INTERLINKING SCIENTIFIC DATA ON A GLOBAL SCALE

Christian Bizer

Research Group Data and Web Science, School of Business Informatics and Mathematics, University of Mannheim, B6, 26, D-68131 Mannheim, Germany
Email: chris@bizer.de

1 INTRODUCTION

A recurrent problem with the way research data is stored, processed, and accessed is that scientific work environments, as well as research data infrastructures, remain isolated solutions that focus on data from specific disciplines or data produced within specific geographic regions, such as the European Union or the United States. Moreover, scientific work environments and research data infrastructures are based on a wide range of different technical architectures; this hampers the exchange of data among systems. In most cases, the research data infrastructures that are currently in operation remain isolated data silos that supply scientists with the data the systems were designed to deliver but do not allow scientists to discover all data that is available world-wide on a topic or to discover data on related topics provided by a different research data infrastructure.

To overcome these limitations and to allow scientists to discover all data that is relevant for their task, the High-Level Expert Group on Scientific Data (HLEG, 2010), which was charged by the European Commission's Directorate-General for Information Society and Media to prepare a "vision 2030" for the evolution of e-infrastructure for scientific data, recommends in its final report the creation of a "global framework in which the data itself becomes the infrastructure–a valuable asset, on which science, technology, the economy, and society can advance".

Interestingly, this data infrastructure is currently already being realized by a global grassroots movement of data providers who are publishing their data on the Web according to the Linked Data principles (Berners-Lee, 2006). All data that is published according to these principles become part of a single global data space. This global data space - the Web of Linked Data - is based on the same architectural principles as the classic document Web (Jacobs & Walsh, 2004). These principles have proven to scale as the success of the document Web has shown over the last 20 years.

Thus this article argues that the Linked Data principles are also likely to fit the requirements of sharing scientific data on global scale.The article discusses the potential use of Linked Data for building a global research data infrastructure along with a set of scientific work environments that provide for discovering and assessing research data and publications on a global scale. The article gives an overview of the scientific- and publication-related data that is already available on the Web of Linked Data. It refers to Linked Data browsers and search engines that can be seen as early prototypes for the future integration of Linked Data features into scientific work environments. Finally, the article outlines a vision on how the Web of Linked Data is facilitated for global research data sharing in 2020, discusses current challenges, and gives recommendations for research as well as the further deployment of Linked Data technologies for sharing research data on a global scale.

1.1 The Linked Data Principles

The term Linked Data refers to a set of best practices for publishing structured data on the Web (Bizer, Heath, & Berners-Lee, 2009; Heath & Bizer, 2011). Tim Berners-Lee introduced these best practices in his Web architecture note on Linked Data (Berners-Lee, 2006), and they have become known as the Linked Data principles. These principles are the following:

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 recommended standards (RDF, SPARQL).
4. Include links to other URIs so that they can discover more things.

The first Linked Data principle advocates using URI references to identify not just Web documents and digital content but also real world objects and abstract concepts. These may include tangible things, such as people,

places, and cars, or those that are more abstract, such as the relationship type of knowing somebody, the set of all temperature values that have been measured at a specific weather station, or just one of these temperature values together with the metadata on when and where it was measured. This principle can be seen as extending the scope of the Web from online resources to encompass any object or concept in the world.

The HTTP protocol is the Web's universal access mechanism. In the classic Web, HTTP URIs are used to combine globally unique identification with a simple, well-understood retrieval mechanism. Thus, the second Linked Data principle advocates the use of HTTP URIs to identify objects and abstract concepts, enabling these URIs to be dereferenced (i.e., looked up) over the HTTP protocol into a description of the identified object or concept. Concrete guidance on how to implement the principle of dereferenceable URIs for the domain of scholarly communication is given by the open archives - object reuse and exchange (OAI-ORE) standard (OAI-ORE, 2008). The paper "Adding e-science assets to the data web" (Van De Sompel et al., 2009) discusses how the OAI-ORE standard is used to bundle publications and other types of e-science data.

In order to enable a wide range of different applications to process Web content, it is important to agree on standardized content formats. The agreement on HTML as a dominant document format was an important factor that made the Web scale. The third Linked Data principle therefore advocates for use of a single-data model for publishing structured data on the Web: the resource description framework (RDF), a simple graph-based data model that has been designed for use in the context of the Web. For serializing RDF data, RDF/XML and RDFa syntax are widely used in the Linked Data context.

The fourth Linked Data principle advocates the use of hyperlinks to connect not only Web documents but also any type of object. For example, a hyperlink may be set between a person and a place or between a place and a company. In contrast to the classic Web where hyperlinks are largely untyped, hyperlinks that connect things in a Linked Data context have types that describe the relationship between the things. For example, a hyperlink of the type "friend of" may be set between two people or a hyperlink of the type "based near" may be set between a person and a place. In the e-science context, these typed hyperlinks can be used, for example, to connect data about a person to her publications, to interlink a publication with supporting experimental data as well as to interlink data describing the same entity, for instance a gene or pathway, between different databases.

Hyperlinks in the Linked Data context are called RDF links in order to distinguish them from hyperlinks between classic Web documents. Across the Web, many different servers are responsible for answering requests attempting to dereference HTTP URIs in many different namespaces and (in a Linked Data context) returning RDF descriptions of the resources identified by these URIs. Therefore, in a Linked Data context, if an RDF link connects URIs in different namespaces, it ultimately connects resources in different data sets.

Just as hyperlinks in the classic Web connect documents into a single global information space, Linked Data use hyperlinks to connect disparate data into a single global data space. These links, in turn, enable applications to navigate the data space. For example, a Linked Data application that has looked up a URI and retrieved RDF data describing a scientific experiment may follow links from that data to data on different Web servers that describe related experiments.

The Linked Data principles enable the implementation of generic applications that operate over the complete data space because the resulting Web of Linked Data is based on standards for the identification of entities, retrieval of entity descriptions, and parsing of descriptions in RDF as a common data model. Examples of such applications include Linked Data browsers, such as Tabulator (Berners-Lee, 2006) or Marbles (Becker & Bizer, 2009), which enable the user to view data from one data source and then follow RDF links within the data to other data sources. Other examples are Linked Data search engines, such as Sig.ma (Tummarello, Cyganiak, Catasta, Danielczyk, Delbru, & Decker, 2010), Falcons (Cheng & Qu, 2009), and VisiNav (Harth, 2010), that crawl the Web of Linked Data and provide data discovery as well as advanced query capabilities on top of the data space.

1.2 Topology of the Web of Linked Data

Linked Data is not just a vision but the Linked Data principles are already applied in various application domains including e-science, libraries, and scholarly communication. The deployment of Linked Data on the Web was initiated by the W3C linking open data (LOD) project, a grassroots community effort founded in January 2007. The founding aim of the project was to identify existing data sets that are available under open

licenses, convert them to RDF according to the Linked Data principles, and publish them on the Web. Figure 1 illustrates the September 2011 scale of the Linked Data cloud originating from the W3C LOD project and classifies the data sets by topical domain, highlighting the diversity of data sets present in the cloud. Each node in the diagram represents a distinct data set published as Linked Data. The arcs indicate that RDF links exist between items in the two connected data sets.

The LOD community maintains a catalogue of known Linked Data sources, the LOD cloud data catalogue (http://datahub.io/group/lodcloud). Altogether, the catalogued data sources serve over 31 billion RDF triples to the Web. A total of 503 million of these triples are RDF links that connect entity descriptions from different data sources. The State of the LOD Cloud document (http://lod-cloud.net/state/) provides further summary statistics on a regular basis about the data sets that are catalogued within the LOD cloud data catalogue.

The Web of Linked Data contains a lot of e-government data as the UK and US governments (http://data.gov.uk/linked-data and http://www.data.gov/semantic) are currently implementing Linked Data as the preferred way to provide access to data produced by public bodies. Other topical domains include geographic data, media data, life-science data, library and scholarly communication data, cross-domain datasets, as well as user-generated content.

The Linked Science workshop series (http://linkedscience.org/events/lisc2013/) showcases the latest developments around using Linked Data technologies for sharing scientific data in general. In addition, several discipline-specific collaboration forums and scientific events have been established. In the following, I will provide references to such efforts in two areas: Life science and scholarly communication.

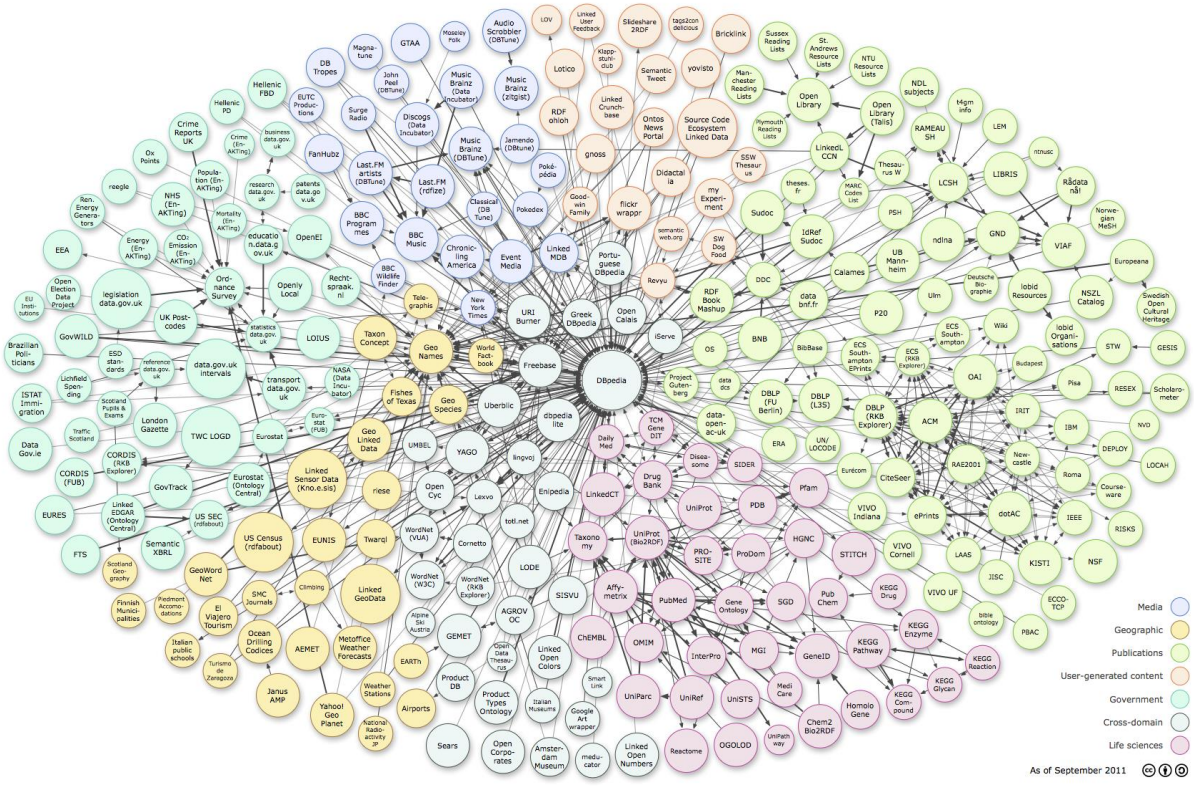

Figure 1: Linked Data cloud, the colours highlight the different topical domains

1.3 Linked Data in Life Science

The usage of Linked Data technologies to share life-science data on a global scale is a good example of how the technologies could also be applied in other scientific disciplines. Linked Data have gained significant uptake in

the life sciences as a technology to connect the various data sets that are used by researchers in this field. In particular, the Bio2RDF project (Belleau, Nolin, Tourigny, Rigault, & Morissette, 2008; Callahan et al., 2013) has interlinked more than 30 widely used 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. The W3C linking open drug data (http://www.w3.org/wiki/HCLSIG/LODD) effort has brought together the pharmaceutical companies Eli Lilly, AstraZeneca, and Johnson & Johnson in a cooperative effort to interlink openly licensed data about drugs and clinical trials, in order to aid drug discovery (Jetzsch, Hassanzadeh, Bizer, Andersson, & Stephens, 2009).

1.4 Libraries and Scholarly Communication

Libraries and scholarly communication are other fields that have seen significant early adoption of Linked Data technologies. The projects in this area aim at integrating library catalogues on a global scale; interlinking the content of multiple library catalogues, for instance, by topic, location, or historical period; interlinking library catalogues with third-party information (picture and video archives or knowledge bases like DBpedia); and at making library data more easily accessible by relying on Web standards. Examples of libraries that have adopted the Linked Data principles include the American Library of Congress and the German National Library, which publish their subject heading taxonomies as Linked Data, while the complete content of LIBRIS and the Swedish National Union Catalogue is available as Linked Data (http://blog.libris.kb.se/semweb/?p=7). Similarly, the OpenLibrary, a collaborative effort to create "one Web page for every book ever published", publishes its catalogue in RDF. Scholarly articles from journals and conferences are also well represented in the Web of Linked Data through community publishing efforts, such as DBLP (http://dblp.l3s.de/), RKBexplorer (http://dblp.rkbexplorer.com/), and the Semantic Web Dogfood Server (http://data.semanticweb.org/). The Linked Data principles together with the OAI-ORE, Dublin Core, SKOS, and FOAF vocabularies form the foundation of the new Europeana data model (http://pro.europeana.eu/edm-documentation). The adoption of this model by libraries, museums, and cultural institutions that participate in Europeana will further accelerate the availability of Linked Data related to publications and cultural heritage artifacts. The necessary technical infrastructure for using Europeana Data Model for publishing library catalogs on the Web of Linked Data is currently being developed within the DM2E project (http://dm2e.eu/). In order to provide a forum and to coordinate the efforts to increase the global interoperability of library data, W3C has started a Library Linked Data Incubator Group (http://www.w3.org/2005/Incubator/lld/). In addition, the Semantic Web in Libraries conference series (http://swib.org/swib13/) showcases the latest developments around employing Linked Data technologies in the library domain.

2 TEN YEAR VISION

The increasing global adoption of Linked Data technologies for sharing scientific data, library data, and e-government data as well as the first generation of Linked Data discovery tools, such as Linked Data search engines, indicate the suitability of the Linked Data architecture for extending the Web with a global scientific data space. Thus, my ten-year vision is as follows.

- Linked Data will develop into the standard technology of sharing scientific data on a global scale and for interconnecting data between different scientific data sources.
- The emerging Web of Linked Data will contain scientific data as well as data from other domains and might become as omnipresent in our daily lives as the classic document Web is today.
- Most open-license scientific data sets will be directly available as Linked Data on the Web. For extremely large data sets from astronomy or physics for which it is inefficient to generate an RDF representation, the Web of Linked Data will contain detailed metadata that will enable the discovery of these data sets.
- Scientific work environments will have Linked Data import and export features and will provide for publishing scientific data directly to the Web of Linked Data. Disciplinary repositories of scientific data as well as data archives will provide Linked Data views on the archived data and will thus make their content available on the Web.
- Scientists will navigate along RDF links between different scientific data sets as well as between publications and supporting experimental data. They will use Linked Data search engines to discover all data on the global scale that is relevant to their question at hand.

3 CURRENT CHALLENGES

3.1 Data Interoperability

Complying with the Linked Data principles solves the problems of syntax and access heterogeneity by standardization. The problem of semantic heterogeneity remains but is eased in the Linked Data context as follows.

- Many Linked Data sources reuse terms from widely-used vocabularies (ontologies) to represent data about common types of entities, such as people, products, reviews, publications, and other creative works. In addition, they use their own, proprietary terms for representing aspects that are not covered by the widely used vocabularies. This partial agreement on terms makes it easier for applications to understand data from different data sources and is a valuable starting point for mining additional correspondences (Heath & Bizer, 2011).
- Many Linked Data sources set identity links (owl:sameAs) pointing at data about the same entity within other data sources. In addition data sources as well as vocabulary maintainers publish vocabulary links that represent correspondences between terms from different vocabularies (owl:equivalentClass, owl:equivalentProperty, rdfs:subClassOf, rdfs:subPropertyOf). Applications can treat these links as integration hints that help them to translate data into their target schema as well as to fuse data from different sources describing the same entity (Heath & Bizer, 2011).

3.2 Data Quality

The Web is an open medium and everybody can publish data on the Web. As with the classic document Web, the Web of Linked Data contains data that is outdated, conflicting, or intentionally wrong (SPAM). Thus, one of the main challenges that Linked Data applications need to handle is to assess the quality of Web data and determine the subset of the available data that should be treated as trustworthy for the task on hand.

3.3 Scientific Work Environments

Up until now, a wide range of generic Linked Data tools, such as Linked Data browsers and Linked Data search engines, have been developed (Heath & Bizer, 2011). What is still mostly missing is the closer integration of Linked Data features into the scientific work environments that are used within the different scientific disciplines. Such features would include import and export (publishing scientific data directly to the Web of Linked Data) as well as data-discovery features. For data discovery it would also make sense to invest in the development of discipline-specific Linked Data search engines that would use focused-crawling approaches to gather all data from the Web that are relevant for a specific discipline and provide the specific search features on top of this data that are relevant for the discipline. An example of a scientific work environment that already provides it content as Linked Data and thus makes it interlinkable with other content is myExperiment, a platform for sharing scientific workflows (http://rdf.myexperiment.org/).

4 RESEARCH DIRECTIONS PROPOSED

4.1 Research on Pay-as-You-Go-Data Integration

Franklin, Halevy, and Maier (2005) have recognized that in large-scale integration scenarios involving thousands of data sources, it is impossible, or at least too expensive, to model a unifying integration schema upfront. They have thus coined the term “data spaces” for information systems that provide for the coexistence of heterogeneous data and do not require an upfront investment into a unifying schema. In such systems, data integration is achieved in a pay-as-you-go manner: as long as no, or only a small number of, mappings has been added to the system, applications can only display data in an non-integrated fashion and can only answer simple queries or even only provide text search. Over time, as more effort is invested in generating mappings, applications can further integrate the data and provide better query answers. The Web of Linked Data is a realization of the data space vision on a global scale. An interesting difference to the original data-space idea is that on the Web terms correspondences are provided by many different parties in the form of vocabulary links (see last section on interoperability). In order to take better advantage of these vocabulary links, more research on pay-as-you-go data integration in the context of the Web is needed.

4.3 Research on Data Quality Assessment in the Web Context

A major challenge for any Linked Data application is to assess the quality of Web data and to decide which subset of the data space the application wants to trust. The range of data-quality assessment methods that are potentially applicable in the Web context is very wide, and the right mixture of these methods will depend on the application context (Bizer & Cyganiak, 2009). What is needed to help applications deal with this challenge is more research on policy-based data-quality assessment frameworks for the context of the Web of Linked Data.

4.4 Development of Discipline-Specific Vocabularies

Many Linked Data sources re-use terms from widely deployed vocabularies to represent data about common things, such as people, organizations, and products. What is still missing in many cases are vocabularies that capture the terms that are relevant for specific research disciplines. Agreeing on such vocabularies would allow interoperability for the base concepts that are relevant for a specific scientific discipline. There are already on-going developments to define discipline-specific RDF vocabularies for life science (http://www.w3.org/2001/sw/hcls/), statistics http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html), and the library world (http://pro.europeana.eu/edm-documentation). It would be beneficial if other scientific communities would follow these examples.

5 CONCLUSION

Structured data is made available on the Web today in various forms. Data are published as CSV data dumps, Excel spreadsheets, and in a multitude of domain-specific data formats. Structured data is embedded into HTML pages using Microformats or Microdata. Various data providers have started to allow direct access to their databases via Web APIs. So what is the rationale for adopting Linked Data instead of, or in addition to, these well-established publishing techniques? In summary, Linked Data provides a more generic, more flexible publishing paradigm that makes it easier for data consumers to discover and integrate data from large numbers of data sources. In particular, Linked Data provides the following.

1. A unifying data model. Linked Data relies on RDF as a single, unifying data model. By providing for the globally unique identification of entities and by allowing different schemata to be used in parallel to represent data, the RDF data model has been specially designed for the use case of global data sharing. In contrast, the other methods for publishing data on the Web rely on a wide variety of different data models, and the resulting heterogeneity needs to be bridged in the integration process.
2. A standardized data access mechanism. Linked Data commits itself to a specific pattern of using the HTTP protocol. This agreement allows data sources to be accessed using generic data browsers and enables the complete data space to be crawled by search engines. In contrast, Web APIs are accessed using different proprietary interfaces.
3. Hyperlink-based data discovery. By using URIs as global identifiers for entities, Linked Data allow hyperlinks to be set between entities in different data sources. These data links connect all Linked Data into a single global data space and enable Linked Data applications to discover new data sources at run-time. In contrast, Web APIs as well as data dumps in proprietary formats remain isolated data islands.
4. Self-descriptive data. Linked Data ease the integration of data from different sources by relying on shared vocabularies, making the definitions of these vocabularies retrievable via dereferencing term URIs, and by allowing terms from different vocabularies to be connected to each other by vocabulary links.

As these are crucial features for realizing the vision of the High Level Expert Group on Scientific Data for building a global data-sharing infrastructure, and as the uptake of Linked Data has already started to acceleratein many communities, I recommend to further promote the usage of Linked Data technologies as the unifying data sharing paradigm for scientific data from all disciplines as well as to open up existing e-science infrastructures so that they can participate in and take advantage of the emerging Web of Linked Data.

6 REFERENCES

Becker, C. & Bizer, C. (2009) Exploring the geospatial semantic web with dbpedia mobile. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web 7*, pp 278–286.

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), pp 706-716.

Berners-Lee, T. (2006) Linked Data - Design Issues. Retrieved from the World Wide Web May 7, 2013: http://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, T. et al. (2006) Tabulator: Exploring and analyzing Linked Data on the semantic web. *Proceedings of the 3rd International Semantic Web User Interaction Workshop.*

Bizer, C. & Cyganiak, R. (2009) Quality-driven information filtering using the wiqa policy framework. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web 7*(1), pp 1–10.

Bizer, C., Heath, T., & Berners-Lee, T. (2009) Linked Data - The Story So Far. *International Journal on Semantic Web & Information Systems 5* (3), pp 1-22.

Callahan, A., Cruz-Toledo, J., Ansell, P., & Dumontier, M. (2013): Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. *Proceedings of The Semantic Web: Semantics and Big Data, 10th International Conference*, pp. 200-212.

Cheng, G. & Qu, Y. (2009) Searching linked objects with falcons: Approach, implementation and evaluation. *International Journal on SemanticWeb and Information Systems (IJSWIS) 5*(3), pp 49-70.

Franklin, M.J., Halevy, A.Y., & Maier, D. (2005) From databases to dataspaces: A new abstraction for information management. *SIGMOD Record 34*(4), p 2733.

Harth, A. (2010) Visinav: A system for visual search and navigation on web data. *Web Semantics: Science, Services and Agents on the World Wide Web 8*(4), pp 348-354.

Heath, T. & Bizer, C. (2011) Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool Publishers, ISBN 978160845431.

High Level Expert Group (HLEG) on Scientific Data (2010) Riding the wave. How Europe can gain from the rising tide of scientific data. *Final report of the High Level Expert Group on Scientific Data*.

Jacobs, I. & Walsh, N. (2004) Architecture of the World Wide Web, Volume One. Retrieved from the World Wide Web May 7, 2013: 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 WWW2009 Workshop on Linked Data on the Web*. Retrieved from the World Wide Web May 7, 2013: http://events.linkeddata.org/ldow2009/papers/ldow2009_paper9.pdf

Open Archives Initiative Object Reuse and Exchange -OAI-ORE (2008) Retrieved from the World Wide Web May 7, 2013: http://www.openarchives.org/ore/documents/ore-production-press-release.pdf.

Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., & Decker S. (2010) Sig.ma: Live views on the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web 8*(4), pp 355-364.

Van de Sompel, H., Lagoze, C., Nelson, M., Warner, S., Sanderson, R., & Johnston P. (2009) Adding e-science assets to the data web. *Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009)*.

(Article history: Available online 1 July 2013)