Collaborative Project

LOD2 – Creating Knowledge out of Interlinked Data

Project Number: 257943 **Start Date of Project:** 01/09/2010 **Duration:** 48 months

Annual Report 2010

In recent times, Linked Data has emerged as a paradigm for integrating and publishing structured data on the Web. Yet, the generation of Linked Data is merely the beginning of a process that seeks to alleviate the potential for data and information overload. Organizations across government and industry embracing Linked Data ultimately seek to produce actionable knowledge that provides foundation for attaining agility levels critical for information age success. Thus, the goal of the LOD2 project is to develop additional infrastructure technology and best practices that fill the chasm between structured-linked-data and applied model logic & reasoning, en route to redefining the Web as we know it. In doing so, LOD2 will integrate and syndicate linked data with large-scale existing applications and showcase the benefits in three application scenarios including Media & Publishing, Corporate Data Intranets and eGovernment.

Website
http://lod2.eu/
http://lod2.eu/BlogPost/
http://twitter.com/lod2project

Project co-funded by the European Commission within the Seventh Framework Programme (2007 – 2013)

Summary of Activities

During the first four months of the project, LOD2 has executed activities and made achievements in the following areas:

- requirements elicitation for the three major use cases, e.g. through the description of concrete use case scenarios and the execution of an online survey,
- state-of-the-art analysis regarding the development and deployment of Linked Data technologies through literature reviews, online surveys among tool providers and the testing of technical enhancements in existing software,
- the development of an "LOD2 roadmap" that identifies deficiencies and white spots of existing technologies and indicates how they can be solved from an LOD2 perspective,
- documentation of tools and software services involved in LOD2 and preparing the Debian packaging of the LOD2 stack,
- start of the development, testing and integration of functional architecture for the LOD2 stack prototype, which provides a flexible technological basis for developing customized semantic social software applications.

Through the next year the LOD2 project will provide a first release of the LOD2 technology stack on which concrete applications in media, enterprise and e-government will build. Likewise, PubLink set up as Linked Open Data Starter Service by the LOD2 consortium at the beginning of the project will take up its work; aiming to substantially extend the publication and use of Linked Data within and beyond the focus domains.

Important Work Area

1. Major Use Cases in LOD2

The LOD2 project revolves around three major use cases in the fields of media & publishing, corporate data intranets and e-government. The use cases serve to develop, test and deploy the technological infrastructure necessary for the integration and syndication of linked data with large-scale existing applications and will thus showcase the benefits provided by LOD2 in a range of ways.

i) Media & Publishing

Large amounts of data resources from the legal domain are used to test and explore the commercial value of linked data in media and publishing. This data will be interlinked and merged automatically. Data from external sources will be used to semantically enrich the existing datasets. Adequate licensing and business models are also investigated with respect to the management of interoperable metadata.

ii) Enterprise Data Web

Linked Data is a natural addition to the existing document and web service intranets and extranets. Corporate data intranets based on Linked Data technologies can help to substantially reduce data integration costs. Using the LOD2 stack for linking internal corporate data with external references from the LOD cloud will allow a corporation to significantly increase the value of its corporate knowledge with relatively low effort.

iii) Linked Governmental Data

The project will showcase the wide applicability of the LOD2 Stack through the design, specification, implementation, testing and user evaluation of a case study targeting ordinary citizens of the European Union. LOD2 will establish a network of European governmental data registries in order to increase public access to high-value, machine-readable data sets generated by European, national as well as regional governments and public administrations. The semi-automatic classification, interlinking, enrichment and repair methods developed in LOD2 will create a significant benefit, since they allow governmental data to be more easily explored, analyzed and mashed together.

2. Requirements, Design and LOD2 Stack Prototype

This work area focuses on the identification of requirements for the three major use cases carried out within the LOD2 project. To get a clear picture of the concrete deployment environment for the Linked Data technologies, requirements have mainly been elicited by defining general application areas and elaborating on the technical and organizational affordances in which the technologies will be deployed. So far reasonable services, software components and processes have been identified for the media and publishing case. The schema also elicits non-technological aspects such as organizational issues (role models, workflows), impacts and risks that might evolve when implementing Linked Data in existing infrastructures. This schema serves as a blueprint for the requirements elicitation in the Enterprise Use Case and in the Open Government Data Use Case, thus helping to describe requirements in a unified and comparable way. Moreover, an online survey has been set up to collect input from a broader range of stakeholders interested in a semantically powered Open Government Data catalogue, while various use case scenarios have been collected for the Enterprise Use Case.

To make progress regarding the technological infrastructure necessary to realize all proposed functionalities and services, a state-of-the-art analysis has been conducted on the basis of a literature review and an online survey among relevant tool providers. The analysis discusses critical aspects in the development and deployment of Linked Data technologies along the defined use cases. The following five areas have been of interest: 1) Storing and Querying Large Knowledge Bases, 2) Knowledge Base Creation, Enrichment and Repair, 3) Reuse, Interlinking and Knowledge Fusion, 4) Web Interfaces for Browsing and Authoring, and 5) Metadata Economics. These areas have been analyzed according to the functionalities and technical specifications of existing technologies and standards, while deficiencies and white spots of existing technologies and how they can be bridged from an LOD2 perspective have been identified. This analysis has led to the conception of a roadmap which helps to define the technological specifications and to model the architecture and the system design. Likewise, the technical specifications of all tools and software services involved in LOD2 are documented in a comparable and standardized fashion. In terms of feasibility the goal is to compile a set of well-integrated Debian packages which shall make it easier for end users to deploy, implement and adapt the various software components.

3. Storing and Querying Very Large Knowledge Bases

Another focus of the LOD2 project is on (i) the upgrade and deployment of the initial LOD2 public cloud – a Virtuoso hosted digest of all major datasets available on the emerging Web of Data, (ii) on the state of the art in RDF data management and on (iii) benchmarking available tools for RDF data management.

Regarding (i), the public LOD cloud has been enhanced by adding new datasets and features. A new viewing modality was introduced with Microsoft Silverlight Pivot viewer to enhance, amongst others, retrieval and faceted browsing. In the future more datasets will be deployed in the public cloud, which also requires LOD2 researchers at NUIG, hosting the datasets on its Webstar cluster, to potentially upgrade its hardware. The LOD Cloud hosting is expected to move to Virtuoso 7 during the first quarter of 2011, featuring columnar compression and dramatically improved space efficiency and thus greatly enhancing the service level of the present LOD deployment.

Regarding (ii), the interest is twofold, namely establishing the state of the art in RDF storage/query engines as well as the state-of-the-art in large-scale RDF storage/query benchmarking. A survey among companies and organizations producing RDF engines has been conducted to identify holes in existing benchmarks for RDF engines. The questionnaire also aims to promote practices for vendor publishing of benchmark results with well-defined reporting and verification processes.

Regarding (iii), the aim is to create benchmarking environments that will provide a level playing field in order to compare diverse RDF storage implementations (the LOD2 state-of-the-art laboratory). Work has been focused on new benchmarks. A new extension, which focuses on Business Intelligence questions, has been drafted for the Berlin SPARQL Benchmark (BSBM-BI). Similarly, there is a discussion on a future information retrieval benchmark. It is the goal to run BSBM-BI on a wide range of RDF systems intended to set precedents for future vendor-driven publishing of results with provisions for result verification. Some aspects of benchmarking are currently not covered well in RDF benchmarks, including testing systems in their ability to query across multiple RDF datasources/datasets, using more complex graph traversals, and the ability to have benchmark data and queries exploit data that is ragged and correlated, as commonly found in practice. Synthetic benchmarks often lack such correlations. For this purpose, a "Social Intelligence Benchmark" is defined with the aim to do BI-style analytical queries on large, well-connected social networking graph data. It will investigate analytics on highly connected and correlated graphs.

Software is enhanced for future research and development work. An important part is the preparation for the coming benchmarks by completing the development of the Virtuoso column-wise storage format. Effort has also been put in working on an initial MonetDB/SparQL frontend and on run-time graph query optimization. The former will be supported by the entire LOD2 consortium and the latter by seeking synergy with other EU projects (PlanetData and Teleios).

4. Knowledge Base Creation, Enrichment and Repair

Since the project start in September progress has been made on several software projects that will be of particular interest for the transformation of unstructured data into RDF and Linked Data.

For instance, a *metadata extension to D2R server* (a tool exposing relational databases as RDF) has been developed as a basis for the publication of structured information as Linked Data. It will be included in the next release of D2R. It provides a mechanism for adding metadata to the RDF graphs served by the Linked Data interface of D2R Server (e.g. licensing and provenance).

Likewise, text annotation standards have been investigated for the development of the core components of *NLP2RDF*, which comprises tools and techniques for text understanding. The aim is to convert their output into RDF and thus expose data hidden within the web of documents. In this respect an annotation format is developed to integrate different annotations in a single formalism. While existing proposals such as EARMARK offer acceptable "stand-off" solutions for linguistic annotations that can be reused, an individual annotation scheme is also created for inline annotations. Furthermore, a Web Service is provided for the annotation of texts to allow easy integration.

Regarding using the Wikipedia knowledge extraction *DBpedia Live* as a test bed for reasoning over large LOD knowledge bases development has begun by acquiring new hardware and advancing the underlying DBpedia software framework. Thus, a running prototype exists and is currently tested and refined.

Committed to the open source idea, LOD2 members have founded a DBpedia Internationalization Committee (http://wiki.dbpedia.org/Internationalization) to involve the DBpedia Community into the development progress. The role of LOD2 members will be to orchestrate the community efforts of external developers. The main benefits are two-fold: 1. the framework will be tested extensively, especially for UTF-8 support. 2. Dissemination will be fostered automatically (as it increases the personal identification with the project).

Finally, large knowledge bases are often prone to modeling errors and problems. To improve data quality in this respect, the ORE (Ontology Enrichment and Repair) tool is envisaged as a technical approach that helps to spot inconsistencies and delineate repair plans. So far a website and project for the ORE project have been created with an initial codebase integrating several ontology debugging algorithms.

User Involvement, Promotion and Awareness

User involvement plays a central role in the execution of our use cases. During the first months of the project user involvement has generally been secured by way of online surveys. One survey has been conducted by the Open Knowledge Foundation to gather input from people interested in open government data to inform work on PublicData.eu as part of the requirements elicitation for the e-government use case. Another questionnaire has been set up by the Semantic Web Company to analyze state-of-the-art uses of Linked Data technologies. This survey aimed to collect information from tool providers on technological specifications relevant to LOD2. Likewise, a survey was carried out by LOD2 partner OpenLink among companies and organizations producing RDF engines to better understand the state of the art in RDF storage/query engines and simultaneously to identify gaps in existing benchmarks for RDF engines. A workshop was finally held by one of the use case partners in which staff members directly concerned by the integration of LOD2 technologies were invited to discuss functionalities and services that should be met by these technologies. This discussion has intensively helped to provide a detailed and realistic use case description as well as to define a concrete deployment environment.

Since the start of the project great effort has been undertaken to promote the work that is taking place as part of LOD2. The project is already well-known throughout various communities such as the Semantic Web community, the Open Data movement or governmental bodies and LOD2 relevant industry sectors. This prominence has not only to do with the fact that the LOD2 project is realized by leading Linked Open Data technology researchers, companies, and service providers from across seven European countries. But from the very beginning of the project the LOD2 consortium has also permanently engaged in dissemination activities as described in the following.

One reason for the project's recognition throughout various communities is that a public lod2 mailing list (lod2@lists.okfn.org) has been created for anyone to join and discuss all aspects of the project. Hosted on lists.okfn.org by LOD2 member Open Knowledge Foundation, the list is frequently used (with more than 30 subscribers) and has already included helpful and enlightening discussions on LOD2 topics. Several other communication channels have also been opened and are used continuously (e.g., twitter: @lod2project, the project website and weblog: http://lod2.eu, slideshare, etc.). The LOD2 weblog has served greatly as a platform for exchange. Apart from regularly announcing important LOD2 events, key activities of LOD2 members and particular achievements are documented for the public. Another way to make the LOD2 project widely known has been pursued by our use case partners. The publishing company Wolters Kluwer Deutschland (WKD), for instance, works on bringing the ideas of LOD2 into the media and publishing community. Apart from organizing a workshop to communicate the ideas and technologies of LOD2 and train media specialists outside the project, a first press release focusing on WKD's role in the project was also covered well by the national press in Germany.

At the start of the project a first press release for LOD2 was created for all partners. It was translated into several languages and distributed by the LOD2 partners in their countries. The launch of the project was not only covered well by the national press, but also by specialized journals in countries like Austria, Belgium and Germany. All of these clippings are available at http://lod2.eu for download.

Moreover, the project partners have presented a number of papers and posters at renowned conferences (e.g., I-Semantics 2010, EKAW2010, ISWC2010 and ESTC2010) and also already published papers in prestigious journals of the scientific field. This first output also includes the award-winning paper *Knowledge Engineering for Historians on the Example of the Catalogus Professorum Lipsiensis*, which was presented by Thomas Riechert at the ISWC 2010 in Shanghai and honoured with the Best In-Use Paper Award. A complete and continuously updated list of publications is available at http://lod2.eu.

Finally, the LOD2 consortium has already participated in several events relevant for the LOD2 project. Members have been active giving presentations at conferences and workshops. Moreover, they have promoted the project's progress as well as objectives and pursued networking activities both within and beyond the scientific community. Events attended by LOD2 members include, among others, the I-Semantics Conference (Sept 2010, Graz, Austria), ICT 2010 (Brussels), TOPIX Open Government Data event (Italy), Open Data tent at Mozilla Drumbeat (Spain), Open Government Data Hackathon (Berlin, Vienna, London - et al.), EC Open Data Workshop (Nov 2010, Luxembourg), ISWC 2010 (Nov 2010, Shanghai, China), Eurovoc Conference (Nov 2010, Luxembourg), OGD Camp 2010 (Nov 2010, London, UK), Online Information 2010 (Nov-Dec 2010, London, UK), ESTC2010 (Dec 2010, Vienna, Austria). LOD2 members also participated in several meet-ups all over Europe (such as Open Data Camp Berlin, eGovCamp Vienna, eGovCamp Turin, etc.). Most importantly, to raise awareness for LOD2 ideas and technologies within the scientific community as well as among the broader public, they also organized and carried out workshops (e.g., Linked Data Workshop at the I-Semantics 2010, Linked (Enterprise) Data Workshop at the ESTC 2010, Web Science Workshop at the annual meeting of the German Society for Computer Science 2010, and the OGD Camp in London).

Additionally, several promotion materials such as a leaflet and an LOD2 sticker have been created and produced for the support of the above training and dissemination activities. Several templates for LOD2 printables etc. have been designed and made available for all partners (slide templates, templates for letters etc.). In this way, the LOD2 project relies on a corporate identity that helps its recognition greatly throughout the community.

In order to increase awareness about Linked Data and the scale of published datasets, the LOD2 members have set up PubLink as a Linked Open Data Starter Service in collaboration with the consortia of the EU-FP7 LATC project (Support Action). Publink is a successful dissemination activity to raise attention for the LOD2 project, especially in the area of governmental organizations on a European level. The deadline for applications was 20 December 2010; six applications have been selected from the following organizations:

- Umweltbundesamt GmbH, Austria
- The Greater London Authority, United Kingdom
- Neue Deutsche Biographie, Historische Kommission, Munich, Germany
- The Parliament of Finland
- The City of Vienna, Austria
- Instituto Canario de Estadística (ISTAC), Spain

The on-site consultancy services will start in early 2011. The focus lies on supporting organizations in publishing linked open data (see also http://lod2.eu/Article/Publink.html). It is planned to start a second round of training/consultancy in year two of the project, focusing on consuming linked open data.

Future Work or Exploitation Prospects

In the coming year the focus of the project lies on integrating the different components and approaches into the comprehensive LOD2 Stack. The LOD2 Stack will be a toolset, which is aimed at supporting publishers and users of Linked Data in all aspects of the RDF data management lifecycle ranging from extraction of RDF, storage and querying, authoring, enrichment, linking, quality assessment to browsing and exploration. The LOD2 Stack will be made available as an aligned set of Debian packages, with a common Web user interface and using the Virtuoso triple store as a central RDF data backend. In the end of the second project year, we aim to support applications in the scale of up to 50 Billion facts represented as RDF triples. Using the Debian packaging system as an integration and deployment technology will enable the use of the LOD2 Stack on individual servers, in cloud environments or even in virtualized platforms.

In parallel to the advancement and integration of tools into the LOD2 Stack, we will continue to explore and develop the LOD2 use cases. In particular, we will steadily add new features and showcases to the network of European Open Government Data registries available at http://publicdata.eu.

Further Information

The Consortium

Universität Leipzig (Coordinator)	Germany
Centrum Wiskunde & Informatica	Netherlands
National University of Ireland in Galway	Ireland
Freie Universität Berlin	Germany
OpenLink Software	United Kingdom
Semantic Web Company	Austria
TenForce	Belgium
Exalead	France
Wolters Kluwer Deutschland	Germany
Open Knowledge Foundation	United Kingdom

Contact

For further information visit the LOD2 website at http://lod2.eu or send an email to the project coordinator.

- Dr. Sören Auer (Scientific Project Leader) at auer@uni-leipzig.de
- Nadine Jänicke (Project Manager) at jaenicke@uni-leipzig.de