütze, A. Jentzsch, and F. Naumann. Profiling and mining RDF data with ProLOD++. In *Proceedings of the IEEE 30th International Conference on Data Engineering (ICDE)*, pages 1198–1201, 2014.

[3] N. Bikakis and T. K. Sellis. Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference*, 2016.

[4] B. Bishop, A. Kiryakov, Z. Tashev, M. Damova, and K. I. Simov. OWLIM Reasoning over FactForge. In *Proceedings of the 1st International Workshop on OWL Reasoner Evaluation (ORE)*, 2012.

[5] O. Curé and G. Blin. *RDF Database Systems: Triples Storage and SPARQL Query Processing*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2014.

[6] O. Curé, G. Blin, D. Revuz, and D. C. Faye. WaterFowl: A Compact, Self-Indexed and Inference-Enabled Immutable RDF Store. In *Proceedings of the 11th Extended Semantic Web Conference (ESWC)*, pages 302–316, 2014.

[7] R. Cyganiak, D. Wood, M. Lanthaler, G. Klyne, J. J. Carroll, and B. McBride. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, Feb. 2014.

[8] A. Dadzie and E. Pietriga. Visualisation of Linked Data - Reprise. *Semantic Web*, 8(1):1–21, 2017.

[9] A. Dadzie and M. Rowe. Approaches to Visualising Linked Data: A Survey. *Semantic Web*, 2(2):89–124, 2011.

[10] A. Doan, J. Madhavan, P. M. Domingos, and A. Y. Halevy. Learning to map between ontologies on the semantic web. In *Proceedings of the Eleventh International World Wide Web Conference, WWW 2002, May 7-11, 2002, Honolulu, Hawaii*, pages 662–673, 2002.

[11] M. Dudás, V. Svátek, and J. Mynarz. Dataset Summary Visualization with LODSight. In *The Semantic Web: ESWC 2015 Satellite Events*, pages 36–40, 2015.

[12] S. Gurajada, S. Seufert, I. Miliaraki, and M. Theobald. TriAD: A Distributed Shared-nothing RDF Engine Based on Asynchronous Message Passing. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 289–300, New York, NY, USA, 2014. ACM.

[13] S. Harris, A. Seaborne, and E. Prud'hommeaux. SPARQL 1.1 Query Language. W3C Recommendation, Mar. 2013.

[14] O. Hartig. *Querying a Web of Linked Data - Foundations and Query Execution*, volume 24 of *Studies on the Semantic Web*. IOS Press, 2016.

[15] J. Huang, D. J. Abadi, and K. Ren. Scalable SPARQL Querying of Large RDF Graphs. *PVLDB*, 4(11):1123–1134, 2011.

[16] S. Khatchadourian and M. P. Consens. ExpLOD: Summary-Based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. In *Proceedings of the 7th Extended Semantic Web Conference (ESWC)*, pages 272–287, 2010.

[17] C. A. Knoblock, P. A. Szekely, J. L. Ambite, A. Goel, S. Gupta, K. Lerman, M. Muslea, M. Taheriyan, and P. Mallick. Semi-Automatically Mapping Structured Sources into the Semantic Web. In *Proceedings of the 9th Extended Semantic Web Conference (ESWC)*, pages 375–390, 2012.

[18] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-Driven Evaluation of Linked Data Quality. In *Proceedings of the 23rd International World Wide Web Conference (WWW)*, pages 747–758, 2014.

[19] S. Lohmann, S. Negru, F. Haag, and T. Ertl. Visualizing Ontologies with VOWL. *Semantic Web*, 7(4):399–419, 2016.

[20] F. Maali, R. Cyganiak, and V. Peristeras. Re-using Cool URIs: Entity Reconciliation Against LOD Hubs. In *Proceedings of the WWW2011 Workshop on Linked Data on the Web (LDOW)*, 2011.

[21] P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: Linked Data Quality Assessment and Fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123, 2012.

[22] N. Mihindukulasooriya, M. Poveda-Villalón, R. García-Castro, and A. Gómez-Pérez. Loupe - An Online Tool for Inspecting Datasets in the Linked Data Cloud. In *Proceedings of the ISWC 2015 Posters & Demonstrations Track*, 2015.

[23] B. Motik, Y. Nenov, R. E. F. Piro, and I. Horrocks. Incremental Update of Datalog Materialisation: The Backward/Forward Algorithm. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 1560–1568, 2015.

[24] H. Müller and J.-C. Freytag. Problems, Methods, and Challenges in Comprehensive Data Cleansing. Technical Report HUB-IB-164, Humboldt University Berlin, 2003.

[25] T. Neumann and G. Weikum. RDF-3X: A RISC-Style Engine for RDF. *PVLDB*, 1(1):647–659, 2008.

[26] A. N. Ngomo and S. Auer. LIMES – A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2312–2317, 2011.

[27] I. Niles and A. Pease. Towards a standard upper ontology. In *FOIS*, pages 2–9, 2001.

[28] N. F. Noy. Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, 33(4):65–70, 2004.

[29] E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23(4):3–13, 2000.

[30] J. Subercaze, C. Gravier, J. Chevalier, and F. Laforest. Inferray: Fast In-Memory RDF Inference. *PVLDB*, 9(6):468–479, Jan. 2016.

[31] T. Tudorache, N. F. Noy, S. W. Tu, and M. A. Musen. Supporting Collaborative Ontology Development in Protégé. In *Proceedings of the 7th International Semantic Web Conference (ISWC)*, pages 17–32, 2008.

[32] J. Urbani, S. Kotoulas, J. Maassen, F. van Harmelen, and H. E. Bal. WebPIE: A Web-scale Parallel Inference Engine using MapReduce. *Journal of Web Semantics*, 10:59–75, 2012.

[33] R. Verborgh and M. De Wilde. *Using OpenRefine*. Packt Publishing, 1st edition, 2013.

[34] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Silk – A Link Discovery Framework for the Web of Data. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web (LDOW)*, 2009.

[35] C. Weiss, P. Karras, and A. Bernstein. Hexastore: Sextuple Indexing for Semantic Web Data Management. *PVLDB*, 1(1):1008–1019, 2008.

[36] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment for Linked Data: A Survey. *Semantic Web*, 7(1):63–93, 2016.