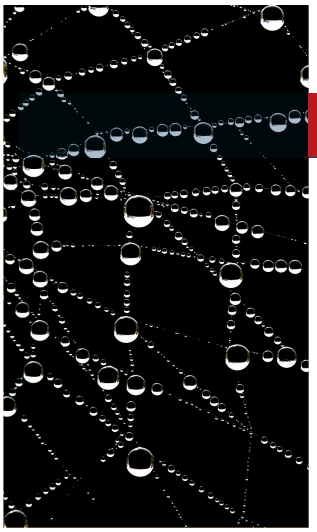

Editor: **Steffen Staab,** University of Koblenz-Landau, staab@uni-koblenz.de

The Emerging Web of Linked Data

Christian Bizer, *Freie Universität Berlin*

The classic World Wide Web is built upon the idea of setting hyperlinks between Web documents. These hyperlinks are the basis for navigating and crawling the Web; they integrate all Web documents into a single global information space.

In recent years, major Web data sources such as Google, Yahoo, eBay, and Amazon have started to provide access to their databases through Web APIs. ProgrammableWeb.com currently lists over 1,300 such APIs. The wealth of data accessible via Web APIs has led 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—Uniform Resource Identifiers (URIs), the Hypertext Transfer Protocol (HTTP), and the Hypertext Markup Language (HTML)—different Web APIs rely on different identification mechanisms and different access mechanisms, and they represent retrieved data in different formats. As most Web APIs do not assign globally unique identifiers to data items, 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, and mashup developers must choose a specific set of data sources for their application. They can't implement applications against all the data available on the Web.

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 data space.[2] Because 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 data space. Because data from different sources is connected by links, it is also possible to crawl the data space, fuse data about entities from different sources, and provide expressive query capabilities over aggregated data, similarly 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 runtime by following data-level 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 to identify Web documents, but also to identify arbitrary real-world entities.[3] Data about these entities is represented using the Resource Description Framework (RDF). Whenever a Web client resolves one of these URIs, the corresponding Web server provides an RDF/XML or RDFa description of the identified entity. These descriptions can contain links to entities described by other data sources. Links take the form of RDF triples, in which the triple's subject is a URI in the namespace of one server, and the triple's object is a URI in the namespace of the other.[4] The triple's predicate URI determines the type of the link. Whenever an application resolves a predicate URI, the corresponding

1541-1672/09/$26.00 © 2009 IEEE
Published by the IEEE Computer Society

server responds with a RDF Schema (RDFS) or Web Ontology Language (OWL) definition of the link type.[5] These descriptions can 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.
- Links connect entities, 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 to find their definitions.
- The Web of Linked Data is open, meaning that applications can discover new data sources at runtime by following links.

The Linking Open Data Effort

Over the last three years, an increasing number of data providers have begun to adopt the Linked Data principles, leading to the creation of a global data space containing billions of assertions about geographic locations, people, companies, books, scientific publications, films, music, television and radio programs, genes, proteins, drugs and clinical trials, online communities, statistical data, census results, and reviews.

The publication of Linked Data is loosely coordinated by the World Wide Web Consortium (W3C) Linking Open Data project, a grassroots community effort founded in January 2007. The project's original and ongoing goal 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 that collects community news, statistics about published data sets, and information about Linked Data publishing tools and applications (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. Anyone can participate in the project 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 links between items in two data sets. Heavier arcs correspond to a greater number of links, and bidirectional arcs indicate that each data set contains outward links to the other.

Figure 1d illustrates the size of the Linking Open Data cloud as of July 2009 and classifies the data sets by topical domain. (See the Linking Open Data wiki for further details about the data sets mentioned in this column.)

Media

A major Linked Data publisher in the media industry is the British Broadcasting Corporation (BBC). The BBC Programmes and Music sites provide data about episodes of radio and TV programs. 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 let applications 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 annotating news texts with URIs from the Linked Open Data cloud referring to places, companies, and people.

Publications

The US Library of Congress and the German National Library of Economics publish their subject heading taxonomies as Linked Data. Linked Data about scholarly publications is available from 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. The RDF Book Mashup, a wrapper around the Amazon and Google Base APIs, provides Linked Data about books. The Open Archives Initiative has based its new Object Reuse and Exchange standard (OAI-ORE) on the Linked Data principles; this standard's deployment is likely to further accelerate the availability of Linked Data related to publications.

Life Sciences

A major provider of Linked Data related to life sciences, the Bio2RDF project has interlinked more than 30 widely used life sciences data sets, including UniProt (the Universal Protein Resource), KEGG (the Kyoto Encyclopedia of Genes and Genomes), CAS (the Chemical Abstracts Service), PubMed, and the Gene Ontology. Altogether, the

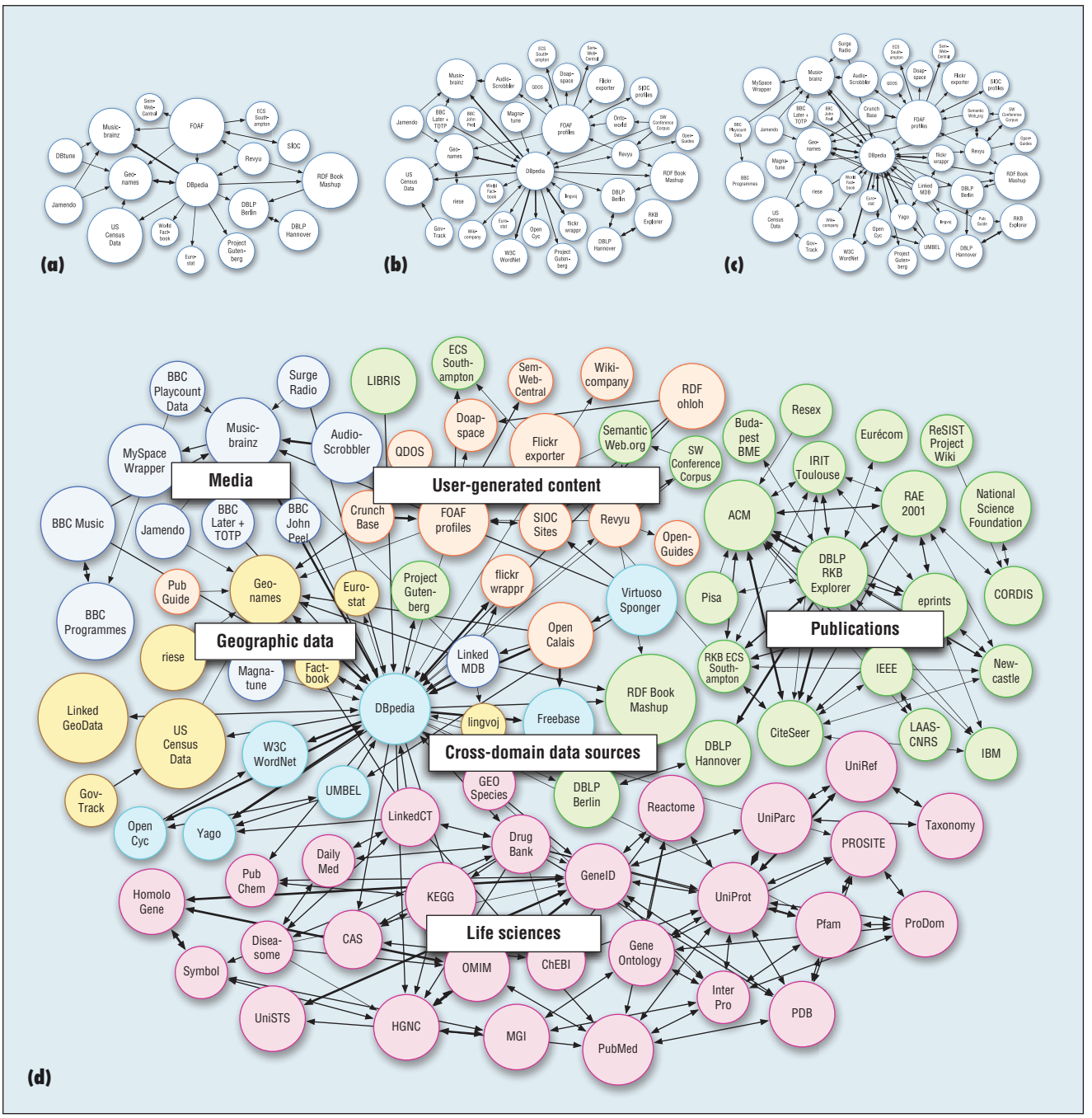

Figure 1. Growth of the Linking Open Data cloud: (a) July 2007, (b) April 2008, (c) September 2008, and (d) July 2009.

Bio2RDF data sets comprise more than two 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 to ease drug discovery.

Geographic Data

Geonames, an open-license geographical database, publishes Linked Data about eight 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. The British Ordnance Survey office has started to publish topological information about the UK's administrative areas as Linked Data. Conversions of the EuroStat, World Factbook, and US Census data sets are also available as Linked Data.

Table 1. Linking Open Data data set statistics, July 2009.

Domain	No. of triples	Percentage of the cloud	No. of links	Percentage 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-generated 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	

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 semiautomated enrichment of blog posts with data-level links pointing to DBpedia, Freebase, MusicBrainz, and Semantic CrunchBase. A further service for annotating Web content with Linked Data URIs is Faviki. These annotations connect the classic document Web with the Web of Linked Data. Applications can use the links to retrieve background information about a blog post's topics or a location depicted by a photo, and then use this information to provide a richer user experience.

Cross-Domain Data Sources

Data sources that provide information spanning multiple domains are crucial for connecting data into a single global data space and avoiding the fragmentation of the data space into distinct topical islands. An example of such a data source is DBpedia, which publishes data extracted from the "infoboxes" commonly seen on the right-hand side of Wikipedia articles. Because DBpedia covers a wide range of topics and has a high degree of conceptual overlap with many other data sets, various data publishers have started to set links from their data sources to DBpedia, making it one of the central interlinking hubs in the Linking Open Data cloud, as Figure 1 shows.

A second major source of cross-domain data is Freebase, an open-license database that users can edit in much the same way they edit Wikipedia. Further cross-domain ontologies available as Linked Data include Wordnet, OpenCyc, YAGO, and UMBEL. These ontologies are interlinked with DBpedia, so applications can mashup data from all these sources.

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

Consuming Linked Data from the Web

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

Generic data browsers let users navigate between data sources along RDF links. They discover new data sources by automatically following *owl:sameAs* links and merge data about an entity from these sources. 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 retrieved by following data-level links from Berners-Lee's Friend of a Friend (FOAF) profile into various other data sources. The colored dots beside the pieces of information indicate the data sources from which the browser merged the data.

Examples of Linked Data search engines, which crawl the Web of Linked Data by following data-level links, include FalconS, SWSE (Semantic Web Search Engine), 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 its BOSS API, and is using the data within SearchMonkey 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 products, 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 ratings Web site that augments reviews with background information from the Web of Linked Data; and Talis Aspire, a resource list management tool for university courses that 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

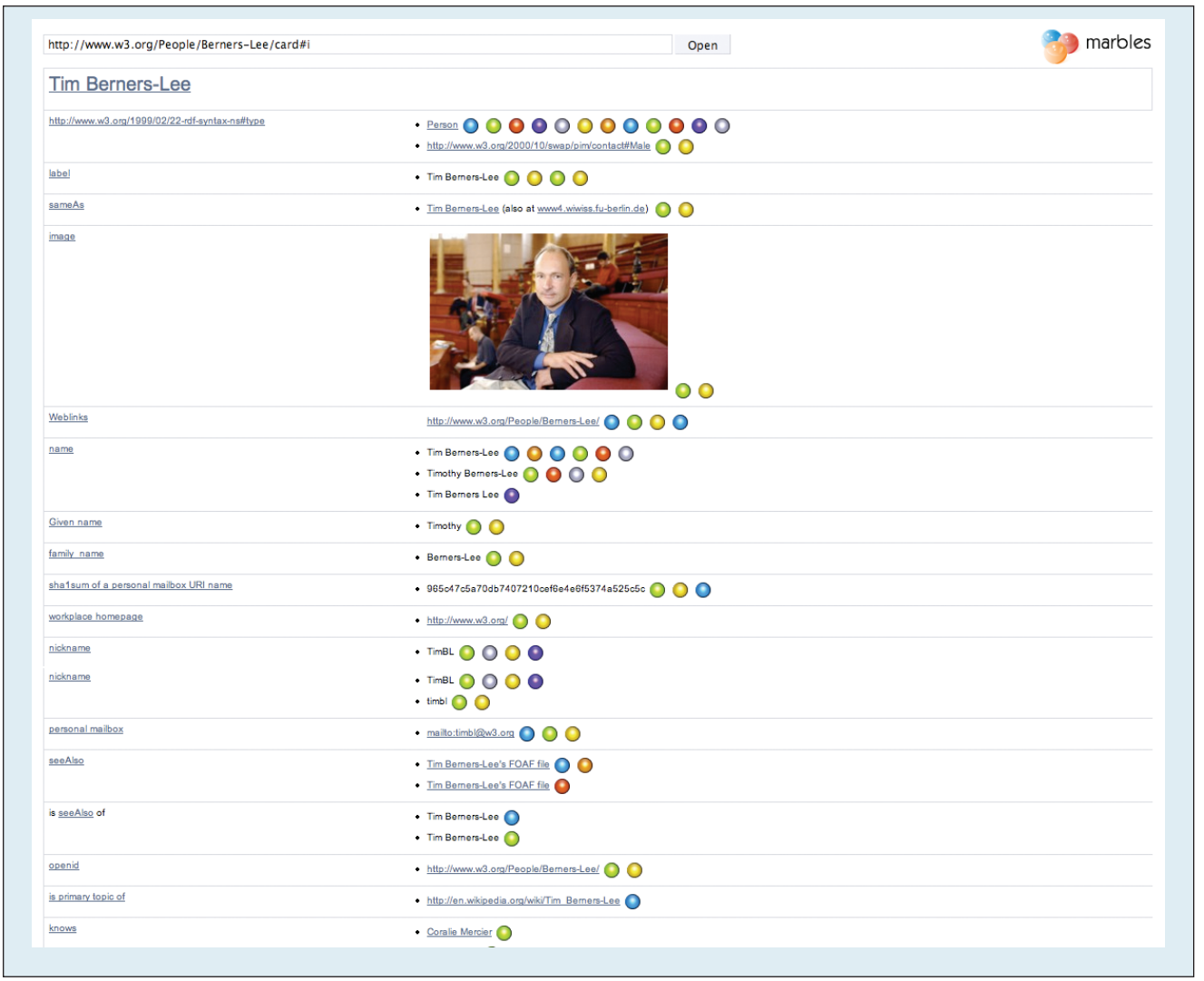

Figure 2. The Marbles Linked Data browser, displaying data about Tim Berners-Lee. Colored dots beside pieces of information indicate the various sources from which the browser merged the data.

Linked Data technologies for easing 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, and crime statistics, to your local representative's voting record. Giving the public easy access to this data enables greater accountability, helps people make informed choices, and lets third parties create tools to analyze and work with the data.

Many public-sector organizations have a 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 accessible only in two forms:

- Human-readable formats—Although HTML and PDF enable access to people, mixing data and its presentation limits the ability of machines to process data.
- Proprietary data formats—Potential consumers must have proprietary software or tools to access the data.

Linked Data has the potential to overcome these barriers. Because Linked Data is exclusively based on open Web standards, data consumers can use generic tools to access, mashup, and visualize data. In addition, Web search engines can pick up data and use it to provide better services to their users. When resolvable HTTP URIs identify data items, it is possible to set data-level links between data sources. This lets different government bodies relate their data to each other while every institution keeps full control of its original data.

Governments are beginning to understand the potential of Linked Data for public-sector applications. UK Prime Minister Gordon Brown

IEEE Intelligent Systems

How to Reach Us

Writers

For detailed information on submitting articles, write for our Editorial Guidelines (isystems@computer.org) or access www.computer.org/intelligent/author.htm.

Letters to the Editor

Send letters to

Dale Strok, Lead Editor
IEEE Intelligent Systems
10662 Los Vaqueros Circle
Los Alamitos, CA 90720
dstrok@computer.org

Please provide an email address or daytime phone number with your letter.

On the Web

Access www.computer.org/intelligent for information about *IEEE Intelligent Systems*.

Subscription Change of Address

Send change-of-address requests for magazine subscriptions to address. change@ieee.org. Be sure to specify *IEEE Intelligent Systems*.

Membership Change of Address

Send change-of-address requests for the membership directory to directory. updates@computer.org.

Missing or Damaged Copies

If you are missing an issue or you received a damaged copy, contact membership@computer.org.

Reprints of Articles

For price information or to order reprints, email isystems@computer.org or fax +1 714 821 4010.

Reprint Permission

To obtain permission to reprint an article, contact William Hagen, IEEE Copyrights and Trademarks Manager, at copyrights@ieee.org.

recently announced the appointment of Tim Berners-Lee as expert adviser on public information delivery. Berners-Lee has published a Web design note about putting government data online[6] and is working with the UK Power of Information Taskforce on realizing these ideas. In the United States, the Obama administration has started similar efforts. The recently launched Data.gov Web site currently provides access to 47 data sets generated by the Federal Government's executive branch.

To work more closely with governments and to support public institutions in using open Web standards, the W3C has formed an E-government interest group. A first result of this group's work 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.

The Web has begun to develop from a medium for publishing and linking documents into a medium for publishing and linking both documents and data. With the fragmentation of the Web into distinct data islands accessible through proprietary Web APIs, we are currently facing a situation similar to the early days of the Web, when services such as CompuServe and AOL tried to restrict users to content provided by a network of hand-selected affiliates. This walled garden approach has failed. Instead, the Web has succeeded as a single global information space that has dramatically changed the way we use information, disrupted business models, and led to profound societal change. With Linked Data, we have the technologies on hand to repeat this story for data. ■

References

1. T. Berners-Lee, "Linked Data—Design Issues," 2006; www.w3.org/DesignIssues/LinkedData.html.
2. C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data—The Story So Far," *Int'l. J. Semantic Web & Information Systems*, to appear, 2009.
3. L. Sauermann and R. Cyganiak, "Cool URIs for the Semantic Web," W3C Interest Group Note, 2008; www.w3.org/TR/cooluris.
4. C. Bizer, R. Cyganiak, and T. Heath, "How to Publish Linked Data on the Web," 2007; http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial.
5. D. Berrueta and J. Phipps, "Best Practice Recipes for Publishing RDF Vocabularies," W3C Working Group Note, 2008; http://www.w3.org/TR/swbp-vocab-pub.
6. T. Berners-Lee, "Putting Government Data Online—Design Issues," 2009; http://www.w3.org/DesignIssues/GovData.html.
7. S. Acar, J. Alonso, and K. Novak, "Improving Access to Government through Better Use of the Web," W3C Interest Group Note, 2009; http://www.w3.org/TR/egov-improving.

Christian Bizer is 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. Contact him at chris@bizer.de.