The Emerging Web of Linked Data

Christian Bizer, Freie Universität Berlin

The classic World Wide Web is built upon the idea to set hyperlinks between Web documents. Hyperlinks are the basis for navigating and crawling the Web. Hyperlinks integrate all Web documents into a single global information space.

In recent years, major Web data sources like Google, Yahoo!, eBay and Amazon have started to provide access to their databases through Web APIs. ProgrammableWeb.com currently lists over 1300 of such APIs. The wealth of data accessible via Web APIs has lead to the development of exciting mashups that combine data from different sources. Unlike the classic Web which is built on a small set of standards – URIs, HTTP and HTML -, different Web APIs rely on different identification mechanisms, different access mechanisms, and retrieved data is represented in different formats. It is generally not possible to set hyperlinks between data items provided by different APIs. Web APIs therefore slice the Web into separate data silos. Mashup developers are forced to choose a specific set of data sources for their application. They cannot implement applications against all data that is available on the Web or will become available during the life-time of the application.

In order to overcome this fragmentation, Tim Berners-Lee outlined a set of best practices for publishing and connecting structured data on the Web: the Linked Data principles [1]. In summary, the Linked Data principles provide guidelines on how to use standardized Web technologies to set data-level links between data from different sources. In analogy to the classic Web, data-level links connect data from different sources into a single global dataspace [2]. As this Web of Linked Data is based on standards for the identification, retrieval and representation of data, it is possible to use generic data browsers to explore the complete dataspace. As data from different sources is connected by links, it is possible to crawl the dataspace, fuse data about entities from different sources and provide expressive query capabilities over aggregated data, similar to how a local database is queried today. Unlike Web 2.0 mashups that work against a fixed set of data sources, Linked Data applications can discover new data sources at run-time by following RDF links and can thus deliver more complete answers as new data sources appear on the Web.

Technologically, the core idea of Linked Data is to use HTTP URIs not only for the identification of Web documents, but for the identification of arbitrary real-world entities [3]. Data about these entities is represented using the Resource Description Framework (RDF). Whenever one of these URIs is dereferenced, the corresponding Web server provides an RDF/XML or RDFa description of the identified entity. These descriptions may contain links to entities described by other data sources. Links take the form of RDF triples where the subject of the triple is a URI in the namespace of one server, while the object of the triple is a URI in the namespace of the other [4]. The predicate URI of the triple determines the type of the link. Whenever a predicate URI is dereferenced, the corresponding server responds with a RDF Vocabulary Definition Language (RDFS) or Web Ontology Language (OWL) definition of the link type [5]. These descriptions may in turn contain links pointing at other vocabularies, thereby defining mappings between related vocabularies.

The Web of Linked Data can be seen as an additional layer that is tightly interwoven with the classic document Web and has many of the same properties:

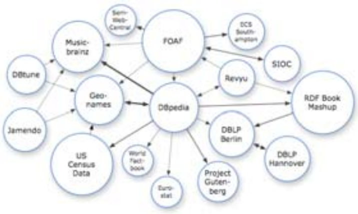
- Anyone can publish data to the Web of Linked Data.
- Entities are connected by links, creating a global data graph that spans data sources and enables the discovery of new data sources.
- Data is self-describing. If an application encounters data represented using an unfamiliar vocabulary, the application can resolve the URIs that identify vocabulary terms in order to find their definition.
- The Web of Data is open, meaning that applications can discover new data sources at run-time by following links.

The Linking Open Data Effort

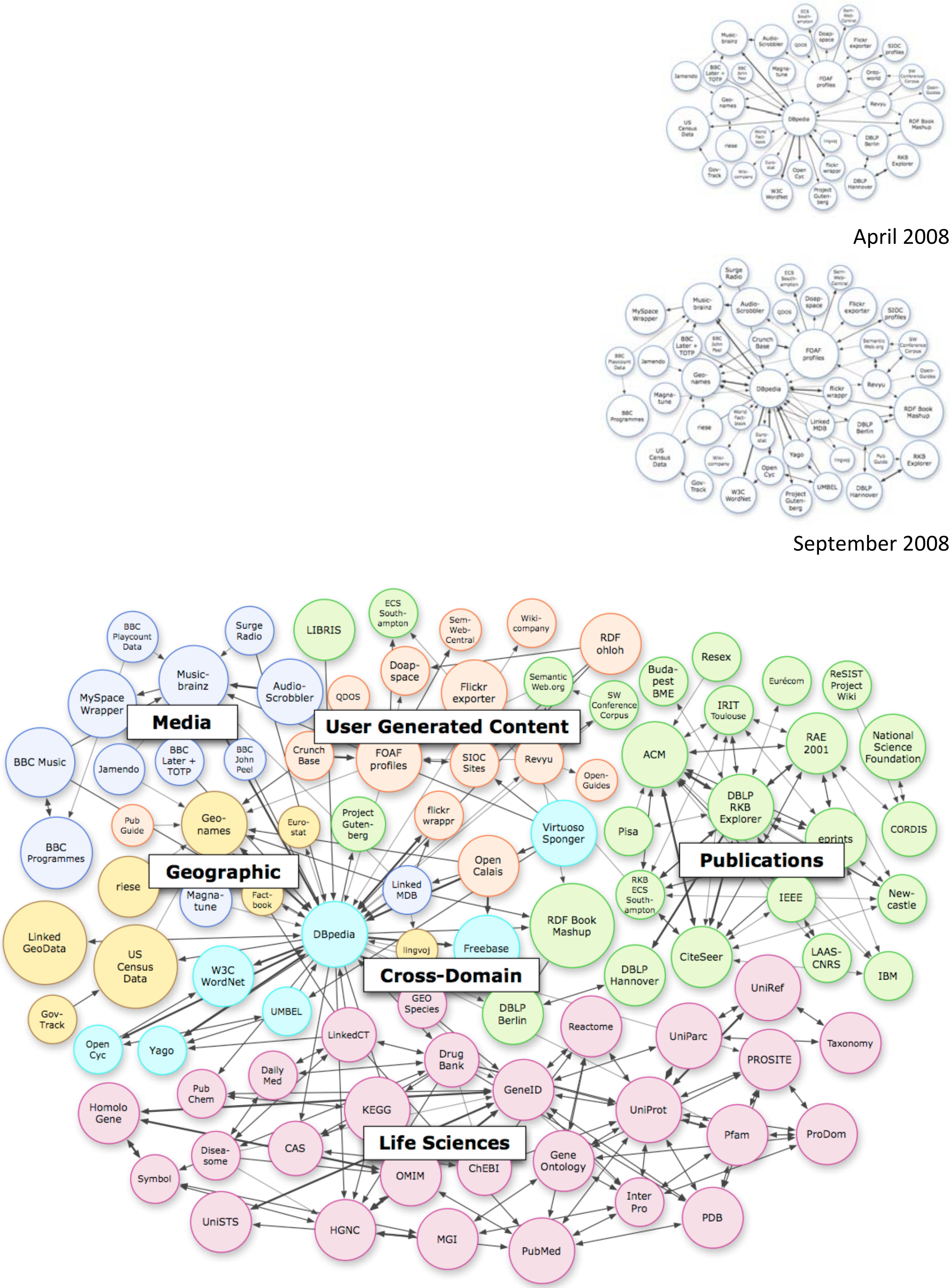
The Linked Data principles are being adopted by an increasing number of data providers over the last three years, leading to the creation of a global dataspace containing billions of assertions about geographic locations, people, companies, books, scientific publications, films, music, television and radio programmes, genes, proteins, drugs and clinical trials, online communities, statistical data, census results, and reviews.

The publication of Linked Data is loosely coordinated by the W3C Linking Open Data project, a grassroots community effort founded in January 2007. The original and ongoing goal of the project is to bootstrap the Web of Linked Data by identifying existing data sets that are available under open licenses, converting them to RDF according to the Linked Data principles, and publishing them on the Web. Participants of the project maintain a wiki which collects community news, statistics about published data sets, and information about Linked Data publishing tools and applications [Endnote: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData]. Discussions about the Web of Linked Data take place on the *public-lod@w3.org* mailing list. The project is open for anyone to participate simply by publishing a data set according to the Linked Data principles and interlinking it with existing data sets.

Figure 1 illustrates the growth of the data cloud originating from the W3C Linking Open Data project. Each node in the diagram represents a distinct data set published as Linked Data. The arcs indicate the existence of links between items in the two data sets. Heavier arcs correspond to a greater number of links, while bidirectional arcs indicate that outward links to the other exist in each data set.



July 2007



April 2008

September 2008

July 2009

Figure 1: Growth of the Linking Open Data cloud.

The colored diagram at the bottom of Figure 1 illustrates the size of Linking Open Data cloud as of July 2009 and classifies the data sets by topical domain. The following sections give an overview of the main data sets as well as ongoing data publication efforts. Further details and references to the mentioned data sets are found in the Linking Open Data wiki.

Media

A major Linked Data publisher within the media industry is the British Broadcasting Corporation (BBC). The BBC /programmes and /music sites provide data about episodes of radio and TV programmes. The data is interlinked with Musicbrainz, an open-license music database, and with DBpedia, a Linked Data version of Wikipedia. The links between BBC /music, Musicbrainz and DBpedia allow applications to retrieve and combine data about artists from all three sources. Further media companies that have announced that they are going to publish Linked Data include the New York Times, CNET and Thomson Reuters. Thomson Reuters has also developed OpenCalais, a service for the annotation of news texts with URIs from the Linked Data cloud referring to places, companies and people.

Publications

The American Library of Congress and the German National Library of Economics publish their subject heading taxonomies as Linked Data. Linked Data about scholarly publications is provided by the L3S Research Center which hosts a Linked Data version of the DBLP bibliography. The ReSIST project publishes and interlinks bibliographic databases such as the IEEE digital library, CiteSeer, and various institutional repositories. Linked Data about books is provided by the RDF Book Mashup, a wrapper around the Amazon and the Google Base APIs. The Open Archives Initiative has based its new Object Exchange and Reuse (OAI-ORE) standard on the Linked Data principles and it is likely that the deployment of this standard will further accelerate the availability of Linked Data related to publications.

Life Sciences

A major provider of Linked Data related to life sciences is the Bio2RDF project which has interlinked more than 30 widely used life sciences data sets including UniProt, KEGG, CAS, PubMed and the Gene Ontology. Altogether, the Bio2RDF data sets comprise more that 2 billion RDF triples. Within the W3C Linking Open Drug Data effort, the pharmaceutical companies Eli Lilly, AstraZeneca, and Johnson & Johnson cooperate to interlink open-license data about drugs and clinical trials in order to ease drug discovery.

Geographic Data

Geonames, an open-license geographical database, publishes Linked Data about 8 million locations. The LinkedGeoData project publishes a Linked Data version of OpenStreetMap providing information about more than 350 million spatial features. Locations in Geonames and LinkedGeoData are interlinked with corresponding locations in DBpedia, if existent. The British Ordnance Survey office has started to publish topological information prescribing the administrative areas within the UK as Linked Data. There are also conversions of the EuroStat, Wolrd Factbook and US Census data sets available as Linked Data.

User Generated Content

An increasing amount of metadata about user generated content from Web 2.0 sites is becoming available as Linked Data. Examples include the flickr wrappr around the Flickr photo sharing service and the SIOC exporters for WordPress, the Drupal content management system and the phpBB bulletin boards. Zemanta provides tools for the semi-automated enrichment of blog posts with data-level links pointing to DBpedia, Freebase, MusicBrainz and Semantic CrunchBase. A futher service for the annotation of Web content with Linked Data URIs is Faviki. These annotations connect the classic document Web with the Web of Linked Data. The links can be used by applications to retrieve background information about the topics of a blog post or a location depicted by a photo, which can in turn be used by the application to provide a richer user experience.

Cross-Domain Data Sources

Data sources that provide information spanning multiple domains are curial for connecting data into a single global data space and to avoid the fragmentation of the dataspace into distinct topical islands. An example of such a data source is DBpedia, which publishes data that has been extracted from the "infoboxes" commonly seen on the right hand side of Wikipedia articles. As DBpedia covers a wide range of topics and has a high degree of conceptual overlap with various other data sets, an various data publishers have started to set links from their data sources to DBpedia, making DBpedia one of the central interlinking hubs within the Linking Open Data cloud (cf. Figure 1). A second major source of cross-domain data is Freebase, an open-license database which users can edit in a similar fashion as they edit Wikipedia today. Further cross-domain ontologies that are available as Linked Data include Wordnet, OpenCyc, YAGO and UMBEL. These ontologies are interlinked with DBpedia, which allows applications to mashup data from all sources.

Table 1 summarizes the amount of Linked Data that is available by July 2009 within each topical domain as well as the number RDF links between data sets. The table is based on the statistics collected by members of the W3C Linking Open Data effort in the project wiki.

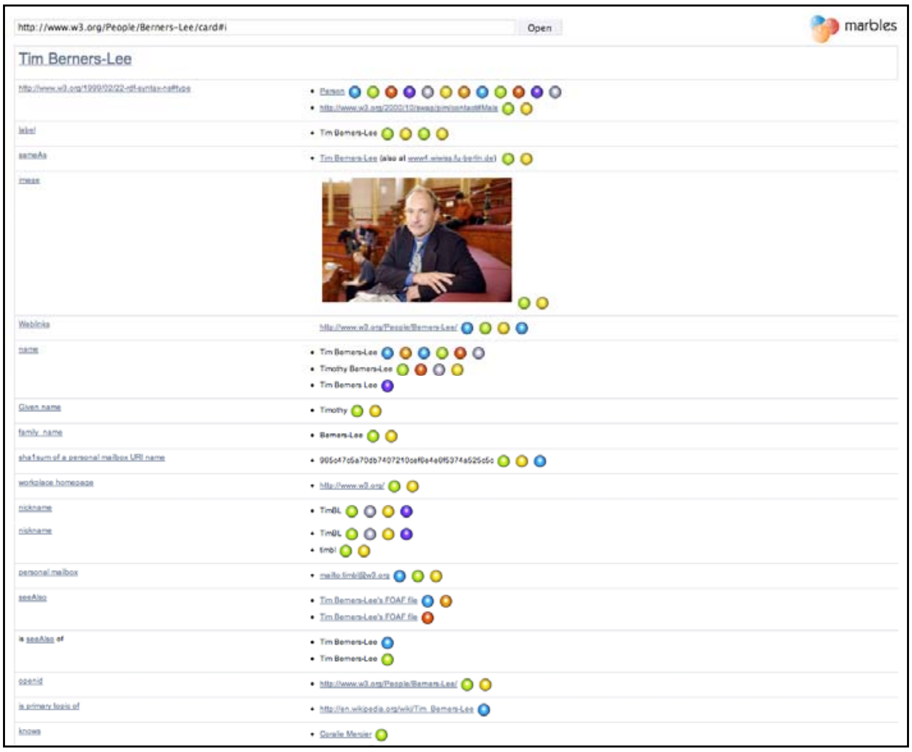
Domain	*Number of Triples*	*% of Cloud*	*Number of Links*	*% of Links*
Media	698.000.000	10,4%	1.238.000	0,8%
Publications	212.000.000	3,2%	4.922.000	3,3%
Life Sciences	2.429.000.000	36,1%	133.199.000	89,4%
Geographic Data	3.097.000.000	46,0%	4.038.000	2,7%
User Generate Content	76.000.000	1,1%	1.559.000	1,0%
Cross-Domain	214.000.000	3,2%	3.992.000	2,7%
Total	*6.726.000.000*		*148.948.000*	

Table 1: Linking Open Data data set statistics as of July 2009.

Consuming Linked Data from the Web

With significant volumes of Linked Data being published on the Web, various efforts are underway to build applications that exploit the Web of Linked Data.

There are generic data browsers which allow users to navigate between data sources along RDF links and which merge data about an entity from multiple sources that has been discovered by automatically following *owl:sameAs* links. Examples of such browsers include Tabulator, Marbles, VisiNav, razorbase, and Fenfire. Figure 2 shows the Marbles Linked Data browser displaying data about Tim Berners-Lee that has been retrieved by following data-level links from Tim's FOAF profile into various other data sources. The colored dots next to pieces of information indicate the data sources from which data was merged.



Tim Berners-Lee

http://www.w3.org/1999/02/22-rdf-syntax-ns#type	• Person • http://www.w3.org/2000/10/swap/pim/contact#Male
label	• Tim Berners-Lee
sameAs	• Tim Berners-Lee (also at www4.wiwiss.fu-berlin.de)
image	
Weblinks	http://www.w3.org/People/Berners-Lee/
name	• Tim Berners-Lee • Timothy Berners-Lee • Tim Berners Lee
Given name	• Timothy
family_name	• Berners-Lee
sha1sum of a personal mailbox URI name	• 965c47c5a70db7407210cef6e4e6f5374a525c5c
workplace homepage	• http://www.w3.org/
nickname	• TimBL
nickname	• TimBL • timbl
personal mailbox	• mailto:timbl@w3.org
seeAlso	• Tim Berners-Lee's FOAF file • Tim Berners-Lee's FOAF file
is seeAlso of	• Tim Berners-Lee • Tim Berners-Lee
openid	• http://www.w3.org/People/Berners-Lee/
is primary topic of	• http://en.wikipedia.org/wiki/Tim_Berners-Lee
knows	• Coralie Mercier

Figure 2: The Marbles Linked data browser displaying data about Tim Berners-Lee.

Examples of Linked Data search engines, that crawl the Web of Linked Data by following data-level links, include FalconS, SWSE, Sindice, Swoogle and Watson. Yahoo! and Google have also started to crawl Linked Data in its RDFa serialization. Yahoo! provides access to crawled data through the Yahoo BOSS API and is using the data within Yahoo Search Monkey to make search results more useful and visually appealing. Google uses crawled RDF data for its Social Graph API and is planning to use crawled data to enhance search results snippets for reviews and people.

Linked Data applications that target specific application domains include DBpedia Mobile, a smart phone application for tourists exploring a city; Revyu, a rating website which augments reviews with background information from the Web of Linked Data; and Talis Aspire, a resource list management tool for university courses which is used by 40.000 students at the University of Plymouth and the University of Sussex.

Linked Data and E-Government

The current uptake of the Linked Data principles in various domains raises the question about the potential of Linked Data technologies for easing the access to public sector data.

Public organizations produce a wealth of highly relevant data ranging from economic statistics, the register of companies, the land register, data about local schools, crime statistics, to the voting record of your local representative. Giving the public easy access to this data enables greater accountability, helps people to make informed choices, and allows 3rd parties create tools to analyze and work with the data.

Many public sector organizations are required by their mandate to make data resulting from their operations accessible to the general public. In practice however, various barriers hinder access to this data. The great majority of public sector data is either not accessible on the Web or if Web-accessible it is mainly found in two shapes:

1. Human readable formats such as HTML or PDF. While enabling access to people, mixing data and its presentation limits the ability of machines to process data.

2. Proprietary data formats, which require potential consumers to have proprietary software or tools to access the data.

Linked Data has the potential to overcome these barriers. Linked Data is exclusively based on open Web standards. This enables data consumers to use generic tools to access, mashup and visualize data. It also enables Web search engines to pick up data and use it to provide better services to their users. By relying on resolvable HTTP URIs for the identification of data items, it is possible to set data-level links between data sources. This allows different government bodies to relate their data to each other while every institution keeps full control of the original data.

The potential of Linked Data for easing access to public sector data is increasingly understood. The UK Prime Minister Gordon Brown has recently announced the appointment of Tim Berners-Lee as expert adviser on public information delivery. Tim Berners-Lee has published a Web design note about putting government data online [6] and is working together with the UK Power of Information Taskforce on realizing these ideas. In the Unitied States, the Obama administration has started similar efforts. The Data.gov website was recently launched and currently provides access to 47 data sets generated by the Executive Branch of the Federal Government. In order to work closer together with governments and to support public institutions in using open Web standards, the World Wide Web Consortium has formed an eGovernment Interest Group. A first result of the work of this group is the W3C note "Improving Access to Government through Better Use of the Web" [7] which highlights the benefits of open, standard-based access to government data and discusses technical options to provide such access.

Outlook

The Web has started to develop from a medium for the publication and linkage of documents into a medium for the publication and linkage of both – documents and data. With the fragmentation of the Web into distinct data island accessible through proprietary Web APIs, we are currently facing a similar situation as in the early days of the Web when services like CompuServe and AOL were trying to restrict users to content provided by a network of hand-selected affiliates. This walled garden approach has failed. Instead, the Web succeeded as a single global information space that has dramatically changed the way we use information, has disrupted business models, and let to profound societal change. With Linked Data, we are having the technologies on hand to repeat this story for data.

References

1. T. Berners-Lee, "Linked Data - Design Issues", http://www.w3.org/DesignIssues/LinkedData.html, 2006
2. C. Bizer, T. Heath, T. Berners-Lee, "Linked Data - The Story So Far", International Journal on Semantic Web & Information Systems, 2009 (in print)
3. L. Sauermann, R. Cyganiak, "Cool URIs for the Semantic Web - W3C Interest Group Note", http://www.w3.org/TR/cooluris/, 2008

4. C. Bizer, R. Cyganiak, T. Heath, “How to publish Linked Data on the Web”, http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/, 2007
5. D. Berrueta, J. Phipps, “Best Practice Recipes for Publishing RDF Vocabularies - W3C Working Group Note”, http://www.w3.org/TR/swbp-vocab-pub/, 2008
6. T. Berners-Lee, “Putting Government Data Online - Design Issues”, http://www.w3.org/DesignIssues/GovData.html, 2009
7. S. Acar, J. Alonso, K. Novak, “Improving Access to Government through Better Use of the Web - W3C Interest Group Note”, http://www.w3.org/TR/egov-improving/, 2009

Bio and Photo

Professor Christian Bizer is the head of the Web-based Systems Group at Freie Universität Berlin. The group explores technical and economic questions concerning the development of global, decentralized information environments. He initialized the W3C Linking Open Data community project and the DBpedia project.