Linked Data - The Story So Far

Christian Bizer, Freie Universität Berlin, Germany
Tom Heath, Talis Information Ltd, United Kingdom
Tim Berners-Lee, Massachusetts Institute of Technology, USA

Abstract

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. These 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. In this article we present the concept and technical principles of Linked Data, and situate these within the broader context of related technological developments. We describe progress to date in publishing Linked Data on the Web, review applications that have been developed to exploit the Web of Data, and map out a research agenda for the Linked Data community as it moves forward.

Keywords: Linked Data, Web of Data, Semantic Web, Data Sharing, Data Exploration

1. Introduction

The World Wide Web has radically altered the way we share knowledge by lowering the barrier to publishing and accessing documents as part of a global information space. Hypertext links allow users to traverse this information space using Web browsers, while search engines index the documents and analyse the structure of links between them to infer potential relevance to users' search queries (Brin & Page, 1998). This functionality has been enabled by the generic, open and extensible nature of the Web (Jacobs & Walsh, 2004), which is also seen as a key feature in the Web's unconstrained growth.

Despite the inarguable benefits the Web provides, until recently the same principles that enabled the Web of documents to flourish have not been applied to data. Traditionally, data published on the Web has been made available as raw dumps in formats such as CSV or XML, or marked up as HTML tables, sacrificing much of its structure and semantics. In the conventional hypertext Web, the nature of the relationship between two linked documents is implicit, as the data format, i.e. HTML, is not sufficiently expressive to enable individual entities described in a particular document to be connected by typed links to related entities.

However, in recent years the Web has evolved from a global information space of linked documents to one where both documents and data are linked. Underpinning this evolution is a set of best practices for publishing and connecting structured data on the Web known as

Linked Data. The adoption of the Linked Data best practices has lead to the extension of the Web with a global data space connecting data from diverse domains such as people, companies, books, scientific publications, films, music, television and radio programmes, genes, proteins, drugs and clinical trials, online communities, statistical and scientific data, and reviews. This Web of Data enables new types of applications. There are generic Linked Data browsers which allow users to start browsing in one data source and then navigate along links into related data sources. There are Linked Data search engines that crawl the Web of Data by following links between data sources and provide expressive query capabilities over aggregated data, similar to how a local database is queried today. The Web of Data also opens up new possibilities for domain-specific applications. Unlike Web 2.0 mashups which work against a fixed set of data sources, Linked Data applications operate on top of an unbound, global data space. This enables them to deliver more complete answers as new data sources appear on the Web.

The remainder of this paper is structured as follows. In Section 2 we provide an overview of the key features of Linked Data. Section 3 describes the activities and outputs of the Linking Open Data project, a community effort to apply the Linked Data principles to data published under open licenses. The state of the art in publishing Linked Data is reviewed in Section 4, while section 5 gives an overview of Linked Data applications. Section 6 compares Linked Data to other technologies for publishing structured data on the Web, before we discuss ongoing research challenges in Section 7.

2. What is Linked Data?

In summary, Linked Data is simply about using the Web to create typed links between data from different sources. These may be as diverse as databases maintained by two organisations in different geographical locations, or simply heterogeneous systems within one organisation that, historically, have not easily interoperated at the data level. Technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets.

While the primary units of the hypertext Web are HTML (HyperText Markup Language) documents connected by untyped hyperlinks, Linked Data relies on documents containing data in RDF (Resource Description Framework) format (Klyne and Carroll, 2004). However, rather than simply connecting these documents, Linked Data uses RDF to make typed statements that link arbitrary things in the world. The result, which we will refer to as the Web of Data, may more accurately be described as *a web of things in the world, described by data on the Web.*

Berners-Lee (2006) outlined a set of 'rules' 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', and provide a basic recipe for publishing and connecting data using the infrastructure of the Web while adhering to its architecture and standards.

The Linked Data Technology Stack

Linked Data relies on two technologies that are fundamental to the Web: Uniform Resource Identifiers (URIs) (Berners-Lee et al., 2005) and the HyperText Transfer Protocol (HTTP) (Fielding et al., 1999). While Uniform Resource Locators (URLs) have become familiar as addresses for documents and other entities that can be located on the Web, Uniform Resource Identifiers provide a more generic means to identify any entity that exists in the world.

Where entities are identified by URIs that use the *http://* scheme, these entities can be looked up simply by dereferencing the URI over the HTTP protocol. In this way, the HTTP protocol provides a simple yet universal mechanism for retrieving resources that can be serialised as a stream of bytes (such as a photograph of a dog), or retrieving descriptions of entities that cannot themselves be sent across the network in this way (such as the dog itself).

URIs and HTTP are supplemented by a technology that is critical to the Web of Data – RDF, introduced above. Whilst HTML 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.

The RDF model encodes data in the form of *subject*, *predicate*, *object* triples. 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.

For example, an RDF triple can state that two people, *A* and *B*, each identified by a URI, are related by the fact that *A* knows *B*. Similarly an RDF triple may relate a person *C* to a scientific article *D* in a bibliographic database by stating that *C* is the author of *D*. Two resources linked in this fashion can be drawn from different data sets on the Web, allowing data in one data source to be linked to that in another, thereby creating a Web of Data. Consequently it is possible to think of RDF triples that link items in different data sets as analogous to the hypertext links that tie together the Web of documents.

RDF links (Bizer & Cyganiak & Heath, 2007) take the form of RDF triples, where the subject of the triple is a URI reference in the namespace of one data set, while the object of the triple is a URI reference in the other. Figure 1 shows two example RDF links. The first link states that a resource identified by the URI *http://www.w3.org/People/Berners-Lee/card#i* is member of another resource called *http://dig.csail.mit.edu/data#DIG*. When the subject URI is dereferenced over the HTTP protocol, the *dig.csail.mit.edu* server answers with a RDF description of the identified resource, in this case the MIT Decentralized Information Group. When the object URI is dereferenced the W3C server provides an RDF graph describing Tim Berners-Lee. Dereferencing the predicate URI *http://xmlns.com/foaf/0.1/member* yields a definition of the link type *member*, described in RDF using the RDF Vocabulary Definition Language (RDFS), introduced below. The second RDF link connects the description of the film Pulp Fiction in the Linked Movie Database with the description of the film provided by DBpedia, by stating that the URI *http://data.linkedmdb.org/resource/film/77* and the URI *http://dbpedia.org/resource/Pulp_Fiction_%28film%29* refer to the same real-world entity - the film Pulp Fiction.

Subject: `http://dig.csail.mit.edu/data#DIG`
Predicate: `http://xmlns.com/foaf/0.1/member`
Object: `http://www.w3.org/People/Berners-Lee/card#i`

Subject: `http://data.linkedmdb.org/resource/film/77`
Predicate: `http://www.w3.org/2002/07/owl#sameAs`
Object: `http://dbpedia.org/resource/Pulp_Fiction_%28film%29`

Figure 1. Example RDF links

The RDF Vocabulary Definition Language (RDFS) (Brickley & Guha, 2004) and the Web Ontology Language (OWL) (McGuinness & van Harmelen, 2004) provide a basis for creating vocabularies that can be used to describe entities in the world and how they are related. Vocabularies are collections of classes and properties. Vocabularies are themselves expressed in RDF, using terms from RDFS and OWL, which provide varying degrees of expressivity in modelling domains of interest. Anyone is free to publish vocabularies to the Web of Data (Berrueta & Phipps, 2008), which in turn can be connected by RDF triples that link classes and properties in one vocabulary to those in another, thereby defining mappings between related vocabularies.

By employing HTTP URIs to identify resources, the HTTP protocol as retrieval mechanism, and the RDF data model to represent resource descriptions, Linked Data directly builds on the general architecture of the Web (Jacobs & Walsh, 2004). The Web of Data can therefore be seen as an additional layer that is tightly interwoven with the classic document Web and has many of the same properties:

- The Web of Data is generic and can contain any type of data.
- Anyone can publish data to the Web of Data.
- Data publishers are not constrained in choice of vocabularies with which to represent data.
- Entities are connected by RDF links, creating a global data graph that spans data sources and enables the discovery of new data sources.

From an application development perspective the Web of Data has the following characteristics:

- Data is strictly separated from formatting and presentational aspects.
- Data is self-describing. If an application consuming Linked Data encounters data described with an unfamiliar vocabulary, the application can dereference the URIs that identify vocabulary terms in order to find their definition.
- The use of HTTP as a standardized data access mechanism and RDF as a standardized data model simplifies data access compared to Web APIs, which rely on heterogeneous data models and access interfaces.
- The Web of Data is open, meaning that applications do not have to be implemented against a fixed set of data sources, but can discover new data sources at run-time by following RDF links.

3. The Linking Open Data Project

The most visible example of adoption and application of the Linked Data principles has been the Linking Open Data project [Endnote: http://esw.w3.org/topic/SweoIG/TaskForces/

CommunityProjects/LinkingOpenData], a grassroots community effort founded in January 2007 and supported by the W3C Semantic Web Education and Outreach Group [Endnote: http://www.w3.org/2001/sw/sweo/]. The original and ongoing aim of the project is to bootstrap the Web of Data by identifying existing data sets that are available under open licenses, converting these to RDF according to the Linked Data principles, and publishing them on the Web.

Participants in the early stages of the project were primarily researchers and developers in university research labs and small companies. Since that time the project has grown considerably, to include significant involvement from large organisations such as the BBC, Thomson Reuters and the Library of Congress. This growth is enabled by the open nature of the project, where anyone can participate simply by publishing a data set according to the Linked Data principles and interlinking it with existing data sets. An indication of the range and scale of the Web of Data originating from the Linking Open Data project is provided in Figure 2. Each node in this cloud diagram represents a distinct data set published as Linked Data, as of March 2009.
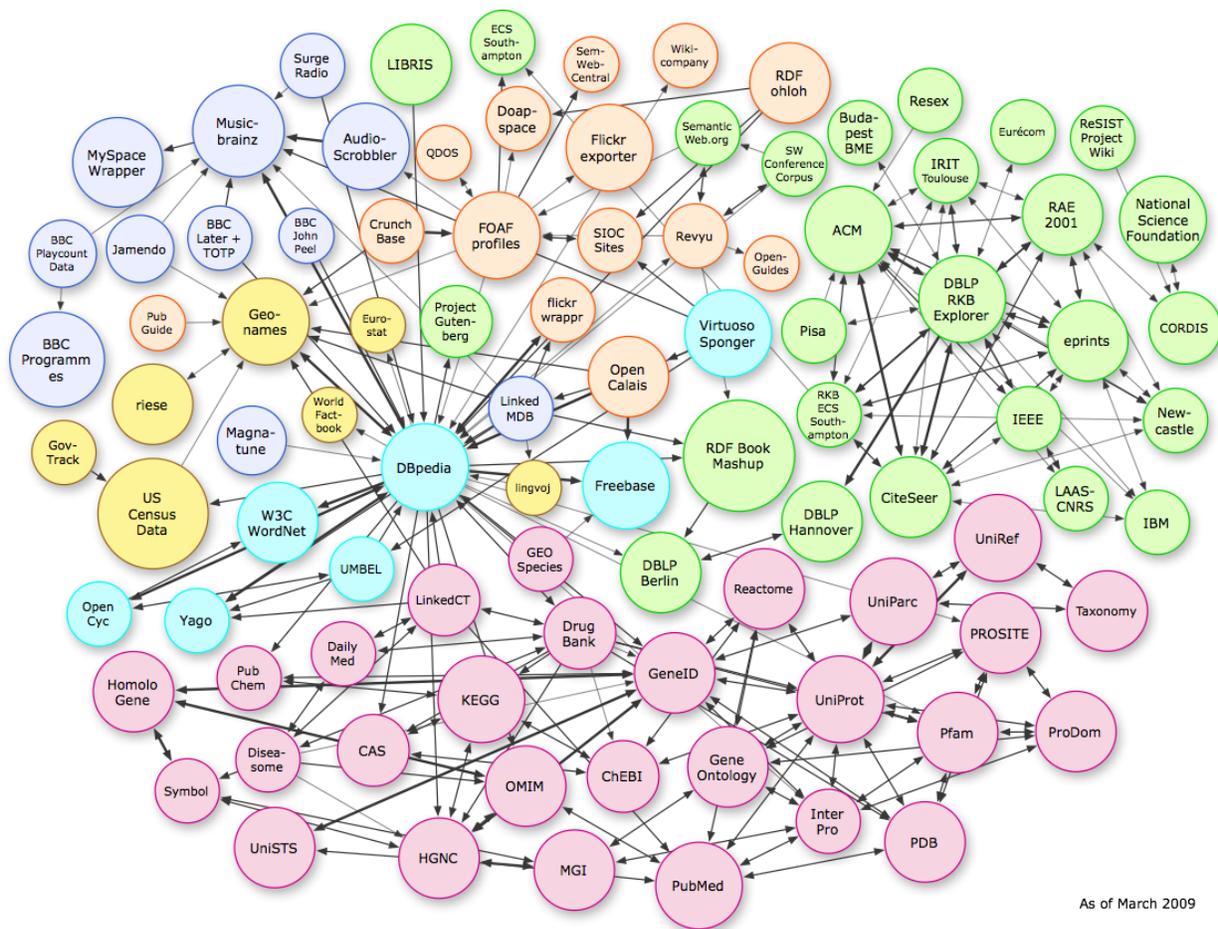

Figure 2. Linking Open Data cloud diagram giving an overview of published data sets and their interlinkage relationships.

The arcs in Figure 2 indicate that links exist between items in the two connected data sets. Heavier arcs roughly correspond to a greater number of links between two data sets, while bidirectional arcs indicate the outward links to the other exist in each data set.

The content of the cloud is diverse in nature, comprising data about geographic locations, people, companies, books (Bizer & Cyganiak & Gauss, 2007), scientific publications (Van de Sompel et al., 2009), films (Hassanzadeh & Consens, 2009), music, television and radio programmes (Kobilarov et al, 2009), genes, proteins, drugs and clinical trials (Belleau et al., 2008, Jentzsch et al., 2009), online communities, statistical data, census results, and reviews (Heath & Motta, 2008).

Calculating the exact size of the Web of Data is challenging due to the fact that much of the data is being generated by wrappers around existing relational databases or APIs and therefore first need to be crawled before it can be counted or analyzed (Hausenblas et al., 2008). Alternatively, the size of the Web of Data can be estimated based on the data set statistics that are collected by the LOD community in the ESW wiki. According to these statistics, the Web of Data currently consists of 4.7 billion RDF triples, which are interlinked by around 142 million RDF links (May 2009). [Endnote: http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/LinkStatistics and http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics]

As Figure 2 shows, certain data sets serve as linking hubs in the Web of Data. For example, the DBpedia data set (Auer et al., 2007) consists of RDF triples extracted from the "infoboxes" commonly seen on the right hand side of Wikipedia articles, while Geonames [Endnote: http://www.geonames.org/ontology/] provides RDF descriptions of millions of geographical locations worldwide. As these two data sets provide URIs and RDF descriptions for many common entities or concepts, they are frequently referenced in other more specialised data sets and have therefore developed into hubs to which an increasing number of other data sets are connected.

4. Publishing Linked Data on the Web

By publishing data on the Web according to the Linked Data principles, data providers add their data to a global data space, which allows data to be discovered and used by various applications. Publishing a data set as Linked Data on the Web involves the following three basic steps:

1. Assign URIs to the entities described by the data set and provide for dereferencing these URIs over the HTTP protocol into RDF representations.
2. Set RDF links to other data sources on the Web, so that clients can navigate the Web of Data as a whole by following RDF links.
3. Provide metadata about published data, so that clients can assess the quality of published data and choose between different means of access.

In the following, we will give an overview about each of these tasks as well as about tools that have been developed to support publishers with each task.

Choosing URIs and RDF Vocabularies

Data providers can choose between two HTTP URI usage patterns to identify entities: 303 URIs and hash URIs. Both patterns ensure that clients can distinguish between URIs that identify real-world entities and URIs that identify Web documents describing these real-world entities (Sauermann & Cyganiak, 2008). In an open environment like the Web, different information providers publish data about the same real-world entity, for instance a geographic location or a celebrity. As they may not know about each other, they introduce

different URIs to identify the same entitiy. For instance, DBpedia uses the URI *http://dbpedia.org/resource/Berlin* to identify Berlin, while Geonames uses the URI *http://sws.geonames.org/2950159/* to identify Berlin. As both URIs refer to the same real-world entity, they are called URI aliases. URI aliases are common on the Web of Data, as it can not realistically be expected that all information providers agree on the same URIs to identify an entity. URI aliases also provide an important social function to the Web of Data as they are dereferenced to different descriptions of the same real-world entity and thus allow different views and opinions to be expressed on the Web. In order to still be able to track that different information providers speak about the same entity, it is common practice that information providers set *owl:sameAs* links to URI aliases they know about.

Different communities have specific preferences on the vocabularies they prefer to use for publishing data on the Web. The Web of Data is therefore open to arbitrary vocabularies being used in parallel. Despite this general openness, it is considered good practice to reuse terms from well-known RDF vocabularies such as FOAF, SIOC, SKOS, DOAP, vCard, Dublin Core, OAI-ORE or GoodRelations wherever possible in order to make it easier for client applications to process Linked Data. Only if these vocabularies do not provide the required terms should data publishers define new, data source-specific terminology (Bizer & Cyganiak & Heath, 2007). If new terminology is defined, it should be made self-describing by making the URIs that identify terms Web dereferencable (Berrueta & Phipps, 2008). This allows clients to retrieve RDF Schema or OWL definitions of the terms as well as term mappings to other vocabularies. The Web of Data thus relies on a pay as you go data integration approach (Das Sarma & Dong & Halevy, 2008) based on a mixture of using common vocabularies together with data source-specific terms that are connected by mappings as deemed necessary.

A common serialization format for Linked Data is RDF/XML (Beckett, 2004). In situations where human inspection of RDF data is required, Notation3 (Berners-Lee, 1998), and its subset Turtle (Beckett and Berners-Lee, 2008), are often provided as alternative, inter-convertible serializations, due to the greater perceived readability of these formats. Alternatively, Linked Data can also be serialized as RDFa (Adida et al., 2008) which provides for embedding RDF triples into HTML. In the second case, data publishers should use the RDFa *about* attribute to assign URIs to entities in order to allow other data providers to set RDF links to them.

Link Generation

RDF links allow client applications to navigate between data sources and to discover additional data. In order to be part of the Web of Data, data sources should set RDF links to related entities in other data sources. As data sources often provide information about large numbers of entities, it is common practice to use automated or semi-automated approaches to generate RDF links.

In various domains, there are generally accepted naming schemata. For instance, in the publication domain there are ISBN and ISSN numbers, in the financial domain there are ISIN identifiers, EAN and EPC codes are widely used to identify products, in life science various accepted identification schemata exist for genes, molecules, and chemical substances. If the link source and the link target data sets already both support one of these identification schema, the implicit relationship between entities in both data sets can easily be made explicit as RDF links. This approach has been used to generate links between various data sources in the LOD cloud.

If no shared naming schema exist, RDF links are often generated based on the similarity of entities within both data sets. Such similarity computations can build on a large body of related work on record linkage (Winkler, 2006) and duplicate detection (Elmagarmid et al., 2007) within the database community as well as on ontology matching (Euzenat & Shvaiko, 2007) in the knowledge representation community. An example of a similarity based interlinking algorithm is presented in (Raimond et al., 2008). In order to set RDF links between artists in the Jamendo and Musicbrainz data sets, the authors use a similarity metric that compares the names of artists as well as the titles of their albums and songs.

Various RDF link generation frameworks are available, that provide declarative languages for specifying which types of RDF links should be created, which combination of similarity metrics should be used to compare entities and how similarity scores for specific properties are aggregated into an overall score. The Silk framework (Volz et al., 2009) works against local and remote SPARQL [Endnote: http://www.w3.org/TR/rdf-sparql-query/] endpoints and is designed to be employed in distributed environments without having to replicate data sets locally. The LinQL framework (Hassanzadeh et al., 2009) works over relational databases and is designed to be used together with database to RDF mapping tools such as D2R Server or Virtuoso.

Metadata

Linked Data should be published alongside several types of metadata, in order to increase its utility for data consumers. In order to enable clients to assess the quality of published data and to determine whether they want to trust data, data should be accompanied with meta-information about its creator, its creation date as well as the creation method (Hartig, 2009). Basic provenance meta-information can be provided using Dublin Core terms or the Semantic Web Publishing vocabulary (Carroll et al., 2005). The Open Provenance Model (Moreau et al., 2008) provides terms for describing data transformation workflows. In (Zhao et al., 2008), the authors propose a method for providing evidence for RDF links and for tracing how the RDF links change over time

In order to support clients in choosing the most efficient way to access Web data for the specific task they have to perform, data publishers can provide additional technical metadata about their data set and its interlinkage relationships with other data sets: The Semantic Web Crawling sitemap extension (Cyganiak et al., 2008) allows data publishers to state which alternative means of access (SPARQL endpoint, RDF dumps) are provided besides dereferenceable URIs. The Vocabulary Of Interlinked Datasets (Alexander et al., 2009) defines terms and best practices to categorize and provide statistical meta-information about data sets as well as the linksets connecting them.

Publishing Tools

A variety of Linked Data publishing tools has been developed. The tools either serve the content of RDF stores as Linked Data on the Web or provide Linked Data views over non-RDF legacy data sources. The tools shield publishers from dealing with technical details such as content negotiation and ensure that data is published according to the Linked Data community best practices (Sauermann & Cyganiak, 2008; Berrueta & Phipps, 2008; Bizer & Cyganiak & Heath, 2007). All tools support dereferencing URIs into RDF descriptions. In addition, some of the tools also provide SPARQL query access to the served data sets and support the publication of RDF dumps.

- **D2R Server.** D2R Server (Bizer & Cyganiak, 2006) is a tool for publishing non-RDF relational databases as Linked Data on the Web. Using a declarative mapping

language, the data publisher defines a mapping between the relational schema of the database and the target RDF vocabulary. Based on the mapping, D2R server publishes a Linked Data view over the database and allows clients to query the database via the SPARQL protocol.

- **Virtuoso Universal Server.** The OpenLink Virtuoso server[Endnote: http://www.openlinksw.com/dataspace/dav/wiki/Main/VOSRDF] provides for serving RDF data via a Linked Data interface and a SPARQL endpoint. RDF data can either be stored directly in Virtuoso or can be created on the fly from non-RDF relational databases based on a mapping.
- **Talis Platform**. The Talis Platform[Endnote: http://www.talis.com/platform/] is delivered as Software as a Service accessed over HTTP, and provides native storage for RDF/Linked Data. Access rights permitting, the contents of each Talis Platform store are accessible via a SPARQL endpoint and a series of REST APIs that adhere to the Linked Data principles.
- **Pubby.** The Pubby server (Cyganiak & Bizer, 2008) can be used as an extension to any RDF store that supports SPARQL. Pubby rewrites URI requests into SPARQL DESCRIBE queries against the underlying RDF store. Besides RDF, Pubby also provides a simple HTML view over the data store and takes care of handling 303 redirects and content negotiation between the two representations.
- **Triplify.** The Triplify toolkit (Auer et al, 2009) supports developers in extending existing Web applications with Linked Data front-ends. Based on SQL query templates, Triplify serves a Linked Data and a JSON view over the application's database.
- **SparqPlug**. SparqPlug (Coetzee, Heath and Motta, 2008) is a service that enables the extraction of Linked Data from legacy HTML documents on the Web that do not contain RDF data. The service operates by serialising the HTML DOM as RDF and allowing users to define SPARQL queries that transform elements of this into an RDF graph of their choice.
- **OAI2LOD Server.** The OAI2LOD (Haslhofer & Schandl, 2008) is a Linked Data wrapper for document servers that support the Open Archives OAI-RMH protocol.
- **SIOC Exporters.** The SIOC project has developed Linked Data wrappers for several popular blogging engines, content management systems and discussion forums such as WordPress, Drupal, and phpBB [Endnote: http://sioc-project.org/exporters].

A service that helps publishers to debug their Linked Data site is the Vapour validation service [Endnote: http://vapour.sourceforge.net/]. Vapour verifies that published data complies with the Linked Data principles and community best practices.

5. Linked Data Applications

With significant volumes of Linked Data being published on the Web, numerous efforts are underway to research and build applications that exploit this Web of Data. At present these efforts can be broadly classified into three categories: Linked Data browsers, Linked Data search engines, and domain-specific Linked Data applications. In the following section we will examine each of these categories.
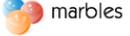
Linked Data Browsers

Just as traditional Web browsers allow users to navigate between HTML pages by following hypertext links, Linked Data browsers allow users to navigate between data sources by following links expressed as RDF triples. For example, a user may view DBpedia's RDF description of the city of Birmingham (UK), follow a 'birthplace' link to the description of the

comedian Tony Hancock (who was born in the city), and from there onward into RDF data from the BBC describing broadcasts in which Hancock starred. The result is that a user may begin navigation in one data source and progressively traverse the Web by following RDF rather than HTML links. The Disco hyperdata browser [Endnote: http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/] follows this approach and can be seen as a direct application of the hypertext navigation paradigm to the Web of Data.

Data, however, provides human interface opportunities and challenges beyond those of the hypertext Web. People need to be able to explore the Web of links between items, but also to powerfully analyze data in bulk. The Tabulator (Berners-Lee et al, 2006; Berners-Lee et al, 2008), for example, allows the user traverse the Web of Data, and expose pieces of it in a controlled way, in "outline mode"; to discover and highlight a pattern of interest; and then query for any other similar patterns in the data Web. The results of the query form a table that can then be analyzed with various conventional data presentation methods, such as faceted browsers, maps, timelines, and so on.

Tabulator and Marbles (Becker & Bizer, 2008) (see Figure 3) are among the data browsers which track the provenance of data, while merging data about the same thing from different sources. While authors such as (Karger & schraefel, 2006) have questioned the use of graph-oriented views over RDF data, as seen in browsers such as FOAFNaut [Endnote: http://www.jibbering.com/foaf/], (Hastrup, Cyganiak & Bojars, 2008) argue that such interfaces fill an important niche, and describe their Fenfire browser that follows this display paradigm.

Figure 3. The Marbles Linked Data browser displaying data about Tim Berners-Lee. The colored dots indicate the data sources from which data was merged.

Linked Data Search Engines and Indexes

In the traditional hypertext Web, browsing and searching are often seen as the two dominant modes of interaction (Olston & Chi, 2003). While browsers provide the mechanisms for navigating the information space, search engines are often the place at which that navigation process begins. A number of search engines have been developed that crawl Linked Data from the Web by following RDF links, and provide query capabilities over aggregated data. Broadly speaking, these services can be divided into two categories: human-oriented search engines, and application-oriented indexes.

Human-oriented Search Engines

Search engines such as Falcons (Cheng & Qu, this issue) and SWSE (Hogan et al., 2007) provide keyword-based search services oriented towards human users, and follow a similar interaction paradigm to existing market leaders such as Google and Yahoo. The user is

presented with a search box into which they can enter keywords related to the item or topic in which they are interested, and the application returns a list of results that may be relevant to the query. However, rather than simply providing links from search results through to the source documents in which the queried keywords are mentioned, both SWSE and Falcons provide a more detailed interface to the user that exploits the underlying structure of the data. Both provide a summary of the entity the user selects from the results list, alongside additional structured data crawled from the Web and links to related entities.

Falcons provides users with the option of searching for objects, concepts and documents, each of which leads to slightly different presentation of results. While the object search (Figure 4.) is suited to searching for people, places and other more concrete items, the concept search is oriented to locating classes and properties in ontologies published on the Web. The document search feature provides a more traditional search engine experience, where results point to RDF documents that contain the specified search terms.

It is worth noting that, while they may be referred to as distinct entities, the document Web and the data Web form one connected, navigable information space. For example, a user may perform a search in the existing document Web, follow a link from an HTML document into the Web of Data, navigate this space for some time, and then follow a link to a different HTML document, and so on.

Figure 4. Falcons object search results for the keyword 'Berlin'.

It is interesting to note that while both SWSE and Falcons operate over corpuses of structured data crawled from the Web, they choose to provide very simple query capabilities

that mimic the query interfaces of conventional Web search engines. While one may intuitively expect the additional structure in the data to be exploited to provide sophisticated query capabilities for advanced users at least, this has not proved to be the case to date, with the exception of Tabulator's style of query-by-example and faceted browsing interfaces for query refinement. SWSE does provide access to its underlying data store via the SPARQL query language, however this is suitable primarily for application developers with a knowledge of the language rather than regular users wishing to ask very specific questions through a usable human interface.

Application-oriented Indexes

While SWSE and Falcons provide search capabilities oriented towards humans, another breed of services have been developed to serve the needs of applications built on top of distributed Linked Data. These application-oriented indexes, such as Swoogle (Ding et al, 2005), Sindice (Oren et al, 2008) and Watson (d'Aquin et al, 2008) provide APIs through which Linked Data applications can discover RDF documents on the Web that reference a certain URI or contain certain keywords. The rationale for such services is that each new Linked Data application should not need to implement its own infrastructure for crawling and indexing all parts of the Web of Data of which it might wish to make use. Instead, applications can query these indexes to receive pointers to potentially relevant documents which can then be retrieved and processed by the application itself. Despite this common theme, these services have slightly different emphases. Sindice is oriented more to providing access to documents containing instance data, while in contrast the emphasis of Swoogle and Watson is on finding ontologies that provide coverage of certain concepts relevant to a query.

Domain-specific Applications

While the Linked Data browsers and search engines described above provide largely generic functionality, a number of services have been developed that offer more domain-specific functionality by 'mashing up' data from various Linked Data sources.

Revyu

Revyu (Heath & Motta, 2008) is a generic reviewing and rating site based on Linked Data principles and the Semantic Web technology stack. In addition to publishing Linked Data, Revyu consumes Linked Data from the Web to enhance the experience of site users. For example, when films are reviewed on Revyu, the site attempts to match these with the corresponding entry in DBpedia. Where a match is made, additional information about the film (such as the director's name and the film poster) is retrieved from DBpedia and shown in the human-oriented (HTML) pages of the site. In addition, links are made at the RDF level to the corresponding item, ensuring that while human users see a richer view of the item through the mashing up of data from various sources, Linked Data-aware applications are provided with references to URIs from which related data may be retrieved. Similar principles are followed to link items such as books and pubs to corresponding entries in external data sets, and to enhance user profiles with FOAF data.

DBpedia Mobile

DBpedia Mobile (Becker & Bizer, 2008) is a location-aware Linked Data browser designed to be run on an iPhone or other mobile device. DBpedia Mobile is oriented to the use case of a

tourist exploring a city. Based on the current GPS position of the mobile device, the application provides a location-centric mashup of nearby locations from DBpedia, associated reviews from Revyu, and related photos via a Linked Data wrapper around the Flickr photo-sharing API. Figure 5 shows DBpedia Mobile displaying data from DBpedia and Revyu about the Brandenburg Gate in Berlin. Besides accessing Web data, DBpedia Mobile also enables users to publish their current location, pictures and reviews to the Web as Linked Data, so that they can be used by other applications. Instead of simply being tagged with geographical coordinates, published content is interlinked with a nearby DBpedia resource and thus contributes to the overall richness of the Web of Data.

Figure 5. DBpedia Mobile displaying information about Berlin

Talis Aspire

Talis Aspire (Clarke, 2009) is a Web-based Resource List Management application deployed to university lecturers and students. As users create lists through a conventional Web interface, the application produces RDF triples which are persisted to an underlying Linked Data-compatible store. The use of Linked Data principles enables items present on one list to be transparently linked to the corresponding items featured on lists at other institutions, thereby building a Web of scholarly data through the actions of non-specialist users.

BBC Programmes and Music

The British Broadcasting Corporation (BBC) uses Linked Data internally as a lightweight data integration technology. The BBC runs numerous radio stations and television channels. Traditionally, these stations and channels use separate content management systems. The BBC has thus started to use Linked Data technologies together with DBpedia and MusicBrainz as controlled vocabularies to connect content about the same topic residing in different repositories and to augment content with additional data from the Linking Open Data cloud. Based on these connections, BBC Programmes and BBC Music build Linked Data sites for all of its music and programmes related brands (Kobilarov et al., 2009).

DERI Pipes

Modelled on Yahoo Pipes, DERI Pipes (Le Phuoc et al. 2009) provides a data level mashup platform that enables data sources to be plugged together to form new feeds of data. The resulting aggregation workflows may contain sophisticated operations such as identifier consolidation, schema mapping, RDFS or OWL reasoning, with data transformations being expressed using SPARQL CONSTRUCT operations or XSLT templates. Figure 6. shows the assembly of a workflow to integrate data about Tim Berners-Lee within the DERI pipes development environment.
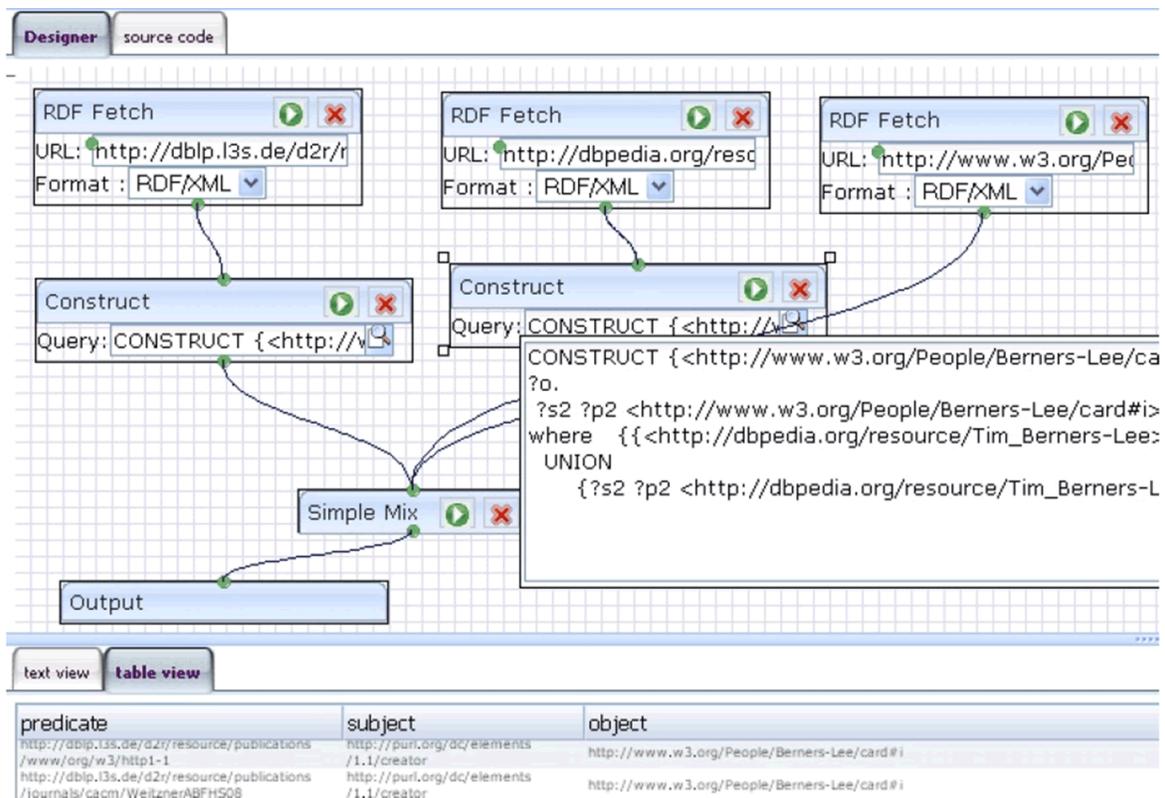

Figure 6. DERI pipes workflow integrating data about Tim Berners-Lee from three data sources.

6. Related Developments (in Research and Practice)

There are several other developments related to Linked Data happening on the Web or being pursued by related research communities. In the following sections, we will compare these developments with Linked Data.

Microformats

Similar to Linked Data, Microformats[Endnote: http://microformats.org/] aim at extending the Web with structured data. Microformats define a set of simple data formats that are embedded into HTML via class attributes. Two major differences between Microformats and Linked Data in its RDFa serialization are: Linked Data is not limited in the vocabularies that can be used to represent data, and the vocabulary development process itself is completely open, while Microformats are restricted to a small set of vocabularies developed through a process closely managed by a specific community. Data items that are included in HTML pages via Microformats do not have their own identifier. This prevents the assertion, across documents and Web sites, of relationships between data items. By using URIs as global identifiers and RDF to represent relationships, Linked Data does not have these limitations.

Web APIs

Many major Web data sources such as Amazon, eBay, Yahoo!, and Google provide access to their data via Web APIs. The website ProgrammableWeb.com currently lists 1309 Web APIs as well as 3966 mashups based on these APIs. Web APIs are accessed using a wide range of different mechanisms, and data retrieved from these APIs is represented using various content formats. In contrast, Linked Data commits itself to a small set of standardized technologies: URIs and HTTP as identification and access mechanism, RDF as content format. Using a single set of technologies instead of relying on diverse interfaces and result formats allows data sources to be more easily crawled by search engines and accessed using generic data browsers. Beside these technical details, there is also a major conceptual difference between Web APIs and Linked Data: most Web APIs do not assign globally unique identifiers to data items. Therefore it is not possible to set links between items in different data sources in order to connect data into a global data space. Mashups based on these APIs are therefore always implemented against a fixed set of data sources. In contrast, Linked Data applications can work on top of an unbounded, global data space. They can discover new data sources by following RDF links and take advantage of new data sources as they appear on the Web without needing to change the application code. Therefore, Linked Data technologies can contribute to connecting the different data silos that currently exist on the Web back into the single global information space.

Dataspaces

A recent concept within the databases community that is very similar to Linked Data is dataspaces (Franklin et al. 2005). Dataspaces provide a target system architecture around which ongoing research on reference reconciliation, schema matching and mapping, data lineage, data quality and information extraction are unified (Halevy et al., 2006). In contrast with other information-integration systems, dataspaces systems offer best-effort answers before complete semantic mappings are provided to the system. A key idea of dataspaces is that the semantic cohesion of a dataspace is increased over time by different parties providing mappings; the same pay as you go data integration approach that currently emerges on the Web of Data. The Web of Data can therefore be seen as a realization of the dataspaces concept on global scale, relying on a specific set of Web standards in order to be

closely aligned with the overall architecture of the Web. It is therefore likely that the Web of Data will benefit considerably from research into dataspaces that is ongoing in the database community.

Semantic Web

The desire to extend the capabilities of the Web to publishing of structured data is not new, and can be traced back to the earliest proposal for the World Wide Web [Endnote: http://www.w3.org/History/1989/proposal.html] and subsequent papers on the topic (Berners-Lee et al., 1994). Trends foreseen at these early stages of the Web's existence included "Evolution of objects from being principally human-readable documents to contain more machine-oriented semantic information" (Berners-Lee et al., 1994), which can be seen as the seeds of an idea that became known as the Semantic Web.

The vision of a Semantic Web has been interpreted in many different ways (e.g. Berners-Lee, Hendler & Lassila, 2001; Marshall & Shipman, 2003). However, despite this diversity in interpretation, the original goal of building a global Web of machine-readable data remains constant across the original literature on the subject. According to (Berners-Lee, 2000, pp.191), "The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web – a web of data that can be processed directly or indirectly by machines". Therefore, while the Semantic Web, or Web of Data, is the goal or the end result of this process, Linked Data provides the means to reach that goal.

By publishing Linked Data, numerous individuals and groups have contributed to the building of a Web of Data, which can lower the barrier to reuse, integration and application of data from multiple, distributed and heterogeneous sources. Over time, with Linked Data as a foundation, some of the more sophisticated proposals associated with the Semantic Web vision, such as intelligent agents, may become a reality.

7. Research Challenges

By publishing and interlinking various data sources on the Web, the Linking Open Data community has created an crystallization point for the Web of Data and a challenging test bed for Linked Data technologies. However, to address the ultimate goal of being able to use the Web like a single global database, various remaining research challenges must be overcome.

User Interfaces and Interaction Paradigms

Arguably the key benefit of Linked Data from the user perspective is the provision of integrated access to data from a wide range of distributed and heterogeneous data sources. By definition, this may involve integration of data from sources not explicitly selected by users, as to do so would likely incur an unacceptable cognitive overhead. While the browsers described in Section 5 demonstrate promising trends in how applications may be developed that exploit Linked Data, numerous challenges remain in understanding appropriate user interaction paradigms for applications built on data assembled dynamically in this fashion (Heath, 2008b). For example, while hypertext browsers provide mechanisms for navigation forwards and backwards in a document-centric information space, similar navigation controls in a Linked Data browser should enable the user to move forwards and backwards between entities, thereby changing the focal point of the application. Linked Data browsers will also need to provide intuitive and effective mechanisms for adding and

removing data sources from an integrated, entity-centric view. Sigma (Catasta & Cyganiak & Tummarello, 2009), a search engine based on the Sindice service, gives an indication of how such functionality could be delivered. However understanding how such an interface can be realised when data sources number in the thousands or millions is a captivating research challenge.

Application Architectures

In principle, Linked Data may be accessed through advance crawling and caching, or on-the-fly at application runtime through link traversal or federated querying. Search engines such as SWSE, Sindice, Falcons, and Watson crawl the Web of Data and provide applications with access to crawled data through APIs. Federated query architectures for Linked Data include DARQ (Quilitz & Leser, 2008) and SemaPlorer (Schenk et al., 2008). The Semantic Web Client Library[Endnote: http://www4.wiwiss.fu-berlin.de/bizer/ng4j/semwebclient/] and SQUIN [Endnote: http://squin.org/] have demonstrated that expressive queries can be answered against the Web of Data by relying on runtime link traversal. The appropriate mixture of these methods will always depend on the specific needs of a Linked Data application. However, due to the likelihood of scalability problems with on-the-fly link traversal and federated querying, it may transpire that widespread crawling and caching will become the norm in making data available to applications in a timely fashion, while being able to take advantage of the openness of the Web of Data by discovering new data sources through link traversal.

Schema Mapping and Data Fusion

Once data has been retrieved from distributed sources, it must be integrated in a meaningful way before it is displayed to the user or is further processed. Today, most Linked Data applications display data from different sources alongside each other but do little to integrate it further. To do so does require mapping of terms from different vocabularies to the applications target schema, as well as fusing data about the same entity from different sources, by resolving data conflicts.

Linked Data sources either use their own schemata or use a mixture of terms from existing, well-known vocabularies together with self-defined terms specific to the particular data source. In order to support clients in transforming data between different schemata, data sources can publish correspondences between their local terminology and the terminology of related data sources on the Web of Data. Current W3C recommendations such as RDF Schema (Brickley & Guha, 2004) and OWL (McGuinness & van Harmelen, 2004) define basic terminology like *owl:equivalentClass*, *owl:equivalentProperty*, *rdfs:subClassOf*, *rdfs:subPropertyOf* that can be used to publish basic correspondences. In many situations, these correspondences are too coarse-grained to properly transform data between schemata. Problems include for instance structural heterogeneity as well as value transformations. An open research issue is therefore the development of languages to publish more fine grained schema mappings on the Web. Ideally, such languages would support transitive mappings and provide for combining partial mappings in order to cover cases where data sources mix terminology from different vocabularies. Candidate technologies for this include the alignment languages presented in (Haslhofer, 2008) and (Euzenat & Scharffe & Zimmermann, 2007) as well as the rules interchange format (RIF)[Endnote: http://www.w3.org/2005/rules/wiki/RIF_Working_Group].

In addition to enhanced support for schema mapping, further research is needed in the area of data fusion for Linked Data applications. Data fusion is the process of integrating multiple data items representing the same real-world object into a single, consistent, and clean representation. The main challenge in data fusion is the resolution of data conflicts, i.e.

choosing a value in situations where multiple sources provide different values for the same property of an object. There is a large body of work on data fusion in the database community (Bleiholder & Naumann, 2008) and an increasing body of work on identity reconciliation in the Web community (Halpin & Thomson, 2008). Specific requirements that distinguish the Web of Data from other data fusion scenarios arise from the autonomy of data sources and the scarceness and uncertainty of quality-related meta-information that is required to assess data quality in order to resolve inconsistencies. Prototypical systems for fusing Linked Data from multiple sources include DERI Pipes (Le Phuoc et al., 2009) and the KnoFuss architecture (Nikolov et al., 2008).

Link Maintenance

The content of Linked Data sources changes: data about new entities is added, outdated data is changed or removed. Today, RDF links between data sources are updated only sporadically which leads to dead links pointing at URIs that are no longer maintained and to potential links not being set as new data is published. Web architecture is in principle tolerant to dead links, but having too many of them leads to a large number of unnecessary HTTP requests by client applications. A current research topic within the Linked Data community is therefore link maintenance. Proposed approaches to this problem range from recalculating links at regular intervals using frameworks such as Silk (Volz et al., 2009) or LinQL (Hassanzadeh et al., 2009), through data sources publishing update feeds (Auer et al., 2009) or informing link sources about changes via subscription models to central registries such as Ping the Semantic Web that keep track of new or changed data items.

Licensing

Applications that consume data from the Web must be able to access explicit specifications of the terms under which data can be reused and republished. Availability of appropriate frameworks for publishing such specifications is an essential requirement in encouraging data owners to participate in the Web of Data, and in providing assurances to data consumers that they are not infringing the rights of others by using data in a certain way. Initiatives such as the Creative Commons [Endnote: http://creativecommons.org/] have provided a framework for open licensing of creative works, underpinned by the notion of copyright. However, as (Miller et al., 2008) discuss, copyright law is not applicable to data, which from a legal perspective is also treated differently across jurisdictions. Therefore frameworks such as the Open Data Commons Public Domain Dedication and License [Endnote: http://www.opendatacommons.org/licenses/pddl/1.0/] should be adopted by the community to provide clarity in this area. In situations where attribution is a condition of data reuse, further research may also be required to explore how this can be achieved in user interfaces that combine data from large numbers of sources.

Trust, Quality and Relevance

A significant consideration for Linked Data applications is how to ensure the data most relevant or appropriate to the user's needs is identified and made available. For example, in scenarios where data quality and trustworthiness are paramount, how can this be determined heuristically, particularly where the data set may not have been encountered previously?

An overview of different content-, context-, and rating-based techniques that can be used to heuristically assess the relevance, quality and trustworthiness of data is given in (Bizer & Cyganiak, 2009; Heath, 2008a). Equivalents to the PageRank algorithm will likely be important in determining coarse-grained measures of the popularity or significance of a

particular data source, as a proxy for relevance or quality of the data, however such algorithms will need to be adapted to the linkage patterns that emerge on the Web of Data.

From an interface perspective, the question of how to represent the provenance and trustworthiness of data drawn from many sources into an integrated view is a significant research challenge. (Berners-Lee, 1997) proposed that browser interfaces should be enhanced with an "Oh, yeah?" button to support the user in assessing the reliability of information encountered on the Web. Whenever a user encounteres a piece of information that they would like to verify, pressing such a button would produce an explanation of the trustworthiness of the displayed information. This goal has yet to be realised, however existing developments such as WIQA (Bizer & Cyganiak, 2009) and InferenceWeb (McGuinness & da Silva, 2003) can contribute to work in this area by providing explanations about information quality as well as inference processes that are used to derive query results.

Privacy

The ultimate goal of Linked Data is to be able to use the Web like a single global database. The realization of this vision would provide benefits in many areas but will also aggravate dangers in others. One problematic area are the opportunities to violate privacy that arise from integrating data from distinct sources. Protecting privacy in the Linked Data context is likely to require a combination of technical and legal means together with a higher awareness of the users about what data to provide in which context. Interesting research initiatives in this domain are Weitzner's work on the privacy paradox (Weitzner, 2007) and the recent work by the TAMI project on information accountability (Weitzner et al., 2008).

Conclusions

Linked Data principles and practices have been adopted by an increasing number of data providers, resulting in the creation of a global data space on the Web containing billions of RDF triples. Just as the Web has brought about a revolution in the publication and consumption of documents, Linked Data has the potential to enable a revolution in how data is accessed and utilised. The success of Web APIs has shown the power of applications that can be created by mashing up content from different Web data sources. However, mashup developers face the challenge of scaling their development approach beyond fixed, predefined data silos, to encompass large numbers of data sets with heterogeneous data models and access methods. In contrast, Linked Data realizes the vision of evolving the Web into a global data commons, allowing applications to operate on top of an unbounded set of data sources, via standardised access mechanisms. If the research challenges highlighted above can be adequately addressed, we expect that Linked Data will enable a significant evolutionary step in leading the Web to its full potential.

References

Adida, B., et al. (2008). RDFa in XHTML: Syntax and Processing - W3C Recommendation. Retrieved June 14, 2009, http://www.w3.org/TR/rdfa-syntax/

Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J. (2009). Describing Linked Datasets. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).

d'Aquin, M., et al. (2008). Toward a New Generation of Semantic Web Applications. IEEE

Intelligent Systems, 23(3):20-28.

Auer, S., et al. (2009). Triplify – Light-Weight Linked Data Publication from Relational Databases. Proceedings of the 18th World Wide Web Conference (WWW2009).

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, C., Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. Proceedings of the 6th International Semantic Web Conference (ISWC2007).

Becker, C., Bizer, C. (2008). DBpedia Mobile - A Location-Aware Semantic Web Client. Proceedings of the Semantic Web Challenge at ISWC 2008.

Beckett, D. (2004). RDF/XML Syntax Specification (Revised) - W3C Recommendation. Retrieved June 14, 2009, http://www.w3.org/TR/rdf-syntax-grammar/

Beckett, D., Berners-Lee, T. (2008). Turtle - Terse RDF Triple Language - W3C Team Submission. Retrieved July 23, 2009, http://www.w3.org/TeamSubmission/turtle/

Belleau, F., Nolin, M., Tourigny, N., Rigault, P., Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics, 41(5):706-16.

Berners-Lee, T. et al. (1994). The World-Wide Web. Communications of the ACM, 37(8):76-82.

Berners-Lee, T. (1997). Cleaning up the User Interface, Section - The ”Oh, yeah?”-Button. Retrieved June 14, 2009, http://www.w3.org/DesignIssues/UI.html

Berners-Lee, T. (1998). Notation3 (N3) A readable RDF syntax. Retrieved July 23, 2009, http://www.w3.org/DesignIssues/Notation3.html

Berners-Lee, T. (2000): Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor. London, Texere.

Berners-Lee, T., Hendler, J., Lassila, O. (2001) The Semantic Web. Scientific American, 284(5):34-43.

Berners-Lee, T., et al. (2005). Uniform Resource Identifier (URI): Generic Syntax. Request for Comments: 3986. Retrieved June 14, 2009, http://tools.ietf.org/html/rfc3986

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved July 23, http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., et. al. (2006), Tabulator: Exploring and Analyzing Linked Data on the Semantic Web. Procedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06).

Berners-Lee, T., et al. (2008). Tabulator Redux: Browsing and Writing Linked Data. Proceedings of the 1st Workshop on Linked Data on the Web (LDOW2008).

Berrueta, D., Phipps, J. (2008). Best Practice Recipes for Publishing RDF Vocabularies - W3C Working Group Note. Retrieved June 14, 2009, http://www.w3.org/TR/swbp-vocab-pub/

Bizer, C., Cyganiak, R. (2006). D2R Server - Publishing Relational Databases on the

Semantic Web. Poster at the 5th International Semantic Web Conference (ISWC2006).

Bizer, C., Cyganiak, R., Heath, T. (2007). How to publish Linked Data on the Web. Retrieved June 14, 2009, http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Bizer, C., Cyganiak, R., Gauß, T. (2007): The RDF Book Mashup: From Web APIs to a Web of Data. Proceedings of the 3rd Workshop on Scripting for the Semantic Web (SFSW2007).

Bizer, C., Cyganiak, R. (2009): Quality-driven Information Filtering using the WIQA Policy Framework. Journal of Web Semantics, 7(1):1-10.

Bleiholder, J., Naumann, F. (2008). Data Fusion. ACM Computing Surveys, 41(1):1-41.

Brickley, D., Guha, R. (2004). RDF Vocabulary Description Language 1.0: RDF Schema - W3C Recommendation. Retrieved June 14, 2009, http://www.w3.org/TR/rdf-schema/

Brin, S., Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30(1-7):107-117.

Carroll, J., Bizer, C., Hayes, P., Stickler, P. (2005): Named graphs. Journal of Web Semantics, 3(4):247-267.

Catasta, M., Cyganiak, R., Tummarello, G. (2009). Towards ECSSE: live Web of Data search and integration. Proceedings of the Semantic Search 2009 Workshop at WWW2009.

Cheng, G., Qu, Y. (this issue). Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. International Journal on Semantic Web and Information Systems, Special Issue on Linked Data.

Clarke, C. (2009). A Resource List Management Tool for Undergraduate Students based on Linked Open Data Principles. Proceedings of the 6th European Semantic Web Conference (ESWC2009).

Coetzee, P., Heath, T., Motta, E. (2008). SparqPlug. Proceedings of the 1st Workshop on Linked Data on the Web (LDOW2008).

Cyganiak, R., Bizer, C. (2008). Pubby - A Linked Data Frontend for SPARQL Endpoints. Retrieved June 14, 2009, http://www4.wiwiss.fu-berlin.de/pubby/

Cyganiak, R., Delbru, R., Stenzhorn, H., Tummarello, G., Decker, S. (2008): Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. Proceedings of the 5th European Semantic Web Conference (ESWC2008).

Das Sarma, A., Dong, X., Halevy, A. (2008). Bootstrapping pay-as-you-go data integration systems. Proceedings of the Conference on Management of Data (SIGMOD2008).

Ding, L., et al. (2005). Finding and Ranking Knowledge on the Semantic Web, Proceedings of the 4th International Semantic Web Conference, November 2005.

Elmagarmid, A., Ipeirotis, P., Verykios, V. (2007). Duplicate Record Detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1):1–16.

Euzenat, J., Scharffe, F., Zimmermann A. (2007). Expressive alignment language and implementation. Knowledge Web project report, KWEB/2004/D2.2.10/1.0.

Euzenat, J., Shvaiko, P. (2007). Ontology Matching. Springer, Heidelberg.

Fielding, R., et al. (1999). Hypertext Transfer Protocol -- HTTP/1.1. Request for Comments: 2616. Retrieved June 14, 2009, http://www.w3.org/Protocols/rfc2616/rfc2616.html

Franklin, M., Halevy, A., Maier, D. (2005). From databases to dataspaces: a new abstraction for information management. ACM SIGMOD Records, 34(4):27-33.

Halevy, A., Franklin, M., Maier, D. (2006). Principles of dataspace systems. Proceedings of the Symposium on Principles of database systems (PODS2006).

Halpin, H., Thomson, H. (2008). Special Issue on Identiy, Reference and the Web. International Journal on Semantic Web & Information Systems, 4(2):1-72.

Hartig, O. (2009). Provenance Information in the Web of Data. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).

Haslhofer, B., Schandl, B. (2008). The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008).

Haslhofer, B. (2008): A Web-based Mapping Technique for Establishing Metadata Interoperability. PhD thesis, Universität Wien.

Hassanzadeh, O., et al. (2009). A Declarative Framework for Semantic Link Discovery over Relational Data. Poster at 18th World Wide Web Conference (WWW2009).

Hassanzadeh, O., Consens, M. (2009). Linked Movie Data Base. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).

Hastrup, T., Cyganiak, R., Bojars, U. (2008). Browsing Linked Data with Fenfire. Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008).

Hausenblas, M., Halb, W., Raimond, Y., Heath, T. (2008). What is the Size of the Semantic Web? In Proceedings of the International Conference on Semantic Systems (I-Semantics2008).

Heath, T. (2008a). Information-seeking on the Web with Trusted Social Networks – from Theory to Systems. PhD Thesis, The Open University.

Heath, T. (2008b). How Will We Interact with the Web of Data? In IEEE Internet Computing, Vol. 12(5): 88-91.

Heath, T., Motta, E. (2008). Revyu: Linking reviews and ratings into the Web of Data. Journal of Web Semantics, 6(4):266-273.

Hogan, A., Harth, A., Umrich, J., Decker, S. (2007). Towards a scalable search and query engine for the web. Proceedings of the 16th Conference on World Wide Web (WWW2007).

Jacobs, I., Walsh, N. (2004): Architecture of the World Wide Web, Volume One - W3C Recommendation. Retrieved June 14, 2009, http://www.w3.org/TR/webarch/

Jentzsch, A., Hassanzadeh, O., Bizer, C., Andersson, B., Stephens, S. (2009). Enabling Tailored Therapeutics with Linked Data. Proceedings of the 2nd Workshop on Linked Data on

the Web (LDOW2009).

Karger, D., schraefel, m.c. (2006). Pathetic Fallacy of RDF. Proceedings of 3rd Semantic Web User Interaction Workshop (SWUI2006).

Klyne, G., Carroll, J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax - W3C Recommendation. Retrieved June 14, 2009,http://www.w3.org/TR/rdf-concepts/

Kobilarov, G., et al. (2009). Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Conections. Proceedings of the 6th European Semantic Web Conference (ESWC2009).

Le Phuoc, D., Polleres, A., Morbidoni, C., Hauswirth, M., Tummarello, G. (2009). Rapid semantic web mashup development through semantic web pipes. Proceedings of the 18th World Wide Web Conference (WWW2009).

Marshall, C., Shipman, F. (2003). Which semantic web? Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HT2003).

McGuinness, D., van Harmelen, F. (2004). OWL Web Ontology Language - W3C Recommendation. Retrieved June 14, 2009, http://www.w3.org/TR/owl-features/

McGuinness, D., da Silva, P. (2003). Infrastructure for Web Explanations. Proceedings of the 2nd International Semantic Web Conference (ISWC2003).

Miller, P., Styles, R., Heath, T. (2008). Open Data Commons, a License for Open Data. Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008).

Moreau, L., et al. (2008). The Open Provenance Model. Technical report, Electronics and Computer Science, University of Southampton.

Nikolov, A., et al. (2008): Integration of Semantically Annotated Data by the KnoFuss Architecture. Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management.

Olston, C., Chi, E. (2003). ScentTrails: Integrating Browsing and Searching on the Web. ACM Transactions on Computer-Human Interaction, 10(3):177-197.

Oren, E., et al. (2008). Sindice.com: A document-oriented lookup index for open linked data. Journal of Metadata, Semantics and Ontologies, 3(1):37-52.

Quilitz, B., Leser, U. (2008). Querying distributed RDF data sources with SPARQL. Proceedings of the 5th European Semantic Web Conference (ESWC2008).

Raimond, Y., Sutton, C., Sandler, M. (2008). Automatic Interlinking of Music Datasets on the Semantic Web. Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008).

Sauermann, L., Cyganiak, R. (2008): Cool URIs for the Semantic Web. W3C Interest Group Note. Retrieved June 14, 2009, http://www.w3.org/TR/cooluris/

Schenk, S., et al. (2008). SemaPlorer—Interactive Semantic Exploration of Data and Media based on a Federated Cloud Infrastructure. Proceedings of the Semantic Web Challenge at

ISWC 2008.

Van de Sompel, H., Lagoze, C., Nelson, M., Warner, S., Sanderson, R., Johnston, P. (2009). Adding eScience Assets to the Data Web. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).

Volz, J., Bizer, C., Gaedke, M., Kobilarov, G. (2009): Silk – A Link Discovery Framework for the Web of Data. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).

Weitzner, D. (2007): Beyond Secrecy: New Privacy Protection Strategies for Open Information Spaces. IEEE Internet Computing, 11(5):94-96.

Weitzner, D., et al. (2008). Information Accountability. Communications of the ACM, 51(6):82-87.

Winkler, W. (2006). Overview of Record Linkage and Current Research Directions. US Bureau of the Census, Technical Report.

Zhao, J., Klyne, G., Shotton, D. (2008). Provenance and Linked Data in Biological Data Webs. Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008).

Bios and Photos

Christian Bizer

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. The results of his work include the Named Graphs data model, which was adopted into the W3C SPARQL standard, the Fresnel display vocabulary implemented by several data browsers, and the D2RQ mapping language which is widely used for mapping relational databases to the Web of Data. He initialized the Linking Open Data community project and the DBpedia project.

Tom Heath

Dr. Tom Heath is a researcher in the Platform Division of Talis Information Ltd, a leading provider of Linked Data storage, management and publishing technologies, where he

coordinates internal research focusing on collective intelligence and human-computer interaction in a Linked Data and Semantic Web context. He is a leading member of the Linking Open Data community project, and creator of the Linked Data-enabled reviewing and rating site Revyu.com, winner of the 2007 Semantic Web Challenge. Tom has a PhD in Computer Science from The Open University.

Tim Berners-Lee

Sir Tim Berners-Lee invented the World Wide Web, an internet-based hypermedia initiative for global information sharing while at CERN, the European Particle Physics Laboratory. He wrote the first web client and server in 1990. His specifications of URIs, HTTP and HTML were refined as Web technology spread. Tim is professor at the Laboratory for Computer Science and Artificial Intelligence (CSAIL) at the Massachusetts Institute of Technology (MIT) and the Computer Science Department at the University of Southampton, UK. In 2001 he became a Fellow of the Royal Society.