Contents

Chapter 1

Linked Science: Interconnecting Scientific Assets

Tomi Kauppinen

Institute for Geoinformatics, University of Muenster, Germany

Alkyoni Baglatzi

Institute for Geoinformatics, University of Muenster, Germany

Carsten Keßler

Institute for Geoinformatics, University of Muenster, Germany

1.1 Introduction

The way scientific results are published needs to be improved. First of all, scientific research settings have changed dramatically towards digital environment, [1] which calls for changes to the traditional, paper-based publishing. The idea of "nanopublications" [15], for example, suggests a transition from the one-sided text publishing of scientific results to a more structured,

machine-understandable and data oriented perspective of research and findings.

There is a need for change because the state of the art is insufficient in too many ways: methods, datasets, results and even publications are not described in a machine-understandable way [12, 8, 16]. Moreover, they are not openly accessible, and as a result scientific settings can hardly be reproduced. It is therefore 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—or anyone else trying to reproduce a scientific setting—in her task. As a result it takes too much time to get scientific results into practice [1], or to produce new knowledge on top of the existing knowledge. The communities need better and more open science, and faster in order to cope e.g. with the huge challenges related to our environment (climate change, natural disasters) and society (health, food, poverty).

Linked Science[1]—or Linked Open Science to emphasize the need for transparency—is an approach to interconnect all scientific assets. By doing this, Linked Science seeks to revolutionize the way thousands of research organizations, and millions of scientists in them work and produce new knowledge. Linked Science is defined as a combination of Linked Data[2] [2, 5], Semantic Web [3] and Web standards, open source and Web-based online environments, Cloud Computing[3], and a machine-understandable technical and legal infrastructure.

In this book chapter, we first describe the background concerning semantic modeling of scientific resources and introduce the Linked Open Data University of Muenster initiative in the next section. In Section 3 we present and define the Linked Science approach. Section 4 presents the Linked Science Core vocabulary for linking research settings and all elements related to them together. We also present a case study concerning geochange and deforestation, and provide the discussion of the related work, as well as a research agenda for the years to come. Section 5 concludes the chapter.

1.2 Background

This section points to related work in the use of semantic technologies in archiving and interlinking scientific assets. Moreover, we introduce the Linked Open Data University of Muenster project as a specific initiative to foster the future development of the research process.

[1] See http://linkedscience.org.
[2] See http://linkeddata.org.
[3] http://en.wikipedia.org/wiki/Cloud_computing

1.2.1 Semantics for Scientific Linkage

The vision of the Semantic Web introduced by Tim Berners-Lee in 2001 [3] is based on the idea of making content on the Web machine-understandable. The transition from the level of character sequences to a level of meaning is supposed to provide new opportunities for improving interoperability, search and intelligent applications.

The basic prerequisite for communication is a basic common and shared understanding. In the vision of the Semantic Web, ontologies play a major role in the chain. According to Gruber, "an ontology is an explicit specification of a conceptualization" [9]. Given this, ontologies aim to assist in formalizing the knowledge of a domain. The goal is to enable efficient sharing and integration of information. In order to lower the barrier of achieving the Semantic Web outside the scientific research community, Berners-Lee has proposed the Linked Data approach [2, 5] that starts from existing data sets (e.g. in relational databases or various flavors of XML) and provides light-weight semantic annotations.

In this way, the content in millions of data sets—freed from *data silos* [4] that do not allow easy access or reuse—become part of the Web of data, will be interlinked and described using vocabularies. In Linked Data, online resources are identified by URIs, which are in an ideal case dereferenceable, so that users and machines accessing the data can discover more knowledge by following links. Shared vocabularies will assist in linking data in a useful and meaningful way. Thus, the idea is that data published as Linked Data facilitates sharing between heterogeneous domains, such as scientific disciplines.

Vocabularies are a key part of the Linked Data principles as they provide means to overcome semantic interoperability problems [4]. An example underlying the necessity of vocabularies can be seen in the deforestation domain [7]. In the National Institute for Space Research in Brazil (INPE)[5], the term "desflorestamento" (in Portuguese) has been introduced to refer to the clearing of vegetation and transition categories, omitting clearing of cerrado (savannah). However, the state government of Mato Grosso uses the term "deforestation" (desmatamento) referring to all three categories (forest, transition categories and savannah).

As a result, the deforestation rates of the two organizations are not comparable because of the disagreement on the basic terms. By formalizing such concepts and publishing them as vocabularies, disambiguities could be identified and eventually resolved, so that knowledge can be shared more efficiently. Naturally these semantic interoperability problems expand even more between and across heterogeneous scientific domains. The Linked Data approach aims to provide new perspectives and solutions for knowledge interconnection to overcome such problems.

[4] Berners-Lee called for *raw data now* in his 2009 TED talk; see http://on.ted.com/9x4B.
[5] See http://www.inpe.br/ingles/index.php.

1.2.2 LODUM: Linked Open Data University of Muenster

The University of Muenster, Germany, is one of the first universities to commit to a institution-wide Linked Open Data program. The goal of the LODUM initiative[6] is to increase transparency and comprehensibility both for research and for administrative matters. Beyond the technical infrastructure required to put this undertaking into practice—server facilities to run a triple store, SPARQL endpoint, data dumps, backups, and synchronization jobs—fostering a change in the mindset of the university's research community is key for the success of LODUM. Open Access must be propagated as the preferred way of publishing results, along with the primary data on the LODUM (or any other) data platform.

While the long-term focus is on scientific data and publications, other data, such as class schedules and administrative data, are also already being integrated into the LODUM infrastructure. These additional data complement the LODUM strategy and allow a larger group of users to benefit from the infrastructure by including students and administrative staff. By implementing Open Access and Linked Open Data principles throughout the university, LODUM aims to make the output of the university more visible and foster collaboration, both between the university's faculties and with partners. Figure 1.1 shows an example of how LODUM interconnects different resources—organizations, people, buildings, rooms, courses, research and publications—coming from heterogeneous sources.

Making these raw data accessible to guarantee reproducibility of the results documented in the publication has to become the rule, not the exception. In some fields, this is already common practice, such as in Bioinformatics where papers can only be submitted to a journal along with an ID for the genome sequence in the focus of the paper. This ID is obtained by uploading the sequence to the Gene database hosted at the National Center for Biotechnology Information.[7]

1.2.3 Activities by Scientific and Publishing Communities

More and more universities and research related organizations are likely to increase efforts of opening their data. Projects such as the VIVO Project at Cornell University[8], LUCERO at the Open University[9], the University of Southampton Linked data [10], Linked Life Data[11], Linked Open Drug Data[12] and Bio2RDF[13] are all actively producing Linked Data related to science,

[6] See http://lodum.de and http://data.uni-muenster.de.
[7] See http://www.ncbi.nlm.nih.gov/gene.
[8] See http://vivo.cornell.edu.
[9] See http://lucero-project.info.
[10] See http://data.southampton.ac.uk.
[11] See http://linkedlifedata.com.
[12] See http://www.w3.org/wiki/HCLSIG/LODD.
[13] See http://bio2rdf.org/.

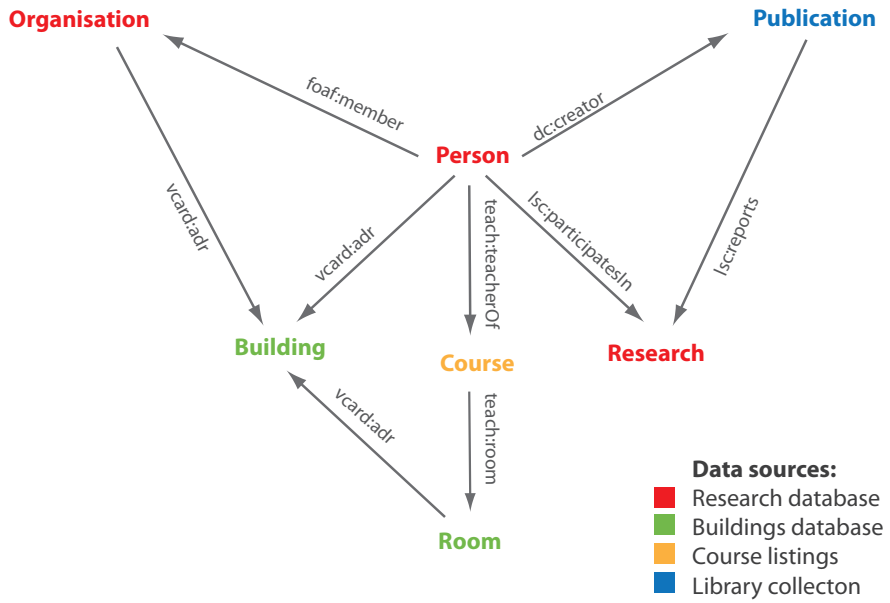

FIGURE 1.1: Example of how LODUM interconnects different resources coming from heterogeneous sources.

research and learning. To increase and enable collaboration and sharing of ideas, the Linked Universities[14] network has been founded as an alliance of European universities engaged into exposing their public data.

Publishers and other related communities are also interested in developing new ways of ensuring reproducibility of research and developing new types of citations. These efforts include for example workshops such as "Beyond the PDF" at the University of California[15], "Data citation principles" at the Harvard University[16], and "Apps for Science"[17] and "Executable Papers"[18] challenges sponsored by Elsevier.

There is a need for efforts in building a community around Linked Science in order to put this approach into practice. One of these efforts that the authors of this book chapter were involved in was the 1st International Workshop of Linked Science (LISC 2011)[19] at International Semantic Web Conference 2011 in Bonn, Germany. The goal was to "discuss and present results of new ways of publishing, sharing, linking, and analyzing such scientific resources

[14]See http://linkeduniversities.org.
[15]See http://sites.google.com/site/beyondthepdf/.
[16]See http://www.iq.harvard.edu/events/node/2462.
[17]See http://appsforscience.com/
[18]See http://www.executablepapers.com/.
[19]See http://linkedscience.org/events/lisc2011.

motivated by driving scientific requirements, as well as reasoning over the data to discover interesting new links and scientific insights". Moreover, related to this, authors also organized a breakout session in the Science Online London 2011[20]—hosted by Nature.com[21] and Digital Science at the British Library in London, UK—in order to develop vocabularies for scientific data sharing and reuse. These events, along with others to be established, need to continue at a regular pace in order to build a community around Linked Science. The two meetings mentioned above have already shown that there is a vibrant community working on various aspects of Linked Science, and that there is a lot of interest to develop this approach further and bring into research practice, curricula, publishers, and funding bodies.

1.3 Linked Science

Linked Science is an approach where scientific resources—e.g. workflows, processes, models, data, methods and evaluation metrics—are semantically annotated and interconnected [12]. In order to achieve linkage, different resources have to be described and connected in an explicit, formalized manner. In practice this needs shared conceptualizations, well-defined ontologies and vocabularies, and reasoning mechanisms to represent, link and share scientific knowledge.

The key motivation behind Linked Science is that it is crucial to efficiently communicate information about scientific findings. Opening up scientific data, methods and results improves transparency, and allows for reasoning to find out links between researchers (e.g. "John has collaborated with Mary"), supports validation of research results (e.g. "The scientific findings about typical deforestation patterns by John can be verified online by anyone") and enables new ways of communicating results as media through interactive visualizations. Ethical aspects and impacts on society and environment are highly important when communicating science. To give an example, observation of deforestation in rain forests, and reasoning about changes can be used to understand and finally reduce deforestation [6].

In Linked Science, data is published on the Web and resources are semantically annotated, methods are available as open source for running against the data, copyrights and licenses are clearly explicated, and all bits and pieces are distributed and in the Cloud. This means that Linked Science is much more than simply providing scientific data as Linked Data. Namely, as argued earlier by Bechhofer et al. [1], publishing data "has requirements of provenance,

[20] http://www.scienceonlinelondon.org/
[21] See http://www.nature.com/.

quality, credit, attribution, methods in order to provide the reproducibility that allows validation of results."

One challenge that Linked Science targets is how to achieve semantic integration and interlinkage of space and time within the scientific studies. Temporal issues in representing and sharing knowledge are crucial as scientific results often refer to a specific period of time. By linking knowledge, which has these immanent properties of space and time, new perspectives for the scientific community emerge. Links are useful in many practical settings. On the Web, links are used to browse Web pages, and to jump from one Web page to the next. Analysis of links is employed by Web search engines in order to rank Web pages given a certain query. Links between current and historical places can be used to align historical and current scientific observations together to analyze e.g. climate change. This could be done similarly as a cultural heritage portal can recommend historical content even if a user is querying by a contemporary place name [13]. The idea in Linked Science approach thus is that through interlinking all of the scientific components together with vocabularies, it is possible to form a huge collection of scientific data. This allows for interesting link and pattern discovery, e.g. links between research institutes, research trends, and so on.

In the following sections we present different aspects of Linked Science.

1.3.1 Distributing, Sharing and Archiving Data

Linked 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" [5]. It allows sharing and use of data, ontologies and various metadata standards: in fact, a common vision is that it will be the *de facto* standard for providing metadata, and the data itself on the Web.

Linked Data is based on *principles*[22] that include using HTTP URIs as names for things so that people can look up those names [and] when someone looks up an URI, providing of useful information, using the standards (RDF, SPARQL), [and] inclusion of links to other URIs, so that they can discover additional information.

In Linked Data all information is encoded using the Resource Description Framework (RDF)[23] as triples of form
<`subject,predicate,object`>. This allows linking of different resources together using predicates, and also defining of 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 triples where this URI is either a subject, predicate or

[22] http://www.w3.org/DesignIssues/LinkedData.
[23] http://www.w3.org/RDF/

object. Linked Science adopts this approach, identifying researchers, research institutes, publications, and research datasets as URIs.

Linked Data thus allows for efficient distribution of data using 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 Web standards, it clearly helps to achieve *compatibility*, in both short- and long-term. In Linked Science, *provenance* [10, 11] information is published as Linked Data, using for example the Open Provenance Model Vocabulary[24]. 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. Moreover, knowledge about the provenance of links themselves—i.e. when a link was created, published, by whom, using which version of which database, or when a link became obsolete [19]—may be used to connect and compare different scientific results.

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 to. All this allows to use methods from the fields of Semantic Computing[25] and Machine Learning[26] in order to analyze semantic and statistical similarities of Linked Data sets and other research resources and thus detect *plagiarism* and *copyright* issues.

1.3.2 Open Source for Reproducible Research

The ability to reproduce research reported in publications is one of the key requirements in science. However, practices for documenting are not fully supporting this ideal. There are numerous reasons for this: data is not easily available, reimplementing methods based on abstract descriptions is time-consuming, error-prone, and sometimes even impossible—the devil is in the details. In Linked Science, the methods and their implementations are provided as an inherent part of the publication.

For example, when a method running statistical analysis is implemented

[24] http://purl.org/net/opmv/ns
[25] See http://en.wikipedia.org/wiki/Semantic_computing.
[26] See http://en.wikipedia.org/wiki/Machine_learning.

using the R Project[27], others can simply run the same method again and thus reproduce results. If they are experts in the field, they can analyze the implementation of the method. In this setting data is also needed—for this, accessing Linked Data via SPARQL-endpoints from within R has recently been made possible by the SPARQL-package [18].

1.3.3 Cloud Computing for Virtualization of Research

Cloud Computing refers to efforts 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. Moreover, one pays only for the traffic and computation needed, and not for the whole infrastructure. Cloud Computing also enables to handle datasets of large size, as the user does not have to handle them in their own environments. Furthermore, the 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 cause that much harm in controlled environments offered by the Cloud. The access to data in Linked Science is provided also visually by Cloud-based services.

When all the predicates for describing the data are given unique URIs, they enable creating more generic browsing and visualization 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 Science can be explored on a timeline, and by using other facets like theme, origin, author and usage history.

1.3.4 Managing Licenses and Copyrights

In Linked Science, all the research data, copyright and license information, the paper itself, actions taken on them and their provenance information will be publicly available 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 fulfill the usage permission criteria of a planned research setting.

License issues have widely been recognized as important in order to ensure transparency. A set of Panton Principles has been published recently to promote clear explication of licenses, and to ensure openness of data. As a summary, Panton Principles[28] state:

1. When publishing data make an explicit and robust statement of your wishes,

[27] See http://www.r-project.org/.
[28] Quoted from http://pantonprinciples.org.

2. Use a recognized waiver or license that is appropriate for data,
3. If you want your data to be effectively used and added to by others it should be open as defined by the Open Knowledge/-Data Definition—in particular non-commercial and other restrictive clauses should not be used.
4. Explicit dedication of data underlying published science into the public domain via PDDL or CCZero is strongly recommended and ensures compliance with both the Science Commons Protocol for implementing Open Access Data and the Open Knowledge/Data Definition.

1.4 Encoding and Linking Scientific Knowledge

1.4.1 Linked Science Core Vocabulary

The Linked Science Core Vocabulary (LSC) [29] is designed for describing scientific resources including elements[30] of research, their context, and interconnecting them. We introduce LSC as an example of building blocks for Linked Science to communicate the linkage between scientific resources in a machine-understandable way. The "core" in the name refers to the fact that LSC only defines the basic terms for science. We argue that the success of Linked Science—or Linked Data in general—lies in interconnected, yet distributed vocabularies that minimize ontological commitments. More specific terms needed by different scientific communities can therefore be introduced as extensions of LSC.

Light-weightiness, simplicity and intuitiveness have been the main design principles of LSC—it focuses on simple properties that can be used to describe the content of a research paper that is, to relate the research, hypotheses, predictions, experiments, data, and publications together. The main classes include Research, Researcher, Publication, Hypothesis, Prediction, and Conclusion. Figure 1.2 shows the main concepts and properties of the vocabulary. Property `<isSupportedBy>` can be used for stating a hypothesis is supported by a certain research. Moreover, one hypothesis `<makes>` one or more predictions, and a research `<produces>` a conclusion. Further on, `<reportedIn>` and `<reports>` are used to relate a publication, i.e. the scientific paper with the research embedded in it. The interconnection between the research and data can be made via the properties `<dataUsed>` and `<dataProduced>`. LSC provides also properties for locating the research in space and time via the properties

[29] See http://linkedscience.org/lsc/ns/.
[30] See e.g. http://en.wikipedia.org/wiki/Scientific_method.

`<isAboutRegion>` and `<isAboutTime>`. Researcher is related to Research via `<participatesIn>`.

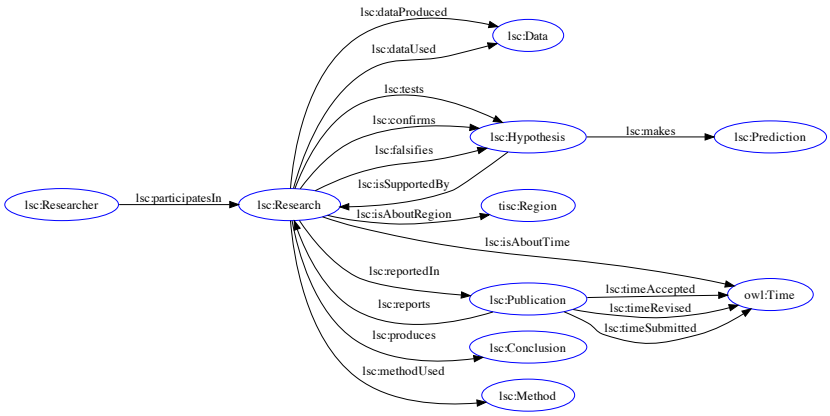

FIGURE 1.2: Graph presenting the concepts and relations of the Linked Science vocabulary.

1.4.2 Case: Describing and Linking a Research Setting

Both publications and studies reported in them can be located in space and time. In the Linked Science approach, methods, data, results are opened up and made available in a machine-processable way. Similar rationale applies to the meta-level information of scientific resources which refer to the metadata of the publication resources and the meta-level information of the research embedded in them.

The spatial location of a publication has two aspects. On the one hand it may be the place it originates from. This could be e.g. a certain university or research institute. On the other hand, location information is introduced by the spatial extent of the study which is described in it, for example deforestation in the Brazilian Amazon. Te same principle applies to the two temporal aspects of a publication—namely the date of the publication, for example the year 2006 and the temporal extent of the research it is about e.g. deforestation data from the year 1987 to 2000.

It is very crucial to be aware of these different aspects when trying to organize and interconnect the knowledge not only within one domain but also in heterogeneous ones. Especially, the visualization of spatial and temporal information of the studies sheds light on the spatial and temporal gaps and overlaps enabling better organization and management of the knowledge.

In the Linked Science Core Vocabulary, we propose terms for representing

the different spatial and temporal aspects of publications so that scientific data may be published according to the Linked Data principles. For that, the properties `<isAboutTime>` and `<isAboutRegion>` are introduced for relating the research in a publication to its spatial and temporal extend. In combination with the The Open Time and Space Core Vocabulary (TISC)[31], a pluralism of spatial expressions can be used as objects, describing the region a research is about.

Figure 1.3 shows time periods of interest of 13 different research settings. The time periods are shown by two granularities, decades (lower time band), and years (upper time band). A lighter area in the lower time band is indicates the temporal extent of the upper time band. Visualization of the time periods enables to detect gaps and overlaps between time periods. For example, in the figure we can see that data from the year 1999 is of interest in six different studies while the data for the year 2002 has been studied only in two research settings. If similar kinds of descriptions were available for all research concerning deforestation in the Brazilian Amazon, it would enable researchers and funding agencies to find gaps and overlaps in a similar fashion.

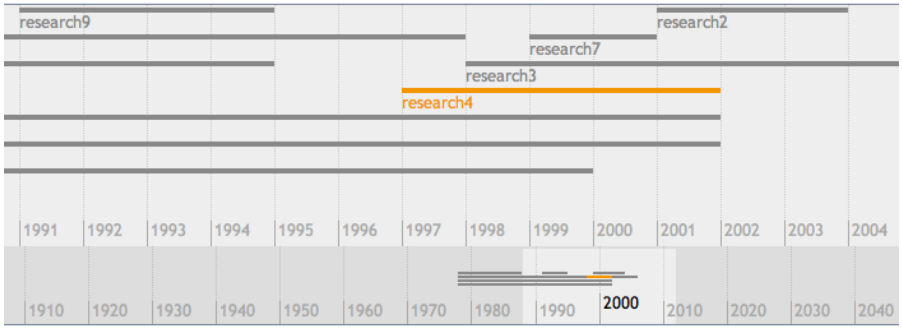

FIGURE 1.3: A timeline presenting the time periods of interest of 13 different researches made concerning the Brazilian Amazon Rainforest.

Figure 1.4 shows an extract of one of the example research settings modelled, namely a paper entitled "Is deforestation accelerating in the Brazilian Amazon?" by Laurance et al., 2001 [14]. Different LSC properties have been in use to model their research, e.g. which methods and data they used, which conclusions the research produced and so on.

1.4.3 Discussion and Related Work

There are interesting related efforts for describing science. For example, The Scientific Knowledge Infrastructure Ontology (SKIo) provides means for

[31] See http://observedchange.com/tisc/ns/.

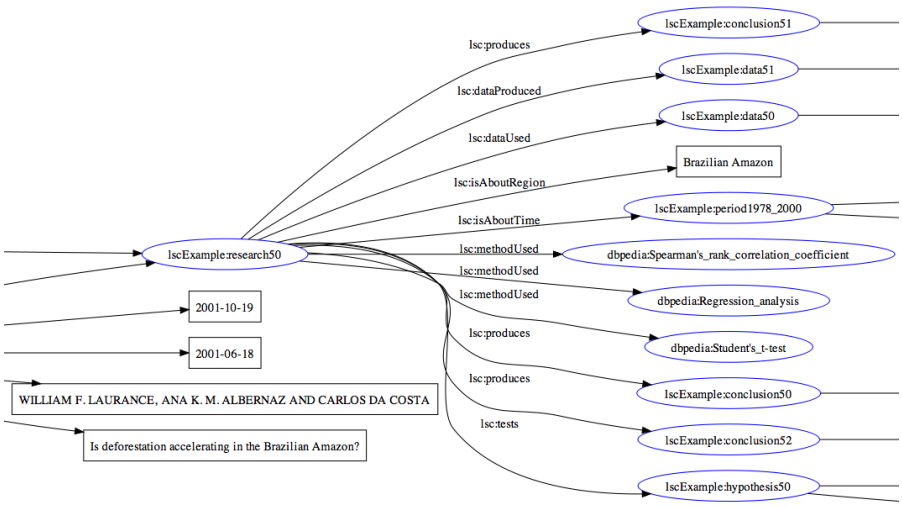

FIGURE 1.4: An example research setting modeled using LSC vocabulary.

describing scientific entities, processes and methods using ontological structures. SKIo is aligned with the foundational ontology DOLCE in order to provide basic distinctions between objects, processes, abstract entities and qualities within the science domain. Moreover, the structures from DOLCE to define causes and effects are employed. The core idea in SKIo is that it can be used to formally define scientific processes—i.e. process semantics—, and to link them to publications. In a climate modeling scenario these formal models may be e.g. used to examine multiple processes causing the same output. This means that SKIo treats processes as functions for similarity reasoning: if two functions with the same inputs—like the amount of rainfall—produce same results, then processes are essentially the same. Moreover, ontology could be used for querying all the environmental processes described in a given set of publications.

The Semantic Publishing and Referencing Ontologies (SPAR) [32] is a set of ontologies intended for describing scientific publications. Citation Typing Ontology (CiTo) [17], for example, can be used to explicate what kind of a reference relation (explicit, implicit or indirect) is in question between two publications. With CiTO, users can indicate in a machine-processable way whether they agree or disagree with other published research.

DataCite[33] 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 Web Journal[34] provides an open review process,

[32]See http://purl.org/spar/.
[33]See http://datacite.org/.
[34]http://www.semantic-web-journal.net/.

meaning that submitted papers are online, enabling anyone to comment them. Moreover, the reviews are also published online, exposing the reviewer names by default.

Linked Data seems to offer a good basis for building Linked Science. For example, CrossRef[35]—a consortium of roughly 3,000 publishers—published millions of pieces of information about publications as Linked Data at the end of April 2011. All these pieces of data may be further linked to research settings they had, methods they used, scientific knowledge they produced, in order to implement the Linked Science approach.

1.4.4 Research Agenda for the Future

Enabling exact descriptions of all relevant scientific resources is crucial in achieving Linked Science. This means that pieces of well-defined information can and should themselves become publications. A research agenda for the future should therefore include finding ways to lower the barrier of publishing small pieces of new knowledge together with verification mechanisms rather than always requiring full ten to twenty page articles. This will ultimately include a data reviewing process, comparable to the established peer reviewing processes for scientific publications.

In Linked Science, browsing and reproducing results should be as easy as it is today to click pages on the Web. The research agenda for the future should thus concentrate in making sharing and reusing of scientific knowledge as easy as possible.

In this chapter, we proposed and presented the Linked Science Vocabulary (LSC) and showed how to use it for describing a research setting. We foresee that the research agenda for years to come should include to further develop and evaluate LSC in different scientific settings. This includes also research about required reasoning mechanisms to interconnect scientific resources.

The listed topics of the 1st International Workshop on Linked Science[36] provide some ideas where the research on Linked Science should focus on. These topics include e.g. formal representations of scientific data, integration of quantitative and qualitative scientific information, ontology-based visualization of scientific data, semantic similarity in scientific applications, semantic integration of crowd sourced scientific data, and connecting scientific publications with underlying research datasets.

To continue this list, provenance, quality, privacy and trust of scientific information are also crucial. Other goals include enrichment of scientific data through linking and data integration, and having case studies on linked science, i.e., statistics, environmental monitoring, etc. There is also a need for developing Linked Data practices for dissemination and archiving of research results, collaboration and research networks, and for research assessment. The

[35] http://www.crossref.org/.
[36] See http://linkedscience.org/events/lisc2011.

development of application scenarios of Linked Science together with all their legal, ethical and economic aspects provide interesting research topics.

Finally, the research agenda should also take into account the whole variety of information in academic settings. For example, for encoding information related to academic offerings, there are vocabularies such as the Academic Institution Internal Structure Ontology (AIISO), the Teaching Core Vocabulary[37] (TEACH), Metadata for Learning Opportunities (MLO), XCRI Course Advertising Profile (XCRI-CAP), the Dublin Core metadata terms, The Friend of a Friend-vocabulary (FOAF), and the Open Provenance Model Vocabulary. These and novel vocabularies should be further developed, evaluated and taken into an efficient use for increasing linkage between scientific resources.

The Linked Science Core Vocabulary naturally faces the same challenges that have prevented adoption of many other ontologies and vocabularies. Our aim is to tackle these challenges by 1) providing a lightweight structure, including only core terms and predicates and therefore minimizing ontological commitment, 2) providing clear examples of the use, 3) enabling the LSC vocabulary to be technically accessible as Linked Data, and by 4) involving the community to further develop and extend it.

1.5 Conclusions

In this book chapter we have laid the foundations for interconnecting scientific resources in order to increase transparency, openness, and reproducibility of science. We first explained the concept of Linked Science, i.e. publishing data using Web techniques, opening and running of methods in the Cloud for reproducing scientific processes and explicating copyright and license issues. We also gave an overview of the technologies required for Linked Science. This includes publishing of data, methods, resources, results, license and provenance information together with the documentation, and to help establishing trust related to them.

Linked Science therefore 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. We gave an example of how scientific knowledge can be encoded as Linked Data by using a combination of existing and developed vocabularies and ontologies. We showed how the approach works with a set of publications related to research about deforestation in the Brazilian Amazon Rainforest. Formalization of the scientific results aimed at sharing the knowledge of the findings. Through visualizations people are able to actually see what time periods and spatial regions are covered by the

[37] See http://linkedscience.org/teach/ns/.

knowledge resulting from the research. Results may therefore be be used to get an overview of the research in certain domain, and further to find out potential gaps that need to be filled by new research settings.

1.6 Glossary

Linked Data: A way to publish data using Web standards and techniques. In the Linked Data approach, all resources are identified and resolvable via HTTP URIs, and there is a linkage between the resources.

Linked Science: An approach where scientific resources—e.g. processes, models, data, methods and evaluation metrics—are semantically linked.

LODUM: The Linked Open Data University of Muenster project works on opening up and linking research and educational data.

Semantic Web: A vision of a Web where meanings of things are explicitly defined such that autonomous agents can share these meanings, connect to and combine results of various semantic services and perform useful actions for a user.

1.7 Acknowledgements

This research has been partially funded by the International Research Training Group on *Semantic Integration of Geospatial Information* (DFG GRK 1498), the Linked Open Data University of Münster (LODUM) project and by the German Academic Exchange Service (DAAD). In addition, the authors would like to thank the insightful comments obtained through the anonymous peer-review process.

Bibliography

[1] S. Bechhofer, J. Ainsworth, J. Bhagat, I. Buchan, P. Couch, D. Cruickshank, D. D. Roure, M. Delderfield, I. Dunlop, M. Gamble, C. Goble, D. Michaelides, P. Missier, S. Owen, D. Newman, and S. Sufi. Why Linked Data is Not Enough for Scientists. In *e-Science (e-Science), 2010 IEEE Sixth International Conference on*, pages 300–307. IEEE, December 2010.

[2] Tim Berners-Lee. Linked Data. Personal view available from http://www.w3.org/DesignIssues/LinkedData.html, 2009.

[3] Tim Berners-Lee, Jim Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.

[4] Y. Bishr. Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science*, 12(4):299–314, 1998.

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data—the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[6] Davis Clodoveu, Gilberto Câmara, and Frederico Fonseca. Beyond SDI: integrating science and communities to create environmental policies for the sustainability of the Amazon. *International Journal of Spatial Data Infrastructure Research (IJSDIR)*, 4:156–174, 2009.

[7] P.M. Fearnside and R.I. Barbosa. Accelerating deforestation in Brazilian Amazonia: towards answering open questions. *Environmental Conservation*, 31(01):7–10, 2004.

[8] P. Groth, A. Gibson, and J. Velterop. The anatomy of a nanopublication. *Information Services and Use*, 30(1):51–56, 2010.

[9] T.R. Gruber et al. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5:199–199, 1993.

[10] Olaf 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*, April 2009.

[11] Olaf Hartig and Jun 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.

[12] Tomi Kauppinen and Giovana Mira de Espindola. Linked Open Science—communicating, sharing and evaluating data, methods and results for executable papers. In *The Executable Paper Grand Challenge, proceedings of The International Conference on Computational Science (ICCS 2011)*, Elsevier Procedia Computer Science series, Singapore, 2011.

[13] Tomi Kauppinen, Jari Väätäinen, and Eero Hyvönen. Creating and using geospatial ontology time series in a semantic cultural heritage portal. In *S. Bechhofer et al.(Eds.): Proceedings of the 5th European Semantic Web Conference 2008 ESWC 2008, LNCS 5021, Tenerife, Spain*, pages 110–123, 2008.

[14] William F. Laurance, Ana K. M. Albernaz, and Carlos Da Costa. Is deforestation accelerating in the Brazilian Amazon? *Environmental Conservation*, 28(04):305–311, 2001.

[15] Barend Mons, Herman van Haagen, Christine Chichester, Peter-Bram t'Hoen, Johan T den Dunnen, Gertjan van Ommen, Erik van Mulligen, Bharat Singh, Rob Hooft, Marco Roos, Joel Hammond, Bruce Kiesel, Belinda Giardine, Jan Velterop, Paul Groth, and Erik Schultes. The value of data. *Nature Genetics*, 43(4), March 2011.

[16] Alexandre Passant, Paolo Ciccarese, John Breslin, and Tim Clark. SWAN/SIOC: Aligning scientific discourse representation and social semantics. In *Workshop on Semantic Web Applications in Scientific Discourse*. The 8th International Semantic Web Conference (ISWC 2009), 2009.

[17] David Shotton. CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics*, 1(Suppl 1):S6+, 2010.

[18] Willem Robert van Hage and Tomi Kauppinen. SPARQL client for R. available from http://cran.r-project.org/package=SPARQL, 2011.

[19] Jun Zhao, Graham Klyne, and David Shotton. Provenance and linked data in biological data webs. In *The 17th International World Wide Web Conference (LDOW2008)*, 2008.