

Linked Open Government Data

Li Ding, *Qualcomm*

Vassilios Peristeras, *European Commission*

Michael Hausenblas, *National University of Ireland, Galway*

This, 2009, is the year for putting government data online. Both US and UK governments made public commitments toward open data.

—Tim Berners-Lee

Public-sector bodies produce and collect government data that records authoritative information about government activities (such as spending and service provision) and regional statistics (such as economic indicators). The emerging open government data (OGD) movement demands proactive release of government data on the Web, free of charge and with minimal constraints on reuse. Key benefits of OGD include facilitating the reuse of government data, opening up new business opportunities, enhancing government transparency and citizen engagement, and distributing the cost of government data processing to communities.

Data.gov, the US national OGD portal (www.data.gov), was launched in May 2009. A few months later, in January 2010, the British government launched Data.gov.uk (<http://data.gov.uk>). The European Commission encourages OGD through the 2003 Public Sector Information Directive and the 2011 Open Data Package

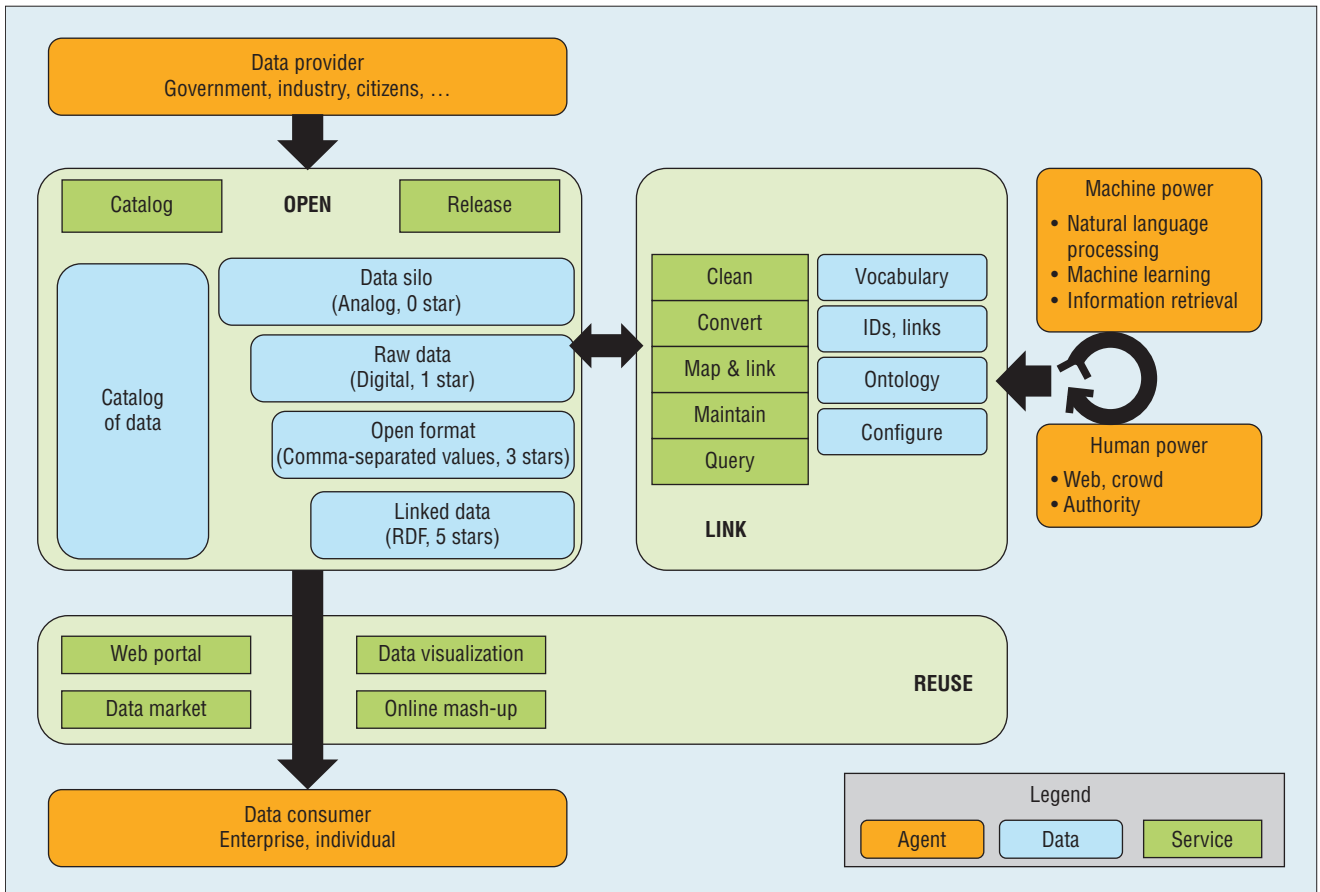


Figure 1. Roadmap of linked open government data, based on community consensus from the 2011 AAAI Fall Symposium on Open Government Knowledge. The three data-processing stages, shown in green, enhance the raw open government data from data providers using the combination of machine power and human power and deliver higher-quality data to a wide range of data consumers via visualizations, mashups, and more.

(http://ec.europa.eu/information_society/policy/psi). As of January 2012, more than 700,000 OGD datasets have been put online by national and local governments from more than 30 countries (http://logd.tw.rpi.edu/demo/international_dataset_catalog_search). One of the major challenges for OGD is the costly integration of government data across domains and political boundaries, because OGD datasets are published in various formats, use different vocabularies, and are accompanied by metadata of varying quality.

Linked open government data (LOGD), pioneered by Data.gov and Data.gov.uk, is emerging on the linked-data Web as a way of facilitating opening, linking, and reusing OGD. Linked data offers minimal

consensus on data representation (using URIs and the Resource Description Framework) and data access (via HTTP), and it enables incremental OGD publishing according to Tim Berners-Lee's "5 Stars of Linked Open Data" (<http://5stardata.info>).¹ LOGD is recognized as a Web-based open ecosystem that organically interconnects the original data owners (such as government agencies), data-processing service providers (such as entity resolution services), and data consumers (enterprises and citizens).² Figure 1 shows a roadmap of LOGD with three data-processing stages:

- In the *open* stage, government agencies play a key role in putting OGD datasets online in reusable formats and maintaining central

OGD catalogs to help citizens finding available and relevant datasets.

- In the *link* stage, community participants (industry and academia, for example) help enhance the quality of the released OGD data. Both human power and machine power can be used to generate additional declarative links (such as standard vocabulary, concept mappings, and references to relevant external data) and value-added services (such as automated entity extraction and resolution).
- In the *reuse* stage, developers pull the published OGD datasets together to build high-value applications. Right now, the value of LOGD deployments is usually exhibited by visual mashups on the Web. In the future, emerging data

markets will become the mechanisms that turn the current volunteer value-adding contributions into a profitable business sector.

LOGD represents a new data integration paradigm for sustainable growth of OGD and consequently can be considered a new enterprise integration application approach. First, it opens up the scope of data integration from traditionally closed enterprise environments such as data warehouses to the entire Web. Users can mash up government data with crowdsourced data, privately owned data, and many other types of non-governmental data. Second, it enables a data-oriented architecture (DOA) that decouples complex data objects into reusable fine-grained linked data on the Web. A service-oriented architecture (SOA) decouples the services used by applications to make them reusable by other applications and systems; a DOA, by contrast, decouples the data to make it reusable. Applying this DOA principle on the Web means that anyone can contribute to LOGD deployment with partial but interlinked contributions, such as declarative mappings from US state names to the corresponding federal information-processing standards codes or a Web service that finds relevant DBpedia (dbpedia.org) entities for a name.

This Special Issue

This special issue features reports from six countries contributed by key government practitioners and academic thought leaders from four continents:

- “Linked Open Government Data: Lessons from Data.gov.uk” walks through the experience of deploying this public data catalog to

illustrate important research challenges in integrating OGD into the linked data Web, and discusses lessons for governments, technical communities, and citizens.

- “US Government Linked Open Data: Semantic.data.gov” is the first official report from the world’s largest open government project—Data.gov, operated by the US government. It describes the background of Data.gov as well as the current and planned use of linked data for organizing knowledge and vocabularies within an OGD portal.
- “Harmonization and Interoperability of EU Environmental Information and Services” reports on the ongoing Infrastructure for Spatial Information in the European Community project, whose goal is a highly interoperable cross-border e-environment framework for the European Union. It unveils the designs and recommendations for enabling semantic interoperability via ontologies, thesauri, and spatio-temporal reasoning.
- “Making Research Data Available in Australia” reviews the architecture and experience involved in building the Australian National Data Service, revealing lessons learned from linking government-funded research data.
- “Open Government Data in Brazil” reports on the newborn Brazilian OGD portal and discusses the need for commonly agreed RDF vocabularies in enabling data links and mashups.
- “Recordkeeping and Linking Government Data in Canada” identifies challenges to LOGD based on the experiences in recordkeeping within the government of Canada. It emphasizes the importance of provenance and shows the requirements for sound recordkeeping.

- “Parallel Identities for Managing Open Government Data” presents a solution for provenance tracking in LOGD using a well-established conceptual model: Functional Requirements for Bibliographic Records, from the information science community.

Instead of covering every aspect of LOGD, we selected these articles to highlight the key challenges and lessons learned from its real-world deployment.

More Data

Although LOGD is perhaps the fastest-evolving part of the linked-data Web, most authors acknowledge a considerable entry barrier to producing LOGD. Open-source software tools have been developed and reused in facilitating cataloging and generating LOGD datasets. In particular, data portals such as the Comprehensive Knowledge Archive Network (<http://ckan.org>) in the UK and the US-India Open Government Platform collaboration (www.data.gov/opengovplatform) help in releasing more OGD datasets. In Brazil, triplification tools such as Triplify (<http://triplify.org>) are helping generate LOGD from raw OGD datasets. In this way, newcomers can easily start their work by reusing contributions from the pioneers.

More Links

The need for linking data is well-understood, as in the effort to link public and research data in Australia. Solutions from the UK, Brazil, and the EU’s Infrastructure for Spatial Information in the European Community project exemplify three different link generation approaches, respectively:

- *backlinking*, which uses plain `rdfs:seeAlso` statements to place

initial links across OGD datasets and leverage social knowledge by utilizing links provided by the community to semantically enrich the relationships and links among the datasets and concepts;

- *data normalization*, which relies on automatically learning concept and entity mappings; and
- *standardization*, which standardizes common metadata and thesauri.

Declarative entity-level links in LOGD should go beyond links to DBpedia, and there are interesting efforts to link OGD datasets by geo-spatial features in the UK and to link OGD datasets to social Web data.³

Reusable Identity and Provenance

In order to track entities, researchers in the LOGD community have investigated meaningful URI naming schemas to facilitate the reuse of entity URIs and address the frequent changes in organograms and the derived lines of order. The report from Canada identifies requirements for preserving and analyzing the provenance metadata of LOGD. "Parallel Identities for Managing Open Government Data" proposes partial solutions to these requirements that leverage library science theory and the consensus from the World Wide Web Consortium's (W3C) provenance standardization efforts.

Accessible Data

Although central dataset catalogs provide entry points for users, it's also important to reach consensus on dataset metadata, according to the lessons learned from Data.gov.uk. At the moment, vocabulary standardization is driven by the W3C Linked Government Data Working Group's project on extending the Digital Enterprise Research Institute's Data Catalog Vocabulary to a standard

dataset catalog vocabulary,⁴ and by the EU Interoperability Solutions for European Public Administration program's Asset Description Metadata Schema for describing semantic assets such as vocabularies, metadata, taxonomies, and code lists. These endeavors also need user-friendly visual interfaces, such as the UK's Geometric Rich Data Interface browser, which effectively enhances the user experience in accessing LOGD datasets.

Collaborative Networks

Several country reports discuss collaboration, communities, links with society, and the international perspective of LOGD. International collaborations between policymakers are growing; for example, the Open Government Partnership initiative currently has 35 participating countries and more than 20 countries ready to join. The UK, US, and Brazilian experiences with close collaboration between governments and the research and academic communities can serve as models for other countries that want to take the step from open to linked government data. The US also offers an interesting example in which the dataset catalog becomes the central point of reference for communities of interest formed to research, study, and exploit the available data. This approach shows a natural evolution of Data.gov from a simple repository of datasets to a live ecosystem of stakeholders coming together to discuss and share requirements and solutions based on real data.

On the basis of the current status of LOGD, we envision three grand challenges closely associated with the three stages (open, link, and reuse) of LOGD processing:

- *A million cataloged datasets.* In the past few years, an increasing

number of government offices, from national to local, have started dedicated websites for cataloging OGD. Finding government data, however, is still challenging to data consumers because of the lack of federation among the catalogs. As of now, Data.gov has already made available around 0.4 million datasets. So, when will there be one million datasets cataloged internationally? Will there be enough universal metadata for finding relevant datasets across the catalogs?

- *A million linked datasets.* Concept mappings are critical to data mash-ups, both in mapping vocabularies ("the term *state_name* is equivalent to *st_name*") and in aligning entity references ("use *dbpedia:Georgia_(U.S._state)* instead of the string 'GA'"). Although Data.gov and Data.gov.uk have published many datasets in RDF, there are only a few vocabulary alignments (via `rdfs:subPropertyOf`, for example) and entity alignments (via `owl:sameAs` statements). When will there be one million linked datasets, each of which is linked to at least one other dataset? Will the simple dataset catalog eventually evolve into a more collaborative LOGD data market in which users can share data as well as data-processing capability?
- *A million LOGD applications.* Once it's available as linked data, we certainly want LOGD being delivered to citizens to realize its value. Right now, we see demos, mashups, applications, and portals leveraging LOGD as part of their data sources. When will there be one million applications online that clearly attribute their direct or indirect use of LOGD? Would provenance metadata be more prevalent at that time?

THE AUTHORS

While it may still take considerable time and community effort to fulfill the three challenges, the open and incremental nature of the LOGD ecosystem has already stimulated a positive feedback loop. According to the country reports, an increasing number of political regions are opening up their government data, more and more techniques have been learned and re-used to link and mash up data, and applications has been built by nongovernment entities to integrate government data and deliver it to citizens. ■

References

1. T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space," *Synthesis Lectures on the Semantic Web: Theory and Technology*, J. Hendler and Y. Ding, eds., Morgan & Claypool, 2011, pp. 1–136.
2. L. Ding et al., "TWC LOGD: A Portal for Linked Open Government Data Ecosystems," *J. Web Semantics*, vol. 9, no. 3, 2011, pp. 325–333.
3. E. Kalampokis, M. Hausenblas, and K. Tarabanis, "Combining Social and Government Open Data for Participatory Decision-Making," *Proc. 3rd Int'l Conf. eParticipation* (ePart 11), LNCS 6847, Springer, 2011, pp. 36–47.
4. F. Maali, R. Cyganiak, and V. Peristeras, "Enabling Interoperability of Government Data Catalogs," *Proc. 9th Int'l. Conf. Electronic Government* (EGOV 10), LNCS 6228, Springer, 2010, pp. 339–350.

Li Ding is a staff engineer at the corporate R&D division at Qualcomm. His research interests include government data, linked data, the Semantic Web, context-aware computing, ontology, provenance, the social Web, semantic search, and knowledge management systems. Ding has a PhD in computer science from the University of Maryland. Contact him at lding@qualcomm.com.

Vassilios Peristeras is a program manager in the Interoperability Solutions for European Public Administration Unit at the European Commission. His research interests include e-government, semantic interoperability, enterprise architecture, and metadata management. Peristeras has a PhD in electronic government from the University of Macedonia. Contact him at vassilios.peristeras@ec.europa.eu.

Michael Hausenblas is a research fellow in the Digital Enterprise Research Institute at the National University of Ireland, where he leads the Linked Data Research Centre. His research interests include linked data, Representational State Transfer, and cloud computing. Hausenblas has a PhD in telematics from Graz University of Technology. Contact him at michael.hausenblas@deri.org.

IEEE computer society

PURPOSE: The IEEE Computer Society is the world's largest association of computing professionals and is the leading provider of technical information in the field.

MEMBERSHIP: Members receive the monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.

COMPUTER SOCIETY WEBSITE: www.computer.org

Next Board Meeting: 5–6 Nov., New Brunswick, NJ, USA

EXECUTIVE COMMITTEE

President: John W. Walz*

President-Elect: David Alan Grier;* **Past President:** Sorel Reisman;* **VP,**

Standards Activities: Charlene (Chuck) Walrad;† **Secretary:** Andre Ivanov

(2nd VP);* **VP, Educational Activities:** Elizabeth L. Burd;* **VP, Member &**

Geographic Activities: Sattupathuv Sankaran;† **VP, Publications:** Tom M.

Conte (1st VP);* **VP, Professional Activities:** Paul K. Joannou;* **VP, Technical**

& Conference Activities: Paul R. Croll;† **Treasurer:** James W. Moore, CSDP;* **2011–2012 IEEE Division VIII Director:** Susan K. (Kathy) Land, CSDP;† **2012–**

2013 IEEE Division V Director: James W. Moore, CSDP;† **2012 IEEE Division**

Director VIII Director-Elect: Roger U. Fujii†

*voting member of the Board of Governors †nonvoting member of the Board of Governors

BOARD OF GOVERNORS

Term Expiring 2012: Elizabeth L. Burd, Thomas M. Conte, Frank E.

Ferrante, Jean-Luc Gaudiot, Paul K. Joannou, Luis Kun, James W. Moore,

William (Bill) Pitts

Term Expiring 2013: Pierre Bourque, Dennis J. Frailey, Atsuhiko Goto,

André Ivanov, Dejan S. Milojicic, Paolo Montuschi, Jane Chu Prey, Charlene

(Chuck) Walrad

EXECUTIVE STAFF

Executive Director: Angela R. Burgess; **Associate Executive Director,**
Director, Governance: Anne Marie Kelly; **Director, Finance & Accounting:**
John Miller; **Director, Information Technology & Services:** Ray Kahn;
Director, Membership Development: Violet S. Doan; **Director, Products**
& Services: Evan Butterfield; **Director, Sales & Marketing:** Chris Jensen

COMPUTER SOCIETY OFFICES

Washington, D.C.: 2001 L St., Ste. 700, Washington, D.C. 20036-4928

Phone: +1 202 371 0101 • **Fax:** +1 202 728 9614

Email: hq.ofc@computer.org

Los Alamitos: 10662 Los Vaqueros Circle, Los Alamitos, CA 90720-1314

Phone: +1 714 821 8380 • **Email:** help@computer.org

MEMBERSHIP & PUBLICATION ORDERS

Phone: +1 800 272 6657 • **Fax:** +1 714 821 4641 • **Email:** help@computer.org

Asia/Pacific: Watanabe Building, 1-4-2 Minami-Aoyama, Minato-ku, Tokyo 107-0062, Japan

Phone: +81 3 3408 3118 • **Fax:** +81 3 3408 3553

Email: tokyo.ofc@computer.org

IEEE OFFICERS

President: Gordon W. Day; **President-Elect:** Peter W. Staecker; **Past**

President: Moshe Kam; **Secretary:** Celia L. Desmond; **Treasurer:** Harold

L. Flescher; **President, Standards Association Board of Governors:** Steven

M. Mills; **VP, Educational Activities:** Michael R. Lightner; **VP, Membership**

& Geographic Activities: Howard E. Michel; **VP, Publication Services &**

Products: David A. Hodges; **VP, Technical Activities:** Frederick C. Mintzer;

IEEE Division V Director: James W. Moore, CSDP; **IEEE Division VIII**

Director: Susan K. (Kathy) Land, CSDP; **IEEE Division VIII Director-Elect:**

Roger U. Fujii; **President, IEEE-USA:** James M. Howard

revised 22 May 2012

