Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 00 (2011) 1–6
Procedia Computer Science

International Conference on Computational Science, ICCS 2011

Linked Open Science—Communicating, Sharing and Evaluating Data, Methods and Results for Executable Papers

Tomi Kauppinen[a], Giovana Mira de Espindola[b]

[a]*Institute for Geoinformatics (ifgi), University of Muenster, Germany*
[b]*National Institute for Space Research (INPE), Brazil*

Abstract

Linked Open Science is an approach to solve challenges of an executable paper. It is a combination of four "silver bullets": 1) publication of scientific data, metadata, results, and provenance information using Linked Data principles, 2) open source and web-based environments for executing, validating and exploring research, 3) Cloud Computing for efficient and distributed computing, and 4) Creative Commons for the legal infrastructure. We will use a realistic scientific research setting related to research on deforestation of the Brazilian Amazon rainforest to provide scenarios to illustrate the application of Linked Open Science.

Keywords:
Academic publishing, Computation, Web of Data, Linked Data, Semantic Web, Open Source, R-project, Cloud Computing, Creative Commons, The Brazilian Amazon Rainforest
2010 MSC: 68U35

1. Introduction

Currently, it is difficult and time-consuming to validate results of any particular scientific effort. The problem is that either or both the implementation of methods and the data behind a scientific paper is not openly available to assist a reviewer in her task. Thus, a review is most often—if not always—a blend of unique ingredients. To validate, and also to enhance scientific results it is thus necessary to try and reproduce a research setting. *An executable paper*[1] is an idea to solve these challenges.

Linked Open Science[2]—or short Linked Science—is our approach to provide the realization for an executable paper as a combination of Linked Data[3] [1, 2], Semantic Web [3] and Web standards, open source and web-based

Email addresses: tomi.kauppinen@uni-muenster.de (Tomi Kauppinen), giovana@dpi.inpe.br (Giovana Mira de Espindola)
URL: http://ifgi.uni-muenster.de/~kauppinen (Tomi Kauppinen), http://www.dpi.inpe.br/~giovana/ (Giovana Mira de Espindola)

[1] http://www.executablepapers.com/
[2] http://linkedscience.org
[3] http://linkeddata.org

online environments, Cloud Computing[4], the legal and technical infrastructure by Creative Commons[5], and naturally also a joint effort by the scientific community and academic publishers.

The remainder of this paper is structured as follows. In Section 2 we describe the Linked Open Science approach for executable papers. In Section 3 the approach is demonstrated by using a real scientific research setting for research about deforestation in the Brazilian Amazon rainforest. We discuss the related work in Section 4 and provide concluding remarks in Section 5.

2. Linked Open Science for Sharing and Validating Research

Linked Open Science aims to be a standardized and generic recipe for executable papers. First of all, the solution is built using the following four "silver bullets":

(i) Linked Data Input data, results and provenance information are published and archived using the Linked Data principles.

(ii) Open Source and Web-based Environments Methods are written for publication in an open source environment such as R-project[6] or using web-based techniques, and exactly the same methods are outlined to the LaTeX-version of the paper by using automated tools.

(iii) Cloud Computing The execution of methods and access to various resources are provided using the Cloud Computing approach.

(iv) Creative Commons Creative Commons Licensing is in use to provide the legal and technical infrastructure for assigning noncommercial licenses and commercial licenses, and public domain statements to scientific work, including data, methods, results and publications.

These ideas are illustrated in Table 1 in relation with the important features of an executable paper. The following subsections discuss them in more detail, and show how these four silver bullets of Linked Open Science enable different features.

2.1. Using Linked Data for Distributing, Sharing and Archiving Data

Linked Open Science relies heavily on Linked Data technologies. In brief, Linked Data is about "using the web to create typed links between data from different sources" [2]. It allows sharing and use of data, ontologies and various metadata standards: in fact, a common envision is that it will be *de facto* standard for providing metadata, and the data itself on the Web.

Linked Data is based on *principles*[7] that include 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 an URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover additional information

In Linked Data all information is encoded using Resource Description Framework (RDF)[8] as triples of form `<subject,predicate,object>`. This allows to link different resources together using predicates, and also define literal values, such as names, for the resources. All resources in Linked Data are identified using URIs. This allows for requesting more information using the URIs in a machine-processable manner. In practice this means that if a software agent requests a URI, then e.g. all the triples where this URI is as a subject could be returned, or alternatively even all

[4] http://en.wikipedia.org/wiki/Cloud_computing
[5] http://creativecommons.org/
[6] http://www.r-project.org/
[7] http://www.w3.org/DesignIssues/LinkedData
[8] http://www.w3.org/RDF/

	Executability	Short and long-term compatibility	Validation	Copyright/licensing	Systems	Size	Provenance	Viruses and code contamination	Detecting plagiarism
Linked Data	✔	✔	✔	✔		✔	✔		✔
Open Source Environment	✔	✔	✔		✔			✔	
Cloud Computing	✔				✔	✔		✔	
Creative Commons				✔			✔		✔

Table 1: Linked Open Science vs. crucial issues of an executable paper.

triples where this URI is either a subject, predicate or object. In the Linked Open Science this means that everything is identified via URIs: researchers, research institutes, publications, and research datasets.

Linked Data thus allows for efficient distribution of data using the Web standards. Scientists can make links from their data to existing datasets, thus connecting scientific resources. There are several benefits of this approach. First of all, distribution of data over the web reduces the space needed to store data in a single, local environment. For example, if a scientist is using geographic information—places with their coordinates and polygonal boundaries and so on—she just links to existing datasets providing this information rather than downloading everything into her local environment.

Thus, with the Linked Data approach the *size* of the data is a smaller issue because most of the data is already on the Web, somewhere, and served automatically by URIs on demand. Linked Data also allows for *executing* and *validating* methods on real data on the (Semantic) Web. Because it is based on WWW-standards, it clearly helps to achieve *compatibility*, in both short- and long-term. In Linked Open Science *provenance* [4, 5] information is published as Linked Data, using for example Open Provenance Model Vocabulary[9]. For example, for datasets this means that it is recorded who has created the data, or has encoded or transformed it, who has published it, and also who has used it. For each piece it is also recorded when the actions were performed at. In addition, the published paper serves as a documentation for data and methods. Note that the scientific data as itself can become a publication, and can be referred and linked upon. All this allows to use methods from the fields of Semantic Computing[10] and Machine Learning[11] in order to analyze semantic and statistical similarities of Linked Data sets and other research resources and thus detect *plagiarism* and *copyright* issues.

2.2. Executing Research in Open Source Environments and on the Web

The idea of using open source environments enables for anyone to not just contribute to the code base but also to *execute* and *validate* the existing implementations of methods and techniques. In Linked Open Science this means using established open source environments—that support many operating systems for *compatibility*—such as R-project[12], or various web-based techniques to implement the methods with. Similarly as today researchers submit LaTeX-source

[9] http://purl.org/net/opmv/ns
[10] http://en.wikipedia.org/wiki/Semantic_computing
[11] http://en.wikipedia.org/wiki/Machine_learning
[12] see e.g. [6]

files of an article together with figures, they will in Linked Open Science submit also methods as R-scripts or as online implementations, and links to datasets. The article versions of the methods are outlined automatically using tools such as Sweave [7] which for R-scripts allow to "create dynamic reports, which can be updated automatically if data or analysis change".

A crucial part of executing research in Linked Open Science is the ability to query data in a distributed manner. Linked Data is served in an increasing manner using SPARQL-endpoints. This means that a SPARQL query [8] can be sent to an endpoint, and the result is RDF. When multiple endpoints are combined together the processing of queries is inherently distributed and lowers the need for local computing power. In Linked Open Science the vision is that scientists may test their hypothesis by building methods using the distributed infrastructure of the Link Data. This makes the whole process transparent.

First of all, in Linked Open Science all the data is on the web for anyone—not just for reviewers—to check and verify the quality, origin, copyrights, and licenses. Secondly, also the implementations of methods are openly accessible, including not just the direct contribution of the author of a paper, but also the more generic methods, e.g. those used for evaluation. For example, it is common in information retrieval literature to report the performance of a method using metrics such as precision and recall. However, the traditional way of reporting these results is error prone: it is not always known how metrics have been implemented for a paper. In Linked Open Science reviewers can see which implementation has been used to verify the results, and if in doubt, can suggest for an alternative one. The vision is that over time this will lead to *de facto* library of evaluation measurement metrics, thus improving the quality of information retrieval research. Similar should then happen to other branches of science.

2.3. Low-cost Computing in the Cloud

Cloud Computing (see e.g. [9]) is an idea to *execute* code on machines around the Web, without the user knowing on which machine(s) the execution actually happens—thus the term "Cloud Computing". It is low-cost because machines are in an efficient use. Cloud Computing also enables to handle datasets of large *size*, as the user does not have to handle them in their own environments. Furthermore, user does not have to maintain all different kinds of *systems* herself: it is likely and also easier to find a needed system environment in the Cloud than setting it up from scratch in a local environment. Furthermore, *viruses* cannot make that much harm when run in controlled environments offered by the Cloud.

The access to data in Linked Open Science is provided also visually by services in the Cloud. When all the predicates for describing the data are given unique URIs it enables for creating more generic visualization and browsing facilities. For example, there are already numerous online applications capable of putting data on a map, if the data uses those URIs for latitude and longitude proposed by W3C. Similarly, data in Linked Open Science can be explored on a timeline, and by using other facets like theme, origin, author and usage history.

2.4. Managing Licenses and Copyrights using the Creative Commons Infrastructure

In Linked Open Science- approach Creative Commons allow for assigning explicit *license* statements and *copyright* terms to methods, datasets, results, and to publications. It enables assigning different terms to all these scientific pieces in different phases of their *provenance*. For example, a scientific work could be initially licensed—e.g. when submitted for review—under the license `CC BY-NC-ND 3.0`[13] but after a successful publication the methods and data could be released to the public domain under the statement `CC0 1.0`[14]. Stating clearly copyright and license issues enables to reduce *plagiarism*.

Thus, in Linked Open Science all the research data, copyright and license information, and the paper itself, and actions taken on them and their provenance information will be opened and represented as Linked Data. In addition to checking copyright and license issues of certain datasets, this approach also allows for querying and filtering datasets using the references to copyright schemes and licenses. Hence, this enables also for finding out interesting datasets which fullfill the usage permission criteria of a planned research setting.

[13] http://creativecommons.org/licenses/by-nc-nd/3.0
[14] http://creativecommons.org/publicdomain/zero/1.0

3. Evaluation and Demonstration

3.1. Scenario

We illustrate the application of the Linked Open Science approach for an executable paper by two scenarios: (i) a research scenario for an analysis of the Brazilian Amazon rainforest data, and (ii) a review scenario. The scenarios are as follows.

(i) Scenario for conducting research and submission of the work

> Maria makes research about the deforestation process in the Brazilian Amazon rainforest. She has just completed a study of this area of more than five million square kilometers. The data she has used in the study is partially provided as Linked Data by existing sources such as GeoNames[15] and partially generated from other data sources by methods that Maria has developed during the research, and implemented in R. Maria decides to submit her work for evaluation to a journal. For this she signs in the journal submission system, registers a new paper, and receives a URI for it. She uploads the article as a PDF, each of the LaTeXsource files, scripts that implement the methods of her study, additional figures, URIs of datasets she has used, and selects Creative Commons licenses for each of the resources. Some of the Linked Data datasets created by others are already online. And some of them she has created during the study, they get now published in the Cloud provided by the publisher.

(ii) Scenario for review and validation of the work

> John has accepted to review a paper on analysis of deforestation process in the Brazilian Amazon rainforest. He opens up the review environment, and can see not just the final article, but also all the related datasets and methods the author (Maria) had used in the study. He can see that exactly the same methods that are reported in the paper are also in the scripts that manipulate the datasets. John executes the methods one by one, and sees how they use and manipulate the datasets in order to produce the results. Results can be depicted as figures, and can also be interacted with by the help of a visualization library provided by the environment. John quickly sees that the research is a novel one, and provides new methods for finding out a new understanding about the deforestation process. He can also look up terms he was not aware of from the Wikipedia[16], automatically linked from the paper. After several other sessions to explore and visualize the datasets and methods, and to read the documentation John decides to accept the paper with minor modifications. He suggests to include more related work to the article. This decision is published to Maria who takes actions, after which the article is accepted, published and archived together with all its datasets, methods, results, provenance, and license information.

4. Discussion and Related Work

There are numerous approaches already in the direction of sharing and communicating science with novel techniques and practices. Linked Open Data University of Muenster (LODUM)[17] is an initiative of the University of Muenster to publish as Linked Data all their data about publications, people, patents, projects, and prizes with references to places and time periods, and defined using properties. Similar attempts are made increasingly, e.g. at the University of Southampton[18] or at the Open University[19] for extracting, interlinking and exposing university content and publishing them online as Linked Data. DataCite[20] is a community-driven effort for enabling citations of datasets using Digital Object Identifiers (DOIs), and to help archiving and accessing research data. The Semantic Publishing and Referencing Ontologies (SPAR)[21] are meant for semantic publishing and referencing. One of them is the Citation

[15] http://geonames.org/
[16] http://www.wikipedia.org/
[17] http://lodum.de/
[18] http://data.southampton.ac.uk/
[19] http://data.open.ac.uk/
[20] http://datacite.org/
[21] http://purl.org/spar/

Typing Ontology (CiTO)[22] [10], meant for characterizing citations by different citation types (explicit, implicit or indirect). Semantic Web Journal[23] provides an open review process, meaning that submitted papers are online, enabling anyone to comment them. Moreover, also the reviews are published online, with the reviewer names exposed. Utopia[24] is another related approach for 1) interacting with data, 2) exploring metadata and contents of articles, 3) and for commenting articles online.

5. Conclusions

In this paper we described Linked Open Science for solving the challenges of an executable paper. This includes publishing of data, methods, resources, results, license, rights assertions and provenance information together with the documentation, and to help establishing trust related to them. Thus Linked Open Science seeks to help authors, reviewers, publishers and the whole scientific community in their challenging tasks, and be the key if not the future form of academic publishing.

Acknowledgments

This research has been partially funded by the International Research Training Group on *Semantic Integration of Geospatial Information* (DFG GRK 1498, see http://irtg-sigi.uni-muenster.de). We also acknowledge Carsten Keßler for his valuable feedback.

6. References

[1] T. Berners-Lee, Linked Data, Personal view available from http://www.w3.org/DesignIssues/LinkedData.html (2009).
[2] C. Bizer, T. Heath, T. Berners-Lee, Linked Data – The Story So Far, International Journal on Semantic Web and Information Systems 5 (3) (2009) 1–22.
[3] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American 284 (5) (2001) 34–43.
[4] O. Hartig, Provenance Information in the Web of Data, in: Proceedings of the Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW), Madrid, Spain, 2009.
[5] O. Hartig, J. Zhao, Publishing and consuming provenance metadata on the web of linked data, in: Proceedings of The third International Provenance and Annotation Workshop, Troy, NY, U.S.A, 2010.
[6] B. D. Ripley, The R project in statistical computing, MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network. 1 (1) (2001) 23–25.
[7] F. Leisch, Sweave, part I: Mixing R and LaTeX, R News 2 (3) (2002) 28–31.
[8] E. Prud'hommeaux, A. Seaborne, SPARQL Query Language for RDF, W3C Recommendation, available from http://w3.org/TR/rdf-sparql-query (January 2008).
[9] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, A view of cloud computing, Commun. ACM 53 (2010) 50–58. doi:10.1145/1721654.1721672.
[10] D. Shotton, CiTO, the Citation Typing Ontology, Journal of Biomedical Semantics 1 (Suppl 1) (2010) S6+. doi:10.1186/2041-1480-1-S1-S6.

[22] http://purl.org/spar/cito/
[23] http://www.semantic-web-journal.net/
[24] http://getutopia.com/