European Public Sector Information Platform

Topic Report No. 2012 / 11

Linked Data

Author: Daniel Dietrich

Published: November 2012

Keywords

Linked Data, Linked Open Data, Open Data, Semantic Web, Public Sector, PSI, PSI Directive, re-use, Technology

1 Executive Summary

Over the last years, activities around Linked Data and the Semantic Web have gained momentum. More and more Public Sector Bodies (PSB) and other Organizations are publishing structured data using linked data standards and technologies, while more and more Companies use advanced Linked Data Technologies to combine, query and analyse huge amounts of Linked Data as part of there services and products. The Linked Data paradigm is evolving from an academic concept 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.

This topic report is about Linked Data and its relevance for the reuse of Public Sector Information (PSI). It introduces the concept and technology and looks into their uptake by governments and businesses in European Union member states. The report showcases recent research projects and examples of actual implementation both in the private and the public sector.

2 Introduction

Data must be accessible and discoverable in order to be reused. There are several measures to increase the accessibility and discoverability of data. These range from very basic measures, like using open standard file formats that can be processed by machines (so search engines can read them properly), to adding metadata to data and optimising data and websites for search (-engines). More advanced measures include publishing and aggregating datasets and their metadata in specialized data catalogues, to publishing data using Linked Data Principles and Technologies.

In 2006, Tim Berners-Lee the inventor of the World Wide Web, introduced the concept of linked data as the new paradigm for the future of the Internet. With Linked Data best practices the Web is evolving from a global information space of linked documents to one where both documents and data are linked. There is consensus that data can only be processed and analysed on large scale following linked data best practice.

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. It builds upon standard Web technologies such as Hypertext

Transfer Protocol (HTTP)[1] and Uniform Resource Identifiers (URI)[2], but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be processed automatically by computers. This enables data from different sources to be connected and queried allowing for better interpretation and analysis.

URIs and HTTP are supplemented by a technology that is critical to the Web of Data – the Resource Description Framework (RDF)[3]. Whilst the Hyper Text Markup Language (HTML)[4], provides a means to structure and link documents on the Web, RDF provides a generic, graph-based data model with which to structure and link data that describes things in the world.

History of the Technology

The World Wide Web Consortium (W3C)[5] published a specification of Resource Description Framework (RDF) data model as a recommendation in 1999 and further developed it since. A definition of RDF can be summed as:

> *"RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications."*[6]

The Semantic Web is a collaborative movement led by W3C promoting common data formats on the Internet. The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web dominated by unstructured and semi-structured documents into a "web of data".

The Semantic Web involves publishing in languages specifically designed for data: Resource Description Framework (RDF), Web Ontology Language (OWL)[7], and Extensible Markup

1 Hypertext Transfer Protocol (HTTP) is an application protocol for distributed, collaborative, hypermedia information systems.[1] HTTP is the foundation of data communication for the World Wide Web. http://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol

2 A Uniform resource identifier (URI), is a string of characters used to identify a name or a resource. Such identification enables interaction with representations of the resource over a network (typically the World Wide Web) http://en.wikipedia.org/wiki/Uniform_resource_identifier

3 W3C, Resource Description Framework (RDF), http://www.w3.org/RDF/

4 W3C, Hyper Text Markup Language (HTML), http://www.w3.org/html/

5 World Wide Web Consortium (W3C), ttp://www.w3.org/

6 W3C, Resource Description Framework (RDF), http://www.w3.org/RDF/

7 W3C, Web Ontology Language (OWL), http://www.w3.org/TR/owl2-overview/

Language (XML)[8]. HTML describes documents and the links between them. RDF, OWL, and XML, by contrast, can describe arbitrary things such as people, meetings, or airplane parts.

The RDF model encodes data in the form of subject, predicate, object triples. Such as "Peter lives in Sofia" or "Sofia is Capital of Bulgaria". The subject and object of a triple are both URIs that each identify a resource, or a URI and a string literal respectively. The predicate specifies how the subject and object are related, and is also represented by a URI.

Linked Data principles

In 2006 Tim Berners-Lee outlined a set of very simple principles for publishing data on the Web in a way that all published data becomes part of a single global data space:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things

These have become known as the 'Linked Data Principles'[9], and provide a basic recipe for publishing and connecting data using the infrastructure of the Web while adhering to its architecture and standards.

The "Open" in Linked Open Data

What is the difference between linked data vs. linked open data? In fact any data can be linked following the Linked Data Principles. However for Data to be reused it must be openly licensed with a license allowing for its commercial and non-commercial reuse. This is why in 2010 Tim Berners-Lee introduced the 5-star scheme for Linked Open Data. The first star (and consequently all other stars) can only be achieved by publishing the data on the web under an Open License.[10]

5-Star Linked Open Data

No star means the data is not available under an open licence, even if it is available on-line.

★ *One star* means a good start: the data is accessible on the Web. It is readable by the human eye, but not by a software agent, because it is in a 'closed' document format, and therefore cannot be easily re-used.

8 W3C, Extensible Markup Language (XML), http://www.w3.org/XML/

9 W3C, Linked Data principles, 2006, http://www.w3.org/DesignIssues/LinkedData

10 According to the Open Definition: "A piece of content or data is open if anyone is free to use, reuse, and redistribute it - subject only, at most, to the requirement to attribute and/or share-alike." A list of licenses that are in compliance with the Open Definition can be found at: http://opendefinition.org/licenses/

★★ *Two stars* mean that the data is accessible on the Web in a structured, machine-readable format. Thus, the re-user can process, export and publish the data easily, still depending however on proprietary software like Word or Excel.

★★★ *Three stars* mean that re-users will no longer need to rely on proprietary software (like CSV instead of Excel). Accordingly, re-users can manipulate the data in any way, without being confined to a particular software producer.

★★★★ *Four stars* mean that the data is now *in* the Web as opposed to *on* the Web through the use of a URI, a Uniform Resource Identifier[11]. As a URI is completely unique, it gives a fine-granular control over the data, allowing for things like bookmarking and linking.

★★★★★ *Five stars* mean that the data is not only *in* the Web but is also linked to other data, fully exploiting its network effects. Through this interlinking, data gets interconnected whereby the value increases exponentially, since it becomes discoverable from other sources and is given a context (e.g., through links to Wikipedia).

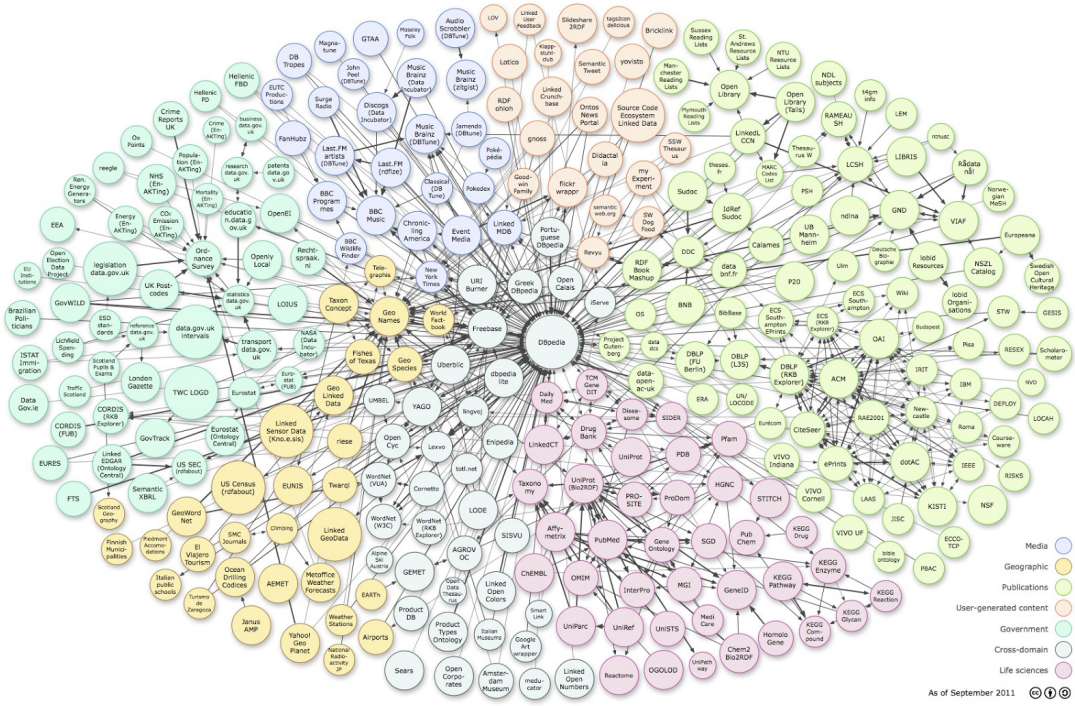
★	make your stuff available on the Web (whatever format) under an open license. See details	example ...
★★	make it available as structured data (e.g., Excel instead of image scan of a table). See details	example ...
★★★	use non-proprietary formats (e.g., CSV instead of Excel). See details	example ...
★★★★	use URIs to identify things, so that people can point at your stuff. See details	example ...
★★★★★	link your data to other data to provide context. See details	example ...

Source: Linked Data Labs, DERI. http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

The Linking Open Data cloud

11 Uniform Resource Identifier, http://en.wikipedia.org/wiki/Uniform_Resource_Identifier

Linked Open Data best practices have been adopted by an increasing number of data providers over the last three years, leading to the creation of a global data space containing billions of assertions - the Web of Data (see figure below).



Source: The Linking Open Data cloud diagram, http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.html

3 State of Play

Over recent years the Linked Open Data paradigm has gained momentum as an approach of publishing data to and reusing data from the web - be it for commercial or non-commons-oriented purposes. To translate this initial progress into broader uptake by market players building viable applications, products and services non Linked Data, the European Commission has supported several interdisciplinary Research and Development initiatives to generally lower the entrance barrier for data publishers and users.

The ultimate goal of all these publicly funded research under the European Commission within the FP7 Information and Communication Technologies Work Programme[12] is to make Linked Data operational and useful for both the private and the public sector. Among these funded research projects are the LOD2, Linked Open Data Around-The-Clock (LATC) and PlanetData.

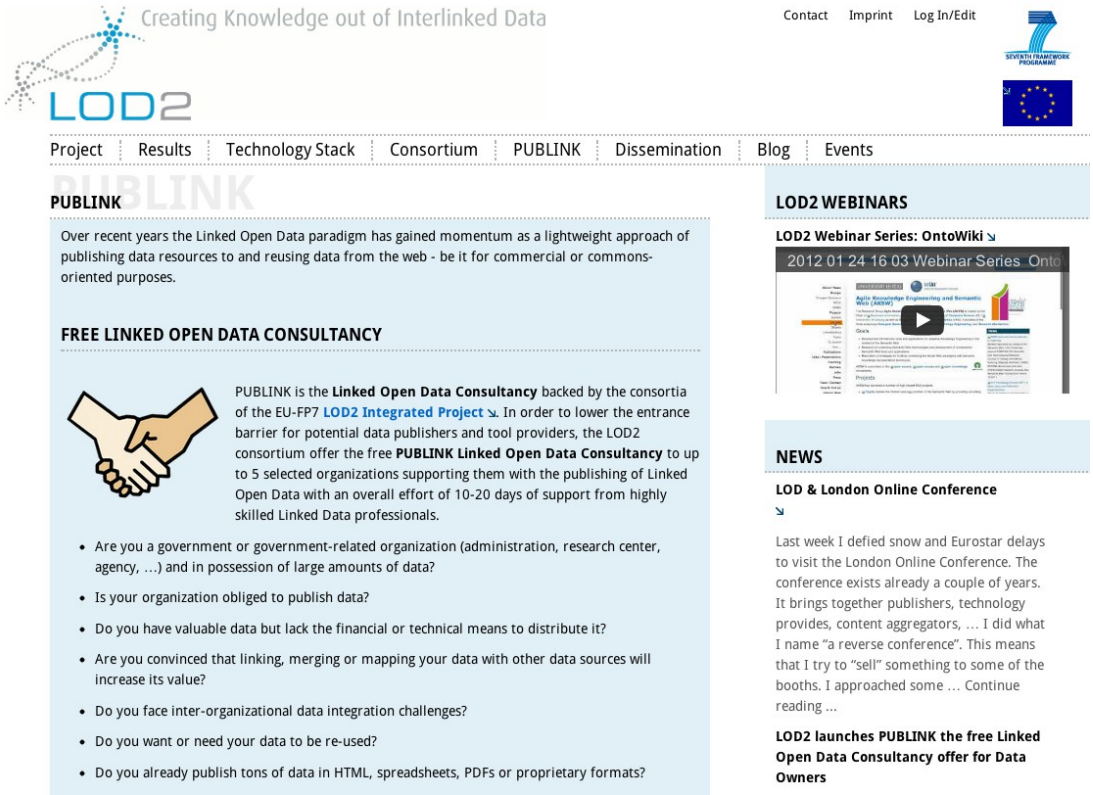
PlanetData aims to establish an interdisciplinary, sustainable European community of researchers, helping organizations to expose their data on the Web in a useful way, by

12 European Commission within the FP7 Information and Communication Technologies Work Programme, http://cordis.europa.eu/fp7/ict/

conducting interdisciplinary research on: a) the Web-friendly representation, robust integration, and scalable transmission, processing, and management of stream-like data, b) the assessment and improvement of the quality of large open data sets of various modalities, and the augmentation of these data sets with self-descriptive metadata giving an account of quality and contextual aspects, and c) the accountable usage of Web-based datasets taking into consideration provenance, privacy, trust, and access rights. http://planet-data.eu/

Source: Planet Data, http://planet-data.eu/

LOD2 is a large-scale integrating project bringing together leading Linked Open Data technology researchers, companies, and service providers from across 11 European countries to address the following 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. In order to lower the entrance barrier for potential data publishers and tool providers, the LOD2 consortium offer the free PUBLINK Linked Open Data Consultancy to up to 5 selected organizations supporting them with the publishing of Linked Open Data with an overall effort of 10-20 days of support from highly skilled Linked Data professionals. http://lod2.eu/



Source: LOD2, http://lod2.eu/

LATC, or 'Linked Open Data Around-The-Clock' is a specific Support Action in the context of the FP7 ICT Programme. Its mission is to support people and organisations to better publish and consume Linked Open Data. The project aims to a) increase the number, the quality and the accuracy of data links between LOD datasets, b) support institutions as well as individuals with Linked Data publication and consumption, and c) create an in-depth test-bed for data intensive applications by publishing datasets produced by the European Institutions[13] as Linked Open Data to seed the EU data cloud.[14] This includes datasets from the EU Commission Financial Transparency System and the Community Research and Development Information Service (CORDIS)[15] as well as datasets from EUR-Lex Access to European Union law[16], Eurostat[17] and European Central Bank (ECB)[18] and many more. http://latc-project.eu/

13 Institutions and Bodies of the European Union on http://institutions.publicdata.eu/

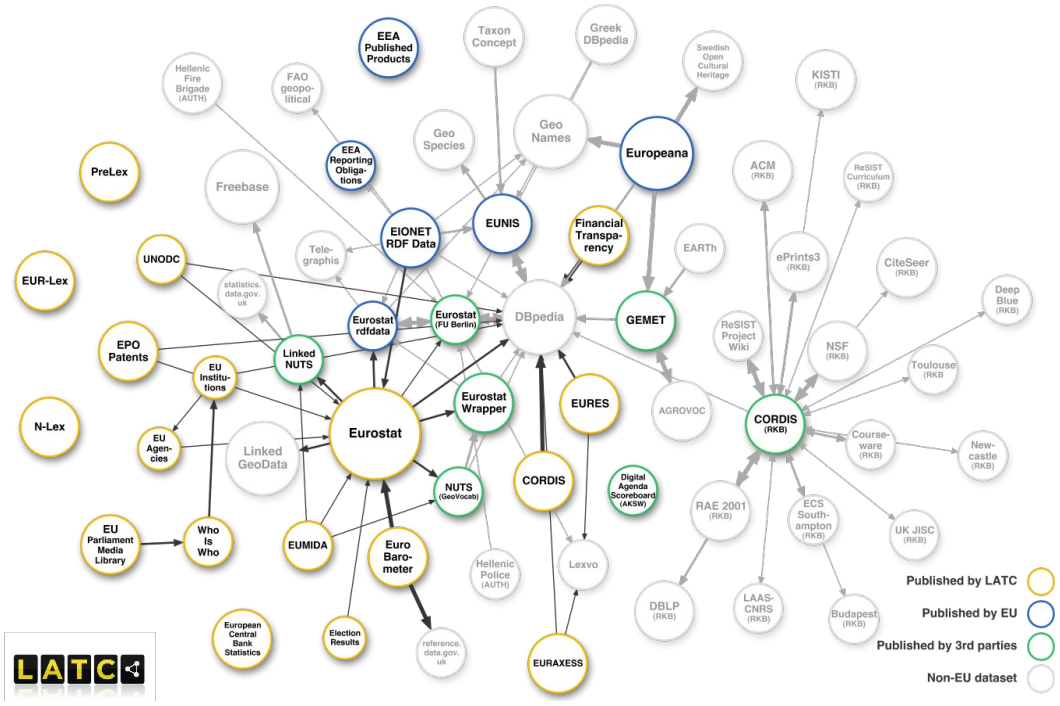
14 LATC, EU Data Cloud, http://latc-project.eu/datasets

15 Community Research and Development Information Service (CORDIS), http://www4.wiwiss.fu-berlin.de/cordis/

16 EUR-Lex Access to European Union law, http://eur-lex.publicdata.eu/

17 Eurostat, http://eurostat.linked-statistics.org/

18 European Central Bank (ECB), http://ecb.publicdata.eu



Source: LATC, European Data Cloud, http://latc-project.eu/datasets

The **European Data Forum** (EDF) is a meeting place for industry, research, policymakers and community initiatives to discuss the challenges of Big Data and the emerging Data Economy and to develop suitable action plans for addressing these challenges. Of special focus for the EDF are Small and Medium-sized Enterprises (SMEs), since they are driving innovation and competition in many data-driven economic sectors. The range of topics discussed at the European Data Forum ranges from novel data-driven business models (e.g. data clearing houses), and technological innovations (e.g. Linked Data Web) to societal aspects (e.g. open governmental data as well as data privacy and security). http://www.data-forum.eu/

Source: European Data Forum, http://www.data-forum.eu/

Talis and the Kasabi case

The UK based Talis Group has played a major role in the research and development of Linked Data over the past years. Part of Talis engagement was Kasabi, a data marketplace service launched in June 2011. The aim of the service was to provide: a) complete data hosting service, including storage and APIs, b) a full Linked Data hosting solution and c) a marketplace for developers to share and gain access to datasets.

The system was built on Talis's proprietary data hosting platform that took care of storage and data indexing. The Kasabi APIs provided an additional layer of functionality including access control and usage monitoring. Kasabi was shut down in July 2012 after Talis decided to refocus its on-going investments on its Education division. While this specific experience cannot be generalized the following two comments by Talis illustrate the slow process of turning the Linked Data paradigm into real-world applications:

> "Usage had remained relatively low throughout the life-time of the service, ultimately contributing to the decision to shutdown the service."[19]

> "In our view, the commercial realities for Linked Data technologies and skills whilst growing is still doing so at a very slow rate, too slow for us to sustain our current levels of investment."[20]

19 Notice of shutting down Kasami, website assessed on 10. January 2013 at: http://blog.kasabi.com/about/

20 Talis Group statement, website assessed on 10. January 2013 at: http://talis-systems.com/

4 Linked Data and PSI

In recent years Governments from around the world began to publish their data as Linked Open Data applying Linked Data Principles to PSI. By 2012 Linked Open Government Data (LOGD) makes up to 10-20 % of the LOD Cloud.

In order to unlock the potential of digital PSI, developers and other prospective users must be able to find datasets they might be interested in reusing. While publishing PSI as raw data on the Internet is the first step it doesn't mean that people can actually find the data easily. Linked Data Principles and Technologies can not only increase the accessibility and discoverability of Data but also introduce options and tool to combine huge amounts of data from multiple sources for advanced analysis and extraction of knowledge.

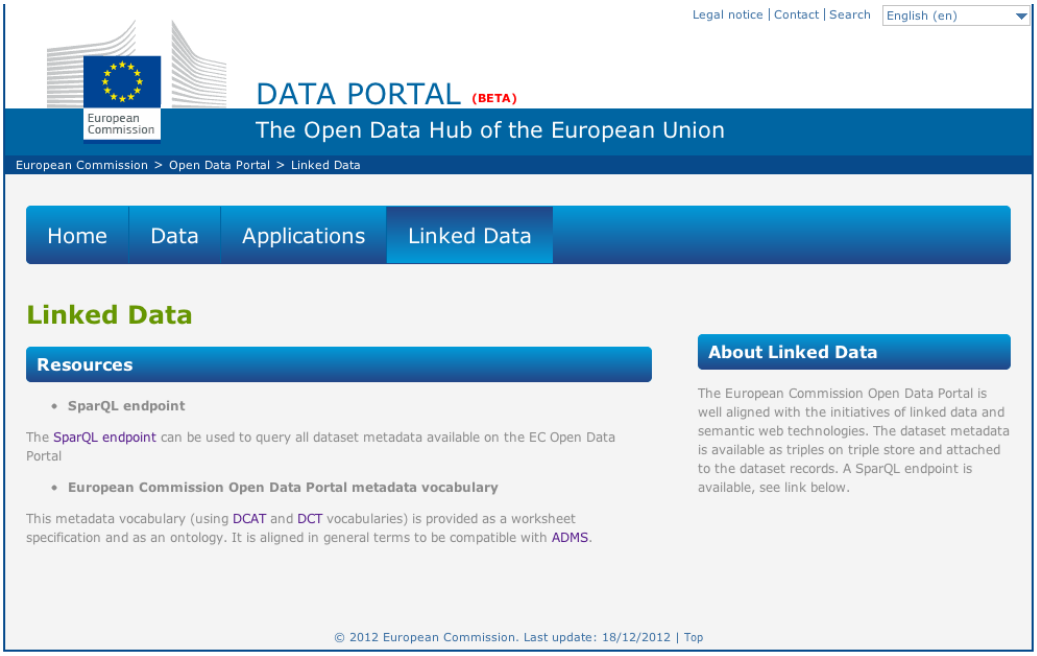
The **European Commission Open Data Portal,** launched late 2012 in beta, is well aligned with the initiatives of linked data and semantic web technologies. The dataset metadata is available as triples on a triple store and attached to the dataset records. A SparQL endpoint[21] can be used to query all dataset metadata available on the EC Open Data Portal. The European Commission Open Data Portal metadata vocabulary (using Data Catalog Vocabulary (DCAT)[22] and DCMI Metadata Terms (DCT)[23] vocabularies) is provided as a worksheet specification and as ontology. It is aligned in general terms to be compatible with Asset Description Metadata Schema (ADMS).[24] http://open-data.europa.eu/open-data/

21 SparQL endpoint of the European Commission Open Data Portal http://open-data.europa.eu/open-data/sparql

22 Data Catalog Vocabulary (DCAT) http://www.w3.org/TR/vocab-dcat/

23 Dublin Core Metadata Initiative, Metadata Terms (DCT) http://dublincore.org/documents/dcmi-terms/

24 The Asset Description Metadata Schema (ADMS) is a metadata vocabulary to describe semantic interoperability assets, http://joinup.ec.europa.eu/asset/adms/home

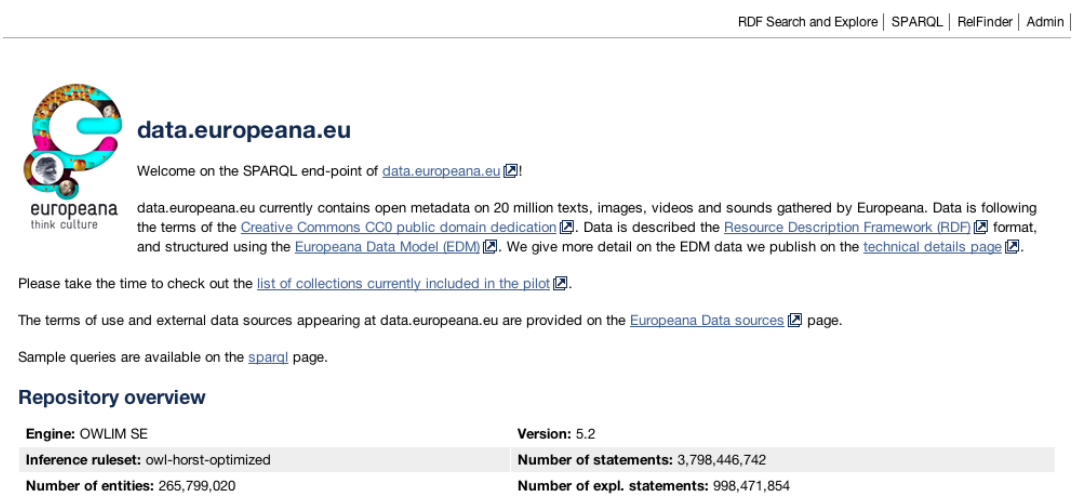


Source: European Commission Open Data Portal, Linked Data Service, http://open-data.europa.eu/open-data/linked-data

Europeana Linked Open Data (LOD)

Europeana is making data available to the public and private sectors alike so they can use it to develop of innovative applications, products and services. The Europeana Linked Open Data (LOD) Pilot[25] contains open metadata on 20 million texts, images, videos and sounds gathered by Europeana. These objects come from data providers who have reacted positively to Europeana's initiative of promoting more open data and have signed the Data Exchange Agreement (DEA). This support for commercial exploration of the metadata aggregated in Europeana is central to Europeana's business strategy. http://data.europeana.eu/

25 Europeana Linked Open Data (LOD) Pilot, http://data.europeana.eu



Source: SPARQL end-point of data.europeana.eu, http://europeana.ontotext.com/

Examples of PSBs publishing LOD

Also at the national and sub-national level various Public Sector Bodies in EU member states have started to publish data following as linked data or linked open data. Positive examples being the German National Library (Deutsche Nationalbibliothek, DNB), which has released a Linked Data version of the German National Bibliography under Creative Commons Zero, making it perfect 5-Star Linked Open Data. According to an article on the openGLAM blog[26], Julia Hauser, from the DNB, explains in an email to the World Wide Web Consortium (W3C): "In 2010 the German National Library (DNB) started publishing authority data as Linked Data. The existing Linked Data service of the DNB[27] is now extended with title data. In this context the licence for linked data is shifted to Creative Commons Zero.[28]"

5 Conclusion

Semantic Web and Linked Data technologies and principles are around for more than a decade and they remain popular research topics. Uptake by the public and private sector and real-world implementations, however clearly fall behind this academic enthusiasm for the technology. Although more and more private companies and Public Sector Bodies start embracing Linked Data Principles and Technologies it appears that real or anticipated barriers of adapting to these technologies remain as too big. Commonly encountered problems may include: the technology appears to be too complicated, the initial investment to big, or the expected benefits too vague to convince both stakeholders in the private sector and the public sector to embrace Linked Data Principles. The question if and when the paradigm of Linked Data Paradigm will unfold its full potential, depends to large

26 German National Library releases more Linked Open Data under a more Open License, http://openglam.org/2012/02/09/191/

27 Linked Data service of the DNB, https://wiki.dnb.de/display/LDS/

28 Creative Commons Zero, http://creativecommons.org/publicdomain/zero/1.0/

extends on the adoption and acceptance of a critical mass of market players. Initiatives such as the EC funded research and innovation projects can clearly help facilitate and accelerate this process.

About the Author

Daniel Dietrich was born in 1973 in Frankfurt, Germany. His academic work covers political science, computer science and communication science in Frankfurt and Berlin. He worked as Research Associate at Technical University Berlin, Department of Internet and Society until the end of 2011. He has been working for the Open Knowledge Foundation (OKFN), since 2009 and is Chairman of the German Chapter of the Open Knowledge Foundation. He is the Project Coordinator for the OKF Project Open Definition as well as the Coordinator of the Working Group on Open Government Data and the Working Group on Open Data in the EU. He is the co-founder of the Open Data Network, a non-profit advocacy organisation to promote Open Data, Open Government and Transparency in Germany, Europe and beyond. In 2011 he became Editor of the ePSI platform.

Copyright information