

## 1. Publishable Summary

---

**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). This 4-year project now going in its last year comprises 15 leading Linked Open Data technology researchers, companies, and service providers from across 11 European countries and South Korea. LOD2 is coordinated by the AKSW research group at the University of Leipzig.

**General Goals.** The semantic web activity has gained momentum with the widespread publishing of structured data as RDF. The Linked Data paradigm has therefore evolved from a practical research idea into a very promising candidate for addressing one of the biggest challenges in the area of intelligent information management: the exploitation of the Web as a platform for data and information integration in addition to document search. In that context LOD2 targets a number of research challenges: improve coherence and quality of data published on the Web, close the performance gap between relational and RDF data management, establish trust on the Linked Data Web and generally lower the entrance barrier for data publishers and users. The LOD2 project tackles these challenges by developing:

- enterprise-ready tools and methodologies for exposing and managing very large amounts of structured information on the Data Web,
- a testbed and bootstrap network of high-quality multi-domain, multi-lingual ontologies from sources such as Wikipedia and OpenStreetMap.
- algorithms based on machine learning for automatically interlinking and fusing data from the Web.
- adaptive tools for searching, browsing, and authoring of Linked Data.

The LOD2 project integrates and syndicates linked data with large-scale, existing applications and showcases the benefits in the three application scenarios publishing, corporate data intranets and Open Government Data.

**Result of the third project year.** The main result of the third project year is the substantially improved LOD2 Stack – an integrated suite of software tools supporting the life-cycle of Linked Data on the Web. The LOD2 Stack was created as a compilation of Debian software packages which can be used with a unified Web user interface. This allows deploying the stack on various infrastructures ranging from virtual machines over full server instances to whole cloud infrastructures. The LOD2 Stack comprises tools covering the life-cycle of linked data on the Web.

- The LOD2 Stack was substantially improved in terms of tool coverage and inter-tool integration. It now comprises 10 additional components integrated in the third year (now 47 components overall). These components are connected via fine-grained dependencies; use the W3C provenance vocabulary to exchange information and the WebID protocol for authentication.
- Through the further integration of previously experimental column store technology and other ground-breaking database innovations for RDF stores (i.e. vectorized execution) the LOD2 consortium got again closer to the goal of closing the performance gap between RDF and relational data management. The performance improvements attained in the third year were demonstrated through a BSBM benchmark run involving 150 Billion RDF triples, which represents a 750-fold increase compared to the largest previously reported results.

- The extraction of RDF from unstructured, semi-structured and structured sources was further improved in the 3<sup>rd</sup> year. With the now standardized NLP Interchange Format (NIF), for example, various Natural Language Processing tools can be loosely coupled and easier integrated. Also, LOD2 Stack extraction tools were internationalized and now support non-Latin based languages such as Cyrillic or Asian languages.
- Data Linking and Integration on the Web was improved with the release of the 2nd Version of the Silk Workbench including the ActiveGenLink workflow which combines active learning and genetic programming release as well as the release of the R2R Mapping framework. The tapping of the 'wisdom of the crowd' is now facilitated through the RDF and crowd-sourcing extensions for OpenRefine (formerly Google Refine).
- The visualization, exploration and browsing widgets of the LOD2 Stack were improved in order to facilitate interactions with end users. With CubeViz the LOD2 Stack now comprises a comprehensive visualization environment for statistical data. With Facete, a scalable visualization tool for spatial was developed integrated into the LOD2 Stack.
- A specifically targeted distribution of the LOD2 Stack for statistical offices and applications was developed in the 3<sup>rd</sup> project year, since the consortium observed a strong interest in Linked Data tooling from stakeholders from the statistics domain.
- The consortium implemented a number of substantial improvements of the pan-European data catalogue Publicdata.eu, which aggregates metadata from more than 30 national, local and regional data catalogues throughout Europe. Now users enjoy a number of novel personalization and data exploration features. Also, a triplification and mapping crowd-sourcing approach was developed and deployed at PublicData.eu now covering more than 10.000 datasets and resulting in 7.3 Billion RDF triples being made available.

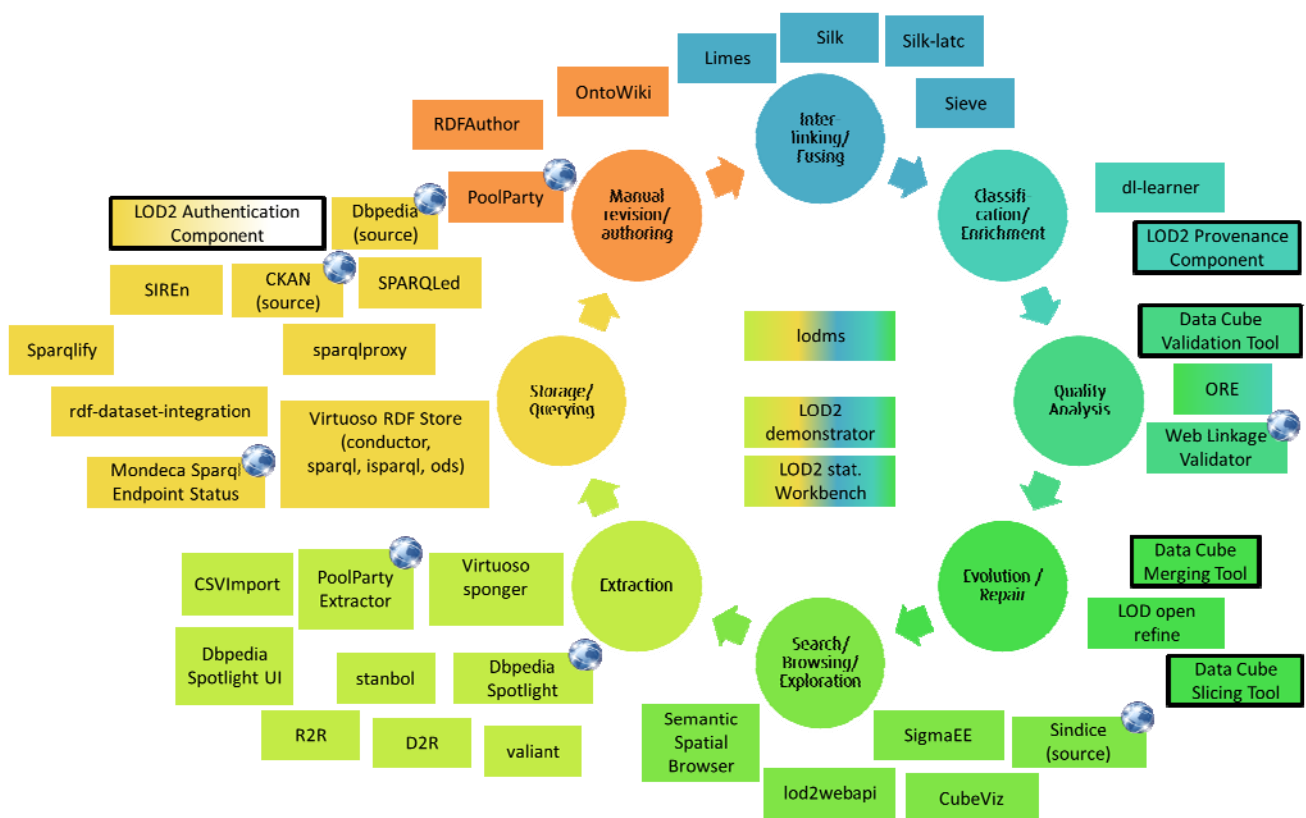


Figure 1: An extended set of 47 LOD2 Stack components support the management of linked data in different life-cycle stages.

In addition to the tools integrated into the LOD2 Stack tool suite also comprises a number of online services and complementary tools and datasets. These include:

- the **CKAN metadata repository**, which interacts bi-directionally with the LOD2 Stack,
- the **DBpedia and LinkedGeoData RDF extraction frameworks**, which have been complemented by LOD2 with Live-SPARQL endpoints interactively reflecting the current state of the two central LOD knowledge bases,
- the **DBpedia Quality evaluation** campaign, which aims at improving DBpedia data quality and also resulted in the TripleCheckMate tool for crowd-sourcing the assessment of Linked Data quality,
- the **DBpedia benchmark**, which measures the performance of triple stores with real data and real queries,
- the **MLODE datasets**, which cover a large number of languages and build the nucleus of a Web of open, interlinked language resources.

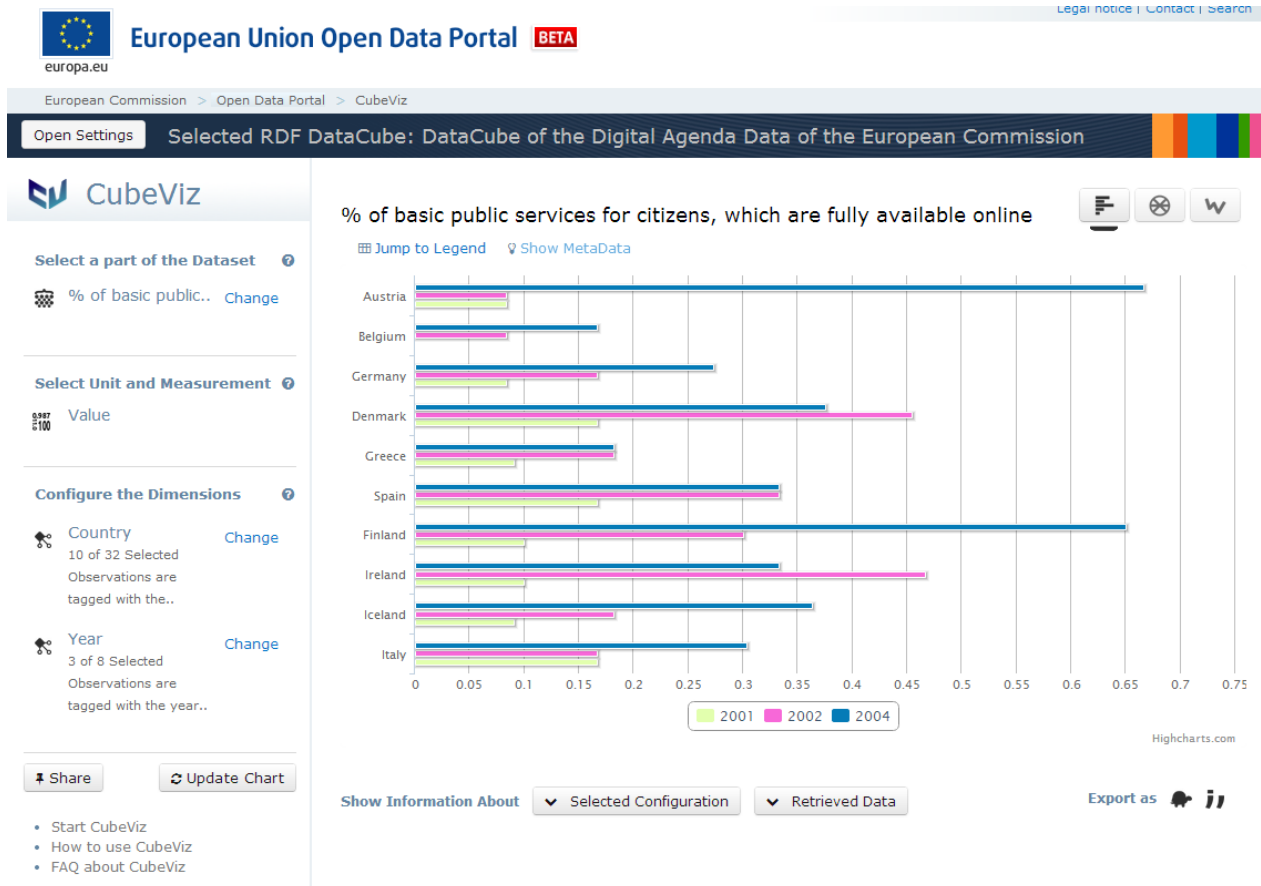


Figure 2: The CubeViz statistical data visualization toolkit, which drives the LOD2 Stack adoption for statistical offices and the EC's open data portal.

The project intensified the exploration of the three LOD2 use-cases by further developing and refining application prototypes of the LOD2 Stack for the three application domains publishing, enterprise and governmental data. The enterprise adoption of the LOD2 Stack was meanwhile successfully deployed in the production environment of a major European corporation.

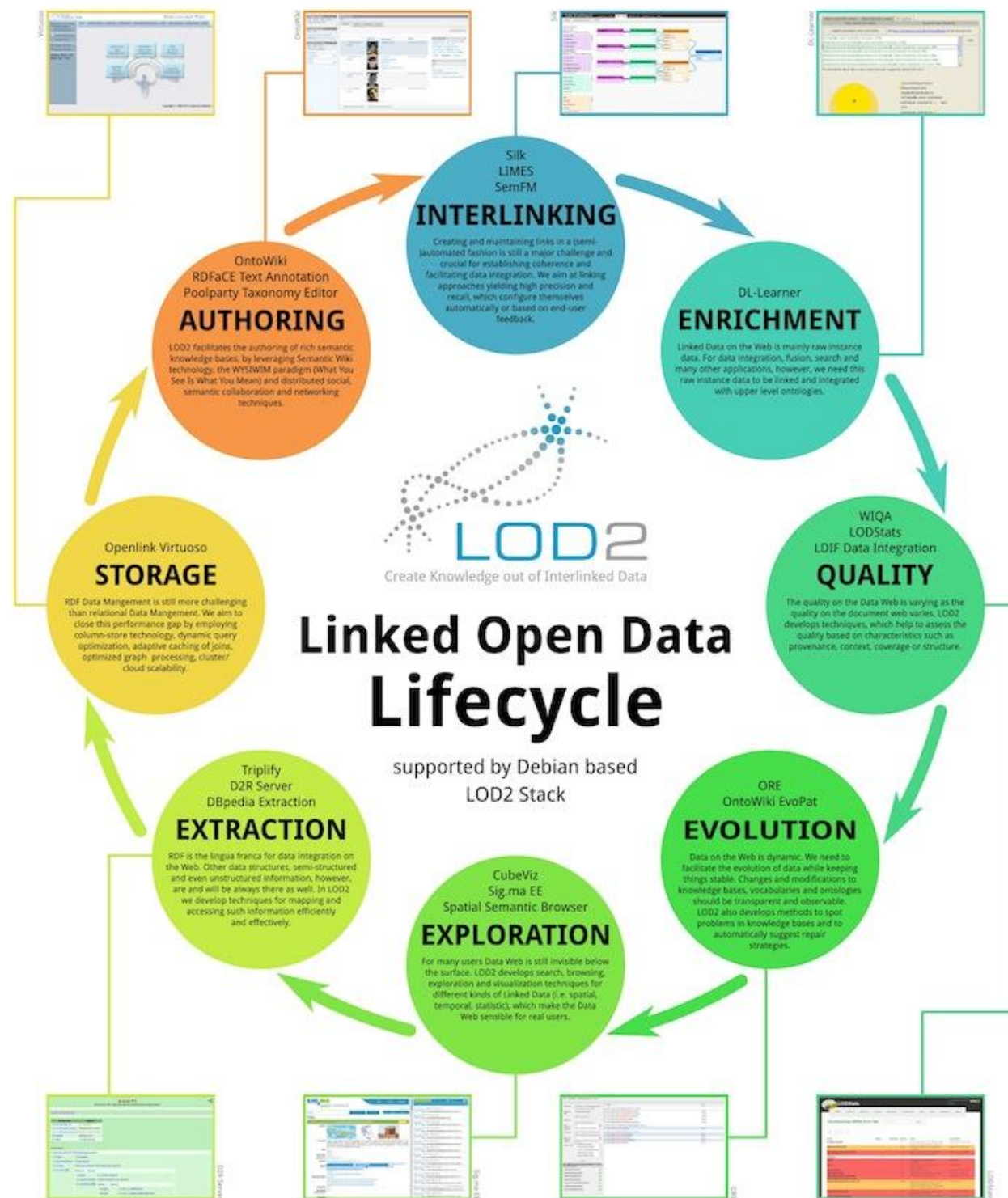
During the third project year, the LOD2 members published again more than 40 peer-reviewed scientific publications, won two Best-Paper awards, organized a number of events (e.g. LDOW workshop at WWW conference, Open Knowledge Conference, I-Semantics, European Data

Forum) and successfully supported a number of organizations in the publishing of Linked Data (e.g. Flemisch Government, National Library of Israel, Stichting Bibliotheek.nl). The consortium contributed to many W3C standards, co-organized two summer schools and hold seven webinars explaining the use and benefits of the LOD2 Stack tools. A comprehensive lecture series OpenCourseWare for the Semantic Data Web was made available by the LOD2 consortium in more than seven languages at: <http://slidewiki.org>

**Outlook.** In the final project year, the project will intensify its efforts to demonstrate the benefit of Linked Data in the three use cases. The refinement and integration of LOD2 Stack components will continue and new research approaches for exploiting synergies between the different phases of the Linked Data life-cycle. LOD2 will co-organize the third edition of the European Data Forum, which will take place on March 19-20 2014 in Athens, Greece.

**The LOD2 consortium** comprises with its academic partners Universität Leipzig, Universität Mannheim, University of Economics Prague, Institute Mihailo Pupin, Korean Advanced Institute of Science and Technology and National University of Ireland in Galway research expertise in Semantic Web technologies, ontological engineering, machine learning, Web search, information retrieval, databases and knowledge stores. With LOD2 partner CWI's reputation in the database realm, LOD2 aims to contribute to cross-fertilization between database and semantic web research. The LOD2 consortium has engineering power provided by the companies TenForce and OpenLink Software as well as dissemination and exploitation expertise provided by its innovative SMEs (Exalead, Openlink, TenForce, I2G and Semantic Web Company), a large corporation (Wolters Kluwer) and the eGovernment and Open Knowledge communities represented by Open Knowledge Foundation.

Further information can be found on the project homepage at: <http://lod2.eu> or from the coordinator Prof. Sören Auer at Universität Leipzig (Phone: +49 (341) 97-32367, Fax: +49 (341) 97-32329, Email: [auer@uni-leipzig.de](mailto:auer@uni-leipzig.de)).



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 the University of Leipzig.



Figure 3: Poster illustrating the Linked Data life-cycle stages and the tools developed to support these.