Creating Knowledge out of Interlinked Data: The LOD2 Tool Stack

Sebastian Tramp, Bert Van Nuffelen, and Sören Auer

Universität Leipzig, Institut für Informatik, AKSW,
Postfach 100920, D-04009 Leipzig, Germany,
`{lastname}@informatik.uni-leipzig.de`
`http://aksw.org`

Abstract. This joint tutorial of the consortium of the FP7 project *LOD2 – Creating Knowledge out of Interlinked Data* will give an overview on the area of creating, managing and using Linked Data sources. As a prerequisite to the main tutorial part, we give an overview of the life cycle of Linked Data usage and its challenges as well as the LOD2 software stack, which is an integrated distribution of aligned tools supporting the life-cycle of Linked Data from extraction, authoring/creation over enrichment, interlinking, fusing to visualization and maintenance. Based on this more foundational explanations, we present a detailed insight into the usage of LOD2 stack both with practical and non-practical parts. More specifically, we present tools and usage scenarios for the following Linked Data life cycle tasks: *extraction* (Triplify / D2R), *storage and querying* (Openlink Virtuoso), *authoring* (OntoWiki), *interlinking* (Silk / LIMES) as well as *enrichment and repair* (ORE). In addition to that, we present the LOD2 stack demonstrator which integrates these and other LOD2 tools for a better user experience.

1 Introduction

LOD2 is a large-scale integrating project co-funded by the European Commission within the FP7 Information and Communication Technologies Work Programme (Grant Agreement No. 257943). Commencing in September 2010, this 4-year project comprises leading Linked Open Data technology researchers, companies, and service providers (15 partners) from across 11 European countries (and one associated partner from Korea) and is coordinated by the AKSW research group at University of Leipzig.

The goals of the LOD2 project include the development of enterprise-ready tools and methodologies for exposing and managing very large amounts of structured information on the Data Web and adaptive tools for searching, browsing, and authoring of Linked Data. These tools are
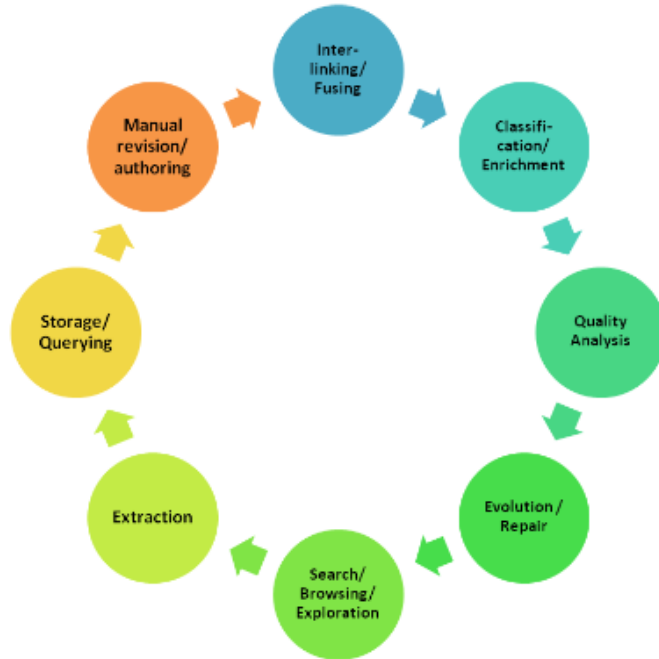
Fig. 1. Linked Data Lifecycle

distributed as the LOD2 stack. The LOD2 stack is an integrated distribution of aligned tools which support the life-cycle of Linked Data (depicted in Figure 1) from extraction, authoring/creation over enrichment, inter-linking, fusing to visualization and maintenance. An holistic presentation of the LOD2 Stack, the Linked Data lifecycle management stages and the tools supporting these stages will benefit the Semantic Web community a lot. Also, the LOD2 Stack (being based on the Debian packaging system) is planned as an open environment, where additional and complementary tools can be easily added thus supporting complex Linked Data management workflows.

2 The Tutorial

The tutorial will give an overview to the complete LOD2 stack and a detailed insight to the integrated tools in Sec.4. For each tool we will give a talk followed by a practical part. Each practical part is guided by an overall usage scenario which covers most parts of the Linked Data life cycle. For the practical parts, we provide a virtual machine where

the LOD2 stack and additional material is already installed (slides, data sources, software packages).

The intended audience covers knowledge workers from the industry which need to integrate and use Linked Data from different open and non-open data sources as well as researchers and PhD students which are new to the area of Linked Data or individual Linked Data management aspects.

This is a full day tutorial and it was never given before on other venues.

3   The Presenters

Dr. Sören Auer leads AKSW research group at Universität Leipzig. His research interests are centered around semantic data web technologies. Sören is author of over 70 peer-reviewed scientific publications. Sören coordinates the EU-FP7-ICT research project LOD2 and is (co-)founder of several high-impact research and community projects such as DBpedia, OntoWiki and LinkedGeoData. He is co-organiser of workshops, programme chair international conferences, area editor of the Semantic Web Journal, serves as an expert for industry, EC, W3C and advisory board member of the Open Knowledge Foundation.

Dr. Bert Van Nuffelen has a PHD in Computer Science, Knowledge Representation from the KULeuven, Belgium (2004). In his research he worked on improving the efficiency of the inference engine for Abductive Constraint Logic Programs [5]. This resulted in a system, called the Asystem, which applied Constraint Programming techniques to the inference engine. In 2005, he joined Mission Critical a Belgian software company, where he developed applications for insurance and financial institutes. These applications were based on Mission Critical's ODASE platform. The core of the ODASE based software design process are the business models expressed as OWL ontologies By interpreting and the reasoning capabilities of the platform applications where build in an agile way. In 2011 he joined Ten-Force where is currently active as semantic technologies software architect and he is the TenForce responsible for the European LOD2 project. There he applies his expertise and experience in developing reasoning engines, RDF stores design, ontology modeling and overall application design for publishing Linked Data.

Sebastian Tramp studied information science and is a research associate at AKSW / University of Leipzig since 2006. He is author of over 30 peer-reviewed publications and presented his research projects on conferences

as the ESWC and the ISWC as well as multiple workshops. His main research interest are Distributed Semantic Social Networks. He is a lecturer at the University of Leipzig and the Leipzig School of Media.

4   The LOD2 Stack

The LOD2 stack comprises new and substantially extended existing tools from the LOD2 partners and third parties. The LOD2 stack is organized as a Debian package repository making the tool stack easy to install on any Debian-based system (e.g. Ubuntu). In our tutorial, we present tools and usage scenarios for the following Linked Data life cycle tasks:

Extraction Triplify [1] provides a building block for the "semantification" of Web applications. Triplify is a small plugin for Web applications, which reveals the semantic structures encoded in relational databases by making database content available as RDF, JSON or Linked Data.

The *D2R Server* [3] is a tool for publishing the content of relational databases on the Semantic Web. D2R Server uses a customizable D2RQ mapping to map database content into RDF triple, and allows the RDF knowledge bases to be browsed and searched - the two main access paradigms to the Semantic Web.

Storage & Querying Virtuoso [4] is an innovative industry standards compliant platform for native data, information, and knowledge management. It implements and supports a broad spectrum of query languages, data access interfaces, protocols, and data representation formats that includes SPARQL, RDF, RDFa and many more. The open-source edition of Virtuoso, which includes a scalable high-performance RDF Quad Store, is be the basis for the LOD2 Stack's knowledge store.

Authoring & Manual Revision OntoWiki [2] is a tool providing support for agile, distributed knowledge engineering scenarios. It facilitates the visual presentation of a knowledge base as an information map, with different views on instance data. It enables intuitive authoring of semantic content, with an inline editing mode for editing RDF content, similar to WYSIWIG for text documents. OntoWiki provides sophisticated means for navigating, visualising and authoring of RDF-based Knowledge Bases. It serves and consumes Linked Data and comprises a comprehensive middleware API for building custom Semantic Web applications.

Interlinking & Fusion The *Silk Link Discovery Framework* [8] supports data publishers in interlinking RDF knowledge bases. Using the declarative Silk - Link Specification Language (Silk-LSL), developers can specify which types of RDF links should be discovered between data sources as well as which conditions data items must fulfill in order to be interlinked. These link conditions may combine various similarity metrics and can take the graph around a data item into account, which is addressed using an RDF path language.

LIMES [7] is a link discovery framework for the Web of Data. It implements time-efficient approaches for large-scale link discovery based on the characteristics of metric spaces. It is easily configurable via a web interface. It can also be downloaded as standalone tool for carrying out link discovery locally.

Enrichment & Repair ORE [6] is a tool for repairing and enriching OWL ontologies. State-of-the-art methods in ontology debugging and supervised machine learning form the basis of ORE and are adapted or extended so as to work well in practice. ORE supports the detection of a variety of ontology modelling problems and guides the user through the process of resolving them. Furthermore, the tool allows to extend an ontology through (semi-)automatic supervised learning. A wizard-like process helps the user to resolve potential issues after axioms are added.

References

1. Sören Auer, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann, and David Aumueller. Triplify: Light-weight linked data publication from relational databases. In *WWW 2009*, pages 621–630. ACM, 2009.
2. Sören Auer, Sebastian Dietzold, and Thomas Riechert. OntoWiki - A Tool for Social, Semantic Collaboration. In *ISWC 2006*, volume 4273 of *LNCS*. Springer, 2006.
3. Christian Bizer. D2R MAP - A Database to RDF Mapping Language, 2003.
4. Orri Erling and Ivan Mikhailov. RDF Support in the Virtuoso DBMS. In *CSSW 2007*, volume 113 of *LNI*, pages 59–68. GI, 2007.
5. Antonis C. Kakas, Bert Van Nuffelen, and Marc Denecker. A-System: Problem Solving through Abduction. In *IJCAI 2001*, pages 591–596. Morgan Kaufmann, 2001.
6. Jens Lehmann and Lorenz Bühmann. ORE - A Tool for Repairing and Enriching Knowledge Bases. In *ISWC 2010*, LNCS. Springer, 2010.
7. Axel-Cyrille Ngonga Ngomo and Sören Auer. LIMES A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In *IJCAI 2011*, pages 2312–2317. IJCAI/AAAI, 2011.
8. Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and Maintaining Links on the Web of Data. In *ISWC 2009*, volume 5823 of *LNCS*, pages 650–665. Springer, 2009.