Linking the Outcomes of Scientific Research: Requirements from the Perspective of Geosciences

Stephan Mäs, Matthias Müller, Christin Henzen, Lars Bernard

Technische Universität Dresden
Department of Geosciences, Professorship of Geoinformation Systems,
Helmholtzstraße 10, 01069 Dresden, Germany
{Stephan.Maes, Matthias_Mueller, Christin.Henzen, Lars.Bernard}@tu-dresden.de

Abstract. The paper discusses the required linked data when scientific geographical data is published on the Web in a legible and traceable manner. The presented work results from the research project GLUES which aims at establishing an interdisciplinary platform for scientific data exchange. To facilitate the search for data and to assist the evaluation of the fitness for use, the published data must be connected to further metainformation, e.g. thematic classifications, semantic definitions of keywords and the data origin. Here the Linked Data paradigm promises manifold advantages, in particular if the reference to simulations or models the data originate from shall be provided.

Keywords: linked scientific data, scientific geodata infrastructure, GDI

1 Introduction

Traditionally, the evaluation of scientific work mainly refers to published articles and their impacts. Nowadays, this is by far not the only outcome of scientific work. In particular the computational branches of science produce data which might be valuable beyond its original scope [1]. The technological process enables scientists to collect and produce much more data about our world than ever before. For example, a single simulation of global climate change over the next 100 years easily produces several Gigabytes of data. If such data and the underlying calculation models and assumptions are sufficiently described it can be valuable input for other scientists. Such exchange can stimulate the reuse of scientific data, the extraction of new information [2], the collaboration amongst scientists and support data-intensive multidisciplinary research. The improved documentation of research results would make scientific work more transparent, in the optimal case even reproducible, allow for the evaluation of fitness for further use [3], and increase its sustainability. Further, Web based visualizations and analysis tools for the comparison of different data sets could support stakeholder work and provide policy makers with insights from scientific research.

Therefore, the publication of scientific data in combination with a reference to the respective publication is necessary, but certainly it is only a first step. To facilitate the search for data and the evaluation of appropriateness for a certain task the data must be connected to further information, in particular concerning the data origin, as well.

In geosciences, researchers might for example search for all available output data of a specific numerical model or for biodiversity data that refer to a particular climate scenario for a time period in the future. To answer such queries the data must be linked with the numerical models, the concrete parameters of the model run, the input data of that run, scenario descriptions, storylines and descriptions of basic assumptions of the model.

This paper summarises some requirements on the publication of scientific data on the Web and how these data must be interlinked with other information. The focus is on the publication of output data of numerical or statistical models of geographically referenced phenomena, such as geodata about land use, biodiversity, water resources, climate, socio-economy or agriculture. Such geodata is usually highly dimensional and time variant and therefore relatively complex. At present, even if such data is discoverable and accessible, the assessment of the data quality with regard to a particular use is difficult. The corresponding scientific articles summarize information about the data producing methods, but they are more focused on new scientific insights and, with regard to a usability evaluation, they do not sufficiently describe the data and its quality in a structured and comprehensible way.

The presented work results from the research project GLUES (Global Assessment of Land Use Dynamics, Greenhouse Gas Emissions and Ecosystem Services). One of the aims of GLUES is to establish a platform for the facilitation of interoperable and interdisciplinary data exchange between scientists. The following chapter provides an insight into the GLUES project. After that some requirements of this data exchange regarding the linkage of scientific data and other information sources are elucidated.

2 Project Background

GLUES is the coordination project of the international interdisciplinary research program ‘Sustainable Land Management’ (LAMA) of the German Ministry of Education and Research. Within this funding measure ten so called regional collaborative projects (RPs) are researching the impacts of climate and socio-economic changes and a corresponding optimization of the use of land and natural resources in different countries and regions. Since this interdisciplinary research is policy-oriented the projects closely cooperate with regional scientists and stakeholders. As coordination project, GLUES is a support action for the RPs. The major aims of GLUES are to support the communication, coordination, facilitation of data exchange and integration of results, by developing a common data platform and consistent scenarios on land use, climate and social-economic change. GLUES will provide a data pool for common use within the LAMA funding measure and a set of consistent global scenarios for the medium and long term projections.

For an effective synthesis of research results the underlying base scenarios and the data sets must be disseminated and shared between the involved research institutions. Technically, the access to modeling and scenario results of GLUES and the RPs will be provided by means of a scientific Geodata Infrastructure (GDI). Such GDI realizes a network of Web services enabling standardized access to distributed geodata in combination with visualization and analysis functions. The GLUES GDI serves three main purposes:

- to allow the involved research teams to effectively disseminate and share their model and analysis results as well as the underlying base scenarios and data sets
- to support a seamless integration of existing resources which for instance serve as input to the scientific models or as reference data for comparative analysis
- to offer robust exploration and analysis tools to support stakeholders in applying the GLUES results and findings in their planning and management activities

Therewith the GLUES GDI provides a common infrastructure to publish, share, reuse and maintain distributed global and regional data sets as well as model results on scenarios of land use, climate change and economic development. It supports the technical collaboration of the GLUES partners and the RPs of the funding measure LAMA and provides the technical basis for outreach activities.

Most of the scientific data that shall be made available in GLUES has a raster or grid structure and is exported from the models/simulations as netCDF or TIFF or in self defined data formats which are not standardized. Some economic datasets have a tabular structure referring to the corresponding administrative units. In the GDI all data is made available in the standardized data formats that are supported by the majority of the GIS on the market.

The Linked Data paradigm[1] for publishing data promises manifold advantages for discovering and retrieving scientific data in the GDI. Therefore it is planned to use corresponding links in the metadata descriptions. The structured connection from the scientific data to the models they are originating from and to other describing metadata allows for more precise descriptions of research results. Therewith, it supports the disambiguation and alignment of common vocabularies and the data and facilitates data integration, aggregation and use [4]. The following chapters summarize, where the publication of scientific geodata demands links to such comprehensive background knowledge.

3 Metadata of Scientific Data

The central component of the GLUES GDI is a data catalogue which supports metadata searches and metadata acquisition. The catalogue enables to search the available metadata for data and service categories, application domains, keywords and names of datasets and services, for instance. If the corresponding Web services are available the catalogue provides direct links for download and visualizations. Registered users of GLUES or the RPs can also publish metadata of their scientific data sets in the catalogue's data base. The metadata is acquired in a structured way according to a common data model (conform to the Infrastructure for Spatial Information in Europe (INSPIRE) regulation [5] and ISO 19115 [6]). The metadata contains elements for:

- identification / descriptive information (e.g. title, type, abstract, data provider contact information, spatial and temporal extent and reference system, dates of publication, revision and creation of the data)
- categorization (topic / thematic classification, keywords)

[1] http://www.w3.org/DesignIssues/LinkedData.html

- access and licensing information (e.g. data licenses, access restrictions)
- distribution (e.g. data format, link to online resources)
- quality (spatial and temporal resolution, lineage information)
- metadata on metadata (e.g. metadata point of contact, metadata date)

Some of these metadata elements, like the spatial and temporal reference system and data license, only allow for standardized or at least unified and commonly used entries. Ideally, such entries link to a corresponding common vocabulary or registry of terms and definitions. For example, the spatial reference systems use EPSG codes[2] to refer to standardized entries.

The thematic classification of the data and the provided keywords should connect to common controlled vocabularies (thesaurus, ontology) for classifications to identify words and semantics. For a spatial data set INSPIRE requires that at least one keyword from the General Environmental Multilingual Thesaurus (GEMET) is provided to describe the relevant spatial data theme [5]. The GLUES GDI supports GEMET as a basic vocabulary to preserve INSPIRE compliance, although it proved to be insufficient for many scientific terms. At present, there are other domain-specific vocabularies available or under development (e.g. WMO BUFR for atmospheric conditions, GEOSS ontology), but generally these are not used among the scientific research groups in LAMA. Some of these nomenclatures reveal inconsistencies in their definitions when they are combined with others, such that the nomenclatures are partially incompatible. Controlled vocabularies for scientific terms, like the science ontology developed in [8] and [9], hardly exist. For environmental modelling in the geosciences this already starts with basic terms like model, scenario, storyline, driver and indicator. Although frequently used different scientific communities have a slightly biased understanding of these terms. Creating a detailed and unambiguous formal description of such terms and, particularly, communicating it to a wider audience is strongly required and a pressing challenge in the near future: while in a face-to-face discussion ambiguities can be resolved this is hardly possible in a catalogue query.

The spatio-temporal scale and the level of detail of the scientific data are diverse and the corresponding descriptions can be complex. Input data of numerical models very often refer to statistical data with common administrative units, like provinces and countries, as spatial resolution. In global economic models these units are often not separately considered and aggregated to larger, equally sized regions to create a uniform sample size. Depending on modelling goals and the expected outputs these aggregated spatial regions can be diverse and are task specific. Nevertheless, the aggregation procedure is hardly documented if the data is published. For a user of datasets with differing aggregation units a comparison and integration is usually extremely time consuming and requires educated guesses. Beside the spatial resolution also the scale of the geographical phenomena can be diverse. Different objectives of models lead to different thematic categories in the data, for example differing nomenclatures of land cover or agricultural products. To support transformation tasks the metadata must contain resolvable links to the corresponding sets of spatial aggregation units and thematic categories. For the aggregation units

[2] http://www.epsg.org/CurrentDB.html

these links could point to gazetteers like for example the Geonames geographical database[3].

For Web services that enable data download and visualization links to data schemata and visualisation schemata are required. Obviously datasets can be much easier combined, if they link to the same data schema. Corresponding (map-) visualizations can be intuitively interpreted if the same visualisation schema is applied.

4 Linking Scientific Data and Models

To evaluate the fitness for use of a dataset information about its origin is vital. Such lineage information is also contained in the ISO standard [7]. It provides elements for linked information, but needs profiling to restrict its broadness for an automated processing of the information, such as linking to related data. Therefore the lineage metadata element has been slightly adjusted in the GLUES catalogue. It provides references to corresponding scientific literature, which is certainly the main requirement when scientific data is published. Further, it contains information about the origin of the data such as the data acquisition method, measurement methods, sensor information (e.g. for remote sensing data) or the applied refinement processes.

For the GLUES GDI, in particular references to numerical models and simulations along with their corresponding input and output datasets are relevant. If such links are systematically provided, the catalogue allows for querying interrelationships between different data sets and models, which can also be visually illustrated. The focus of such visualization can be either on an input dataset, a numerical model / simulation or an output dataset, such that the use of a dataset, the different inputs and outputs of a model or the origin of a dataset are visualized. For example, focusing on information about an input dataset or scenario, it can be shown which models use this data and what outputs they produce. Therewith scientists get a comprehensive view which models provide data for a certain scenario. Hitherto, such information was coupled with an extensive investigation of literature. Beside the scientific work, such comparison can also be of interest for research assessment, since it shows the "impacts" of a dataset.

5 Summary

The Linked Data paradigm brings great possibilities to the publication of scientific geodata. The experiences of the GLUES project show that for such data the correct, formalized and detailed metadata description is in many cases more important than the direct accessibility of the data itself. Such metadata includes references to scientific articles, models, common vocabularies and reference systems. For most of the data the permanent provision on the Web is not efficient, since the number of users is relatively small.

[3] http://www.geonames.org/

An improved discoverability, accessibility and usability evaluation are not only advantageous for scientists but may also provide new metrics for the assessment of research outcomes. Thus, it would also support the work of science managers and strategists in funding organizations or research institutes. Further, research results could be much easier disseminated to stakeholders at all different levels, like those of local, regional or global organizations, but also to the general public.

References

1. Gray, J.: eScience: a transformed scientific method. In: Hey, A. J. G.; Tansley, S.; Tolle, K. M. (Eds.), The fourth paradigm: data-intensive scientific discovery. Redmond, WA, Microsoft Research. (2009)
2. Dadzie, Aba-Sah; Rowe, Matthew: Approaches to Visualising Linked Data: A Survey. In: Semantic Web, Volume 2, Number 2 / 2011, IOS Press
3. Devillers, R.; Beard, K.: Communication and Use of Spatial Data Quality Information in GIS. In: Devillers, R. and Jeansoulin, R. (eds.), Fundamentals of Spatial Data Quality, Geographical Information System Series, chapter 12, pages 237-254, ISTE Ltd., London.
4. Auer, S.; Lehmann, J.; Hellmann, S.: LinkedGeoData - Adding a Spatial Dimension to the Web of Data. In: Proceedings of 7th International Semantic Web Conference ISWC, Lecture Notes in Computer Science, volume 5823, pages 731-746, Springer, (2009)
5. Implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata, (Commission Regulation (EC) No 1205/2008 of 3 December 2008), Official Journal of the European Union, (2008)
6. ISO/TC211: ISO 19115 - geographic information - metadata (2002)
7. ISO/TC211: ISO 19115-2 - geographic information - metadata - part 2: metadata for imagery and gridded data
8. Brodaric, Boyan; Reitsma, Femke; Qiang, Yi: SKIing with DOLCE: toward an e-Science Knowledge Infrastructure. In: Proceedings of the Fifth International Conference (FOIS 2008), Carola Eschenbach and Michael Grüninger (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands, 208-219, 2008
9. Brodaric, Boyan:. Science Knowledge Infrastructure Ontology 3.0. U.K. e-Science Technical Report Series, Report UKeS-2008-03, 40 pp., 2008, accessed at 15th September 2011 http://www.nesc.ac.uk/technical_papers/UKeS-2008-03.pdf.