Representing Research Collaborations and Linking Scientific Project Results in Spatial Data Infrastructures by Provenance Information

Christin Henzen
Chair of Geoinformatics, Technische Universität Dresden, Germany
christin.henzen@tu-dresden.de

Stephan Mäs
Chair of Geoinformatics, Technische Universität Dresden, Germany
stephan.maes@tu-dresden.de

Franziska Zander
Chair of Geoinformatics, Friedrich Schiller University of Jena, Germany
franziska.zander@uni-jena.de

Matthias Schroeder
Zentrum für GeoInformationstechnologie, Deutsches GeoForschungsZentrum GFZ Potsdam, Germany
matthias.schroeder@gfz-potsdam.de

Lars Bernard
Chair of Geoinformatics, Technische Universität Dresden, Germany
lars.bernard@tu-dresden.de

Abstract

Provenance information describes the history of a dataset. In a scientific use case, provenance information can be further used to support reproducibility of data and to show collaborations among researchers. In this paper, we show how to use provenance information in scientific SDIs to show relations among research projects and between datasets. An interdisciplinary research program, in which research results (e.g. global climate change scenarios) are published in several linked geoportals serves as underlying use case. We address issues and requirements on modelling provenance on project, dataset and implementation level.

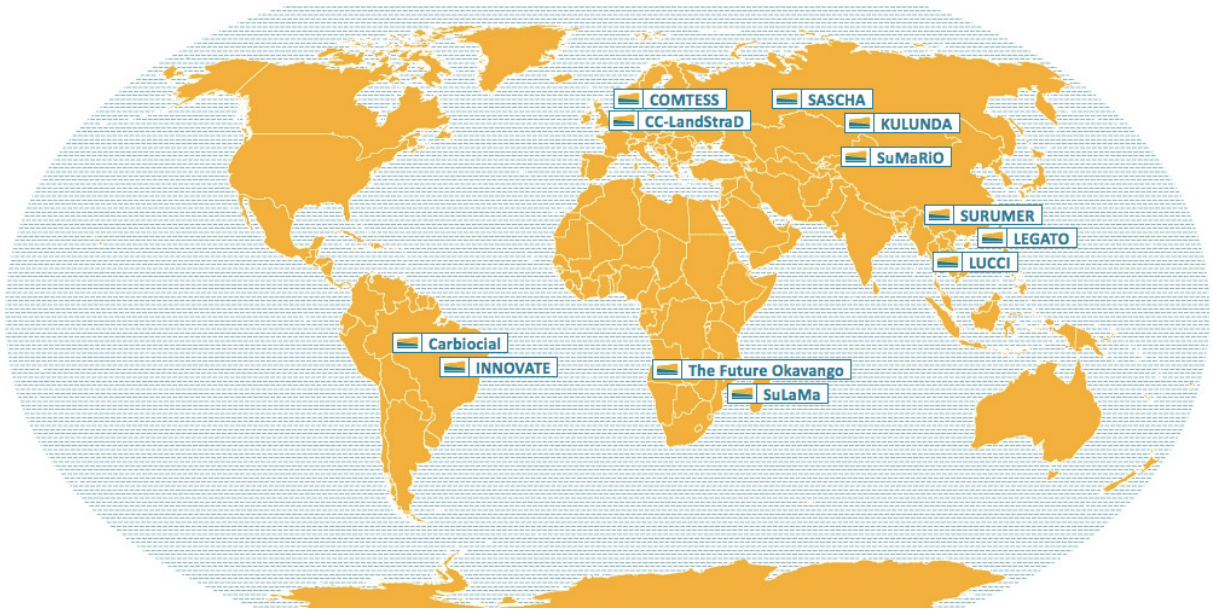
Keywords: Provenance, Spatial Data Infrastructure, Metadata, Research Collaborations.

1 Introduction

Geospatial data provenance describes the history of a dataset – its data sources and data creation processes. It is of major interest for data usability evaluation: Identifying original data sources and learning about previous treatments ([1], [2]).
In available spatial data infrastructures (SDI) provenance information is often neglected. But in particular for data outputs from scientific environmental simulations the provenance information is indispensable to understand data contents and quality. Further, it facilitates transparency, maintenance documentation and might ensures reproducibility ([3], [4], [5]).

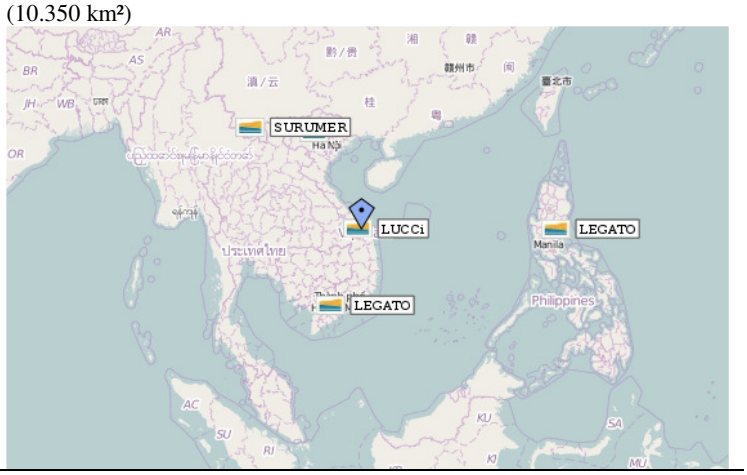
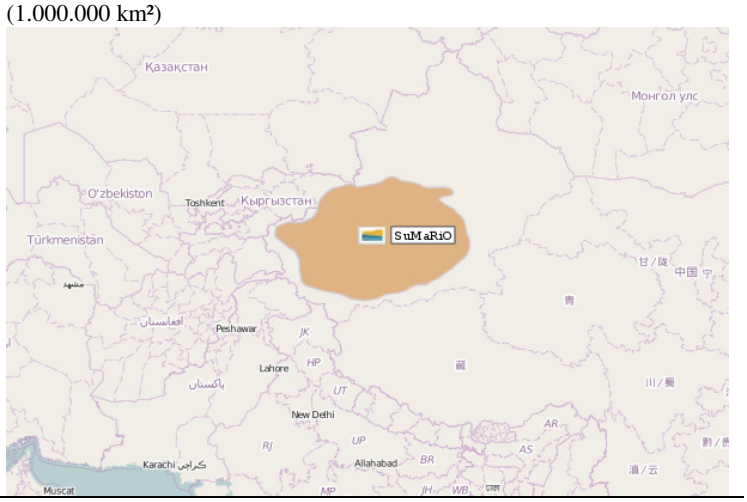
This paper shows how provenance information can be used in scientific SDIs to show relations between datasets, but also among research projects and institutions. While the research

Figure 1: 12 Regional projects of the funding measure *Sustainable Land Management* and their study areas.



Source: http://modul-a.nachhaltiges-landmanagement.de/en/projects/

Table 1: Three projects of the funding measure *Sustainable Land Management*

Project name, production systems on which they work in their study sites, research questions and website	**Study sites**
LUCCi *Industrial and extensive agriculture, forestry, tourism, rubber production* Which role does land-use play for GHG emissions? What sustainable land / water management strategies can cope with climate change impacts? http://www.lucci-vietnam.info/	Vu Gia Thu Bon River Basin, Provinces Da Nang and Quang Nam (10.350 km²) 
SuMaRiO *Industrial and extensive agriculture, forestry, settlement* How to support oasis management along the Tarim River under conditions of climatic and societal change? http://www.sumario.de	Tarim River Basin including the catchment areas of its tributaries, i.e. Aksu River, Yarkand River, Hotan River, Kaikong River, Kaidu River (1.000.000 km²) 
SUSTAINABLE LAND MANAGEMENT GLUES	No study sites; global analysis

collaboration among scientists is in the focus of e-science Infrastructures like myExperiment[1] [6] or iPlant Collaborative[2], it is hardly represented in SDIs. Scientific SDIs can be used to share and disseminate research results and analysis methods [9]. We address the issues and requirements on modelling and implementing data provenance documentations in a scientific SDI.

As underlying use case we utilize research results, in particular datasets, from the international interdisciplinary research program *Sustainable Land Management* (http://modul-a.nachhaltiges-landmanagement.de/en), in which projects sharing geospatial datasets (e.g. of global climate change scenarios) for their research activities. The program addresses global and regional challenges related to land management, climate change and ecosystem services and aims to provide new perspectives on sustainable land management. Within this program, 12 of the projects are regional collaborative projects (RPs) focusing on different regions (Figure 1) and being supported by the scientific coordination project *GLUES*. In this paper we focus on results of three projects: *LUCCi* (*Land Use and Climate Change*

[1] http://www.myexperiment.org
[2] http://www.iplantcollaborative.org

Interactions in Central Vietnam), *SuMaRiO* (*Sustainable Management of River Oases along the Tarim River / China)* and *GLUES* (*Global Assessment of Land Use Dynamics, Greenhouse Gas Emissions and Ecosystem Services*) (Table 1). We focus on the use of provenance information to represent the data exchange and collaboration between the research projects, the history of resulting scientific environmental data and its technical implementation.

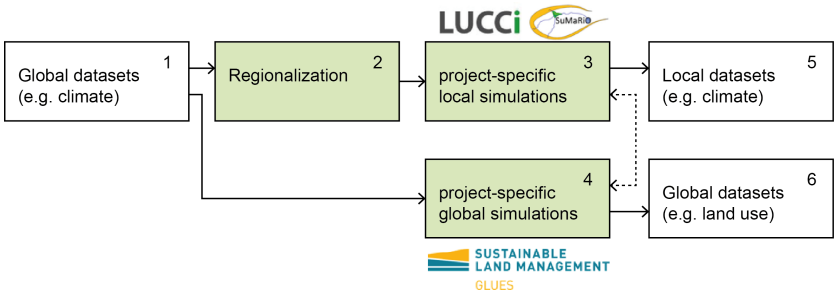
2 Representing collaborations and links among research projects

The assessment of scientific projects commonly refers to the outreach of research results, here geodatasets, and the collaboration between different institutions. In this context, provenance information can be used to answer the following questions:

- How are projects linked to each other (e.g. through common input datasets, scenarios)? How do different projects, researchers or research institutions cooperate?
- How can datasets be used/combined to answer different research questions in different projects? Which project specific solutions are transferable/comparable to other projects?
- In which projects a particular dataset has been used?

In the *Sustainable Land Management* program projects can be linked indirectly by using the same inputs for several project-specific simulation models and directly by using outputs of a project-specific simulation model as input for another project-specific simulation model. To directly visualize the collaboration and data exchange among research projects simplified graphs like in Figure 2 can be generated. This graph shows two typical provenance paths for datasets created by indirectly linked simulation models. One path describes typical RP data analysis - the regionalization of global datasets (Fig. 2: 1-3) and the use of resulting datasets as input for project-specific local simulations (3, 5). The other path describes the creation of *GLUES* project results by using the same global datasets for creating other global datasets, such as land use maps (Fig. 2: 1, 4, 6). In the research program, linking projects directly means that global *GLUES* datasets can be used as input for RPs' simulation models or *GLUES* models can be refined or validated with RP outputs.

Figure 2: Simplified data provenance graph in the *Sustainable Land Management* program on project level.



Focusing on a datasets-specific view, provenance information can be used to answer the following questions:

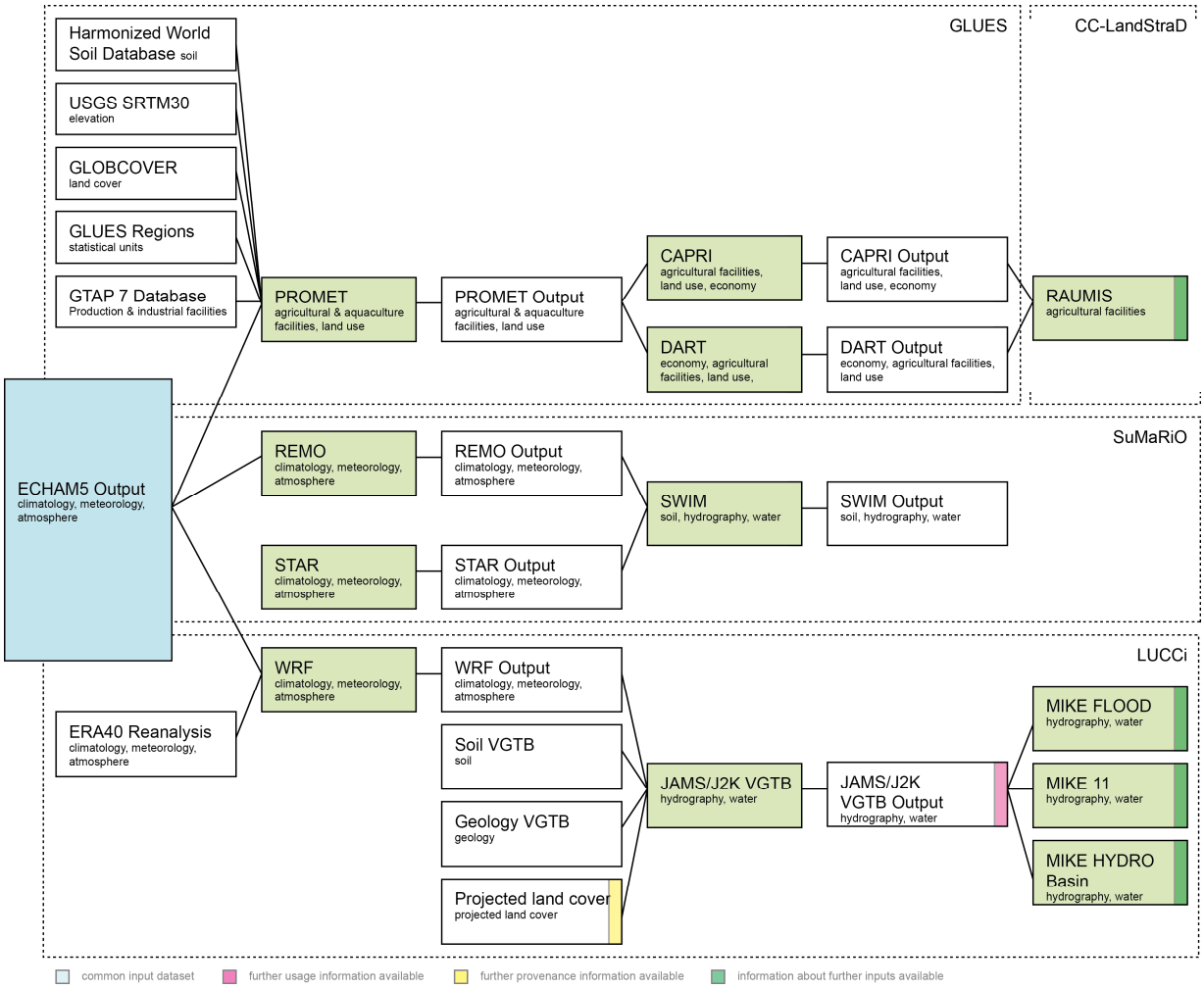
- How was the dataset generated?
- Which data has been used for the generation of a dataset?
- Which data has been generated using a given dataset?
- At which stage of a processing chain is a given dataset?

Within the program several restrictions on provenance modelling on dataset level are made to fit the data and the user requirements, e.g. skip collecting pre-processing steps, such as schema transformation. Datasets, models and relations are characterized as follows [8]:

- A model
  - is represented as an atomic process step
  - can have several inputs and one output, which might summarizes several (atomic) outputs
  - is not directly connected to another model
- A dataset
  - can be input of one or several models
  - can be output of a model and input of another one
  - is identified by a unique id (or name)

Providing complex provenance graphs for distributed datasets and metadata requires the definition of 1) a common level of detail for dataset descriptions and 2) a common mapping strategy that enables to merge distributed provenance information, e.g. data processing across projects. In our use case, commonalities for datasets's provenance are derived from models involved their creation process. Datasets that are used multiple times are recognized by their ids (or names). The mapping process across SDIs is prepared manually by examine published datasets and define and store same ids or names in referencing metadata.
Figure 3 shows a simplified graph of selected *GLUES*, *SuMaRiO* and *LUCCi* geodatasets and the common usage of *ECHAM5 datasets*. In *GLUES,* the global land use and economic simulation models *PROMET, CAPRI* and *DART* are coupled (Figure 3, Table 2). *PROMET*'s data provenance and usage is described as follows: six datasets serve as inputs for *PROMET*, a land use simulation model (Appendix, Table 2).

Figure 3: Simplified provenance of selected *GLUES, SuMaRiO* and *LUCCi* geospatial datasets.



The generated output is used, together with other datasets (here not displayed, e.g. from FAO or OECD) as inputs for the agricultural and economic models *CAPRI* and *DART*. *CAPRI* and *DART* outputs summarize about 500 resp. 200 single datasets. Both models are coupled with other RP models, such as *CC-LandStraD*'s[3] RAUMIS by using their outputs as *RAUMIS* inputs and *RAUMIS* outputs to validate them.

In *SuMaRiO* and *LUCCi* the *ECHAM5* climate datasets of different climate scenarios serve as input data for a regional downscaling with *REMO* reps. *WRF*. In *SuMaRiO* the results were used together with the results of the statistical downscaling model *STAR* to investigate climate and land use change with *SWIM*.

In the selected example for *LUCCi* not only the projected climate scenarios, but also the historical *ERA40* reanalysis datasets were used as input for the regional downscaling with *WRF*. The results were used as climate data input together with soil, geology, elevation datasets, and a combined projected land cover scenario for 2020 to model the impact on the water availability and sediment load in the Vu Gia and Thu Bon (VGTB) catchment with the model *JAMS/J2K VGTB*. Calculated amounts of water at several points inside the catchment and the outlet have been used as input for the application of *MIKE FLOOD*, *MIKE 11* and *MIKE HYDRO Basin*.

3 Challenges

When modelling dataset's relations, in particular provenance across projects, two challenges have to be faced: 1) the definition of a common level of detail for dataset and model descriptions and 2) a concept for the creation of unique identifiers.

In SDIs the level of detail is defined by the implemented metadata standard, which typically does not address all aspects, e.g. whether to describe models as atomic process or as several processes, or whether to model outputs as single dataset or series.

Concepts for unique and persistent identifiers already exist for scientific publications and data, e.g. the digital object identifier (DOI) implemented by several DOI registration services[4]. However, in most SDIs (randomly) generated IDs, e.g. created by catalogue software, are used instead of (data) DOIs [9].

[3] CC-LandStraD is one of the 12 RP in the Sustainable Land Management program http://www.cc-landstrad.de

[4] http://dx.doi.org, www.datacite.org

Today's SDI hardly implement provenance concepts. Even though provenance metadata standards are implemented in standard SDI software, e.g. Geonetwork or pycsw, a suitable processing of provenance information, e.g. visualization of provenance graphs, can hardly be found.

Taking a step back, the automated or facilitated (provenance) metadata acquisition is still a pressing challenge. Especially when dealing with complex provenance information, users need simple clients to publish metadata for their (huge amounts of) data, and might need assistance in linking to provenance sources or usage information.

To enable existing metadata catalogues or other SDI clients to handle provenance information across SDIs, concepts for an (automated) mapping of SDI specific (internal) IDs, e.g. by referencing frequently used datasets or descriptions via registries, are required.

4 Implementation

In the *Sustainable Land Management* program scientific project results are either published in the *GLUES* SDI or in a project specific SDI. Each SDI can contain different base and result datasets, e.g. the *LUCCi* SDI provides measured and simulated time series data, soil profiles, and geodatasets for the study site, whereas the *GLUES* SDI mainly provides global simulation results as geodatasets (and visualization or download services).

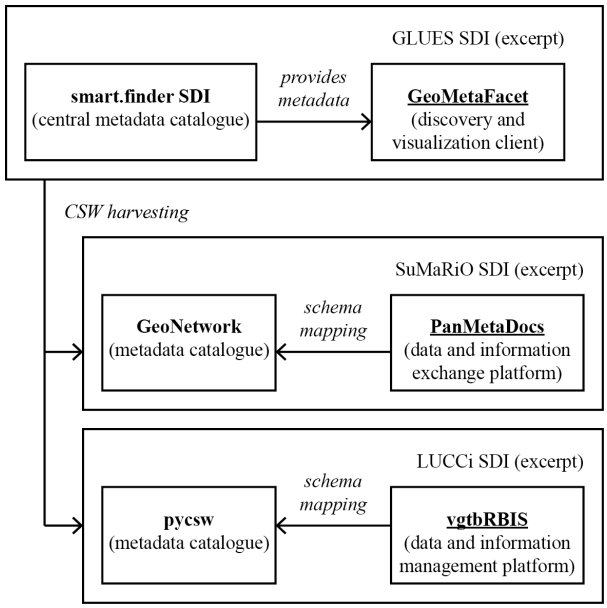
The *GLUES* SDI serves as central SDI node, which links to the project SDIs following the Catalogue Service for the Web (CSW) interface [9] (Figure 4). The *GLUES* metadata catalogue *smart.finder SDI* harvests all available RP catalogues providing metadata according to the ISO19115-2 and enables discovery of provenance information.

Due to the complexity and the high amount of (coupled) metadata, a visual illustration of provenance is essential. *GeoMetaFacet*[5] provides program specific access, discovery and visualization functionality, such as graph-based provenance visualizations[6] of discovered geodatasets [7]. The interactive provenance graph shows a navigable representation of who, what, when and how the data is generated (cp. Figure 3).

In the *SuMaRiO* project a WebGIS for data analysis supplements two metadata management applications. *PanMetaDocs*[7] is a client to exchange scientific information in various resource information types, such as PDF. It supports the project internal data distribution and collaboration between the different working groups of *SuMaRiO*. Further, *PanMetaDocs* provides an OAI-PMH[8] interface, which enables other metadata catalogues to harvest the published information resources [10]. In the *SuMaRio* project, the open source application GeoNetwork is used to harvest the resource metadata stored in PanMetaDocs. GeoNetwork has a CSW interface providing metadata according to the ISO19115-2, which is harvested by the *GLUES* metadata catalogue.

[5] http://geoportal-glues.ufz.de/stories/geometafacet.html
[6] http://geoportal-glues.ufz.de/MetaViz/detail.jsp?id=glues:lmu:metadata:dataset:promet
[7] http://pmd.gfz-potsdam.de/sumario/index.php
[8] http://www.openarchives.org/pmh/

Figure 4: Linked scientific geodata infrastructures of the Sustainable Land Management program (project specific implementation



The *LUCCi* SDI consists of the environmental information system Vu Gia Thu Bon RBIS (VGTB RBIS[9]) and the Open Source and OGC compliant CSW server implementation pycsw[10]. The *VGTB RBIS* is based on the modular structured software platform *RBIS* (River Basin Information System) ([11], [12]). Within RBIS datasets (e.g. time series data, geodata, documents) can be described by detailed metadata based on elements available in ISO 19115-1 [13] with various data type specific extensions and adaptations. Important for the management of provenance information is the extension to store detailed information about data sources, used processing software and version, and the processing step. In order to expose information about stored data in a standardized way on the Web the pycsw server is used. Therefore, public accessible metadata datasets in RBIS are mapped to ISO 19115 and reimported in pycsw.

5 Conclusion & Future work

We showed how provenance information can be used in SDIs to show relations between datasets (and coupled models) and among research projects. Further, we utilized an interdisciplinary research program, in which basic data and simulation model outputs are shared, as underlying use case to show how provenance across projects can be implemented in linked SDIs.

Future work will address the transformation of provenance metadata into linked data and the enhancement with (linked) vocabulary (e.g. FAO's controlled thematic vocabulary *AGROVOC*) to better supply the demand for reproducible algorithms (in particular simulation models).

[9] http://leutra.geogr.uni-jena.de/vgtbRBIS
[10] http://pycsw.org

Open issues for future research demand metadata standard adjustments to meet requirements of modelling simulation model descriptions. Common SDI metadata standards do not provide elements to map meaningful parameters such as scenarios or drivers. Further, simulation models are often slightly adapted to meet a certain use case. Implementing simulation model inheritance in metadata standards would facilitate metadata acquisition by enabling users to reference (and reuse) existing model descriptions and focus on describing adaptions.

References

[1] L. He, P. Yue. Adding Geospatial Data Provenance into SDI – A Service-Oriented Approach. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. Vol.8, No.2, 2015.

[2] C. Henzen, S. Mäs, L. Bernard. Provenance Information in Geodata Infrastructures. Vandenbroucke, Danny (Ed.); Bucher, Bénédicte (Ed.); Crompvoets, Joep (Ed.), *Geographic Information Science at the Heart of Europe*. Lecture Notes in Geoinformation and Cartography. pp.133–151. doi:10.1007/978-3-319-00615-4_8, 2013.

[3] L. Di, P. Yue. Provenance in Earth Science Cyberinfrastructure; A White Paper for NSF EarthCube, 2011.

[4] B. Glavic, K. Dittrich. Data provenance: A categorization of existing approaches. BTW'07, pp.227-241, 2007.

[5] L. Osterweil, L. Clarke, A. Ellison, E. Boose, R. Podorozhny, A. Wise. Clear and Precise Specification of Ecological Data Management Processes and Dataset Provenance; IEEE Transactions on Automation Science and Engineering, Vol.7, No.1, 2010.

[6] C.A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos M, P. Li, D. De Roure. myExperiment: A Repository and Social Network for the Sharing of Bioinformatics Workflows. Nucleic Acids Research 38: W677-W682. doi:10.1093/nar/gkq429, 2010.

[7] C. Henzen, S. Mäs, M. Müller, L. Bernard, H. Tressel, S. Haase. GeoMetaFacet 2.0 - Interaktive nutzerfreundliche Visualisierung von geograpischen Metadaten. Geoinformatik, Hamburg, 2014.

[8] C. Henzen, S. Mäs, L. Bernard. Provenance Information in Geodata Infrastructures. D.Vandenbroucke, (Ed.), B. Bucher, (Ed.), J. Crompvoets (Ed.), Geographic Information Science at the Heart of Europe, 2013. Lecture Notes in Geoinformation and Cartography. pp. 133–151. doi:10.1007/978-3-319-00615-4_8, 2013.

[9] L. Bernard, S. Mäs, M. Müller, C. Henzen, J. Brauner. Scientific Geodata Infrastructures: Challenges, Approaches and Directions. In: International Journal of Digital Earth, doi:10.1080/17538947.2013.781244, 2013.

[9] OpenGIS Catalogue Services Specification, Version 2.0.2, OGC 07-006r1, 2007.

[10] C. Rumbaur, N. Thevs, M. Disse, M. Ahlheim, A. Brieden, B. Cyffka, D. Duethmann, T. Feike, O. Frör, P. Gärtner, Ü. Halik, J. Hill, M. Hinnenthal, P. Keilholz, B. Kleinschmit, V. Krysanova, M. Kuba, S. Mader, C. Menz, H. Othmanli, S. Pelz, M. Schroeder, T. F. Siew, V. Stender, K. Stahr, F. M. Thomas, M. Welp, M. Wortmann, X. Zhao, X. Chen, T. Jiang, J. Luo, H. Yimit, R. Yu, X. Zhang, C. Zhao.
Sustainable management of river oases along the Tarim River (SuMaRiO) in Northwest China under conditions of climate change. Earth Syst. Dynam., 6, pp.83–107, doi:10.5194/esd-6-83-2015, 2015.

[11] S. Kralisch, F. Zander, F. & W.-A. Flügel. OBIS - A Data and Information Management System for the Okavango Basin. J. Oldeland (Ed.), C. Erb (Ed.), M. Finck (Ed.) & N. Jürgens (Ed.), Biodiversity and Ecology 5, pp.213-220. doi:10.7809/b-e.00276, 2013.

[12] F. Zander, S. Kralisch, C. Busch, W.-A. Flügel, W.-A. (2011), Environmental data management with the River Basin Information System, in Proc. of the 19th International Congress on Modelling and Simulation, edited by Chan F., Marinova, D. and Anderssen, R-S., pp.3191-3197, 2011.

[13] International Standards Organization. International Standard ISO 19115 Geographic information – Metadata. Reference Number ISO19115:2005(E), 2005.

Appendix

Table 2: Overview of coupled simulation models

Simulation model and website	Short description
ECHAM5 (ECMWF Hamburg) http://www.mpimet.mpg.de/en/science/models/echam.html	5th generation of an atmospheric general circulation model, which forms the atmospheric component of the MPI-ESM
PROMET (PROcess of radiation Mass and Energy Transfer) http://www.geographie.uni-muenchen.de/department/fiona/forschung/projekte/promet_handbook	grid distributed, continuous, integrated land surface processes model to simulate natural processes and impacts of human interventions
DART (Dynamic Applied Regional Trade) https://www.ifw-kiel.de/academy/data-bases/dart_e	recursive-dynamic computable general equilibrium (CGE) model of the world economy to analyse international climate policies
CAPRI (Common Agricultural Policy Regionalised Impact Analysis Model) http://www.capri-model.org	global agricultural sector model to support decision making related to the common agricultural policy
RAUMIS (Regional Agricultural and Environmental Information System) https://www.ifw-kiel.de/narola/narola-modelle/raumis	to analyse impacts of alternative agricultural policies on agricultural land use, production, income and factor use in Germany on a regional scale
REMO (Regional Model) http://www.remo-rcm.de	atmospheric model, which is coupled to three different hydrology models and three ocean/sea-ice models
STAR (STAtistical Regional climate model) https://www.pik-potsdam.de/research/climate-impacts-and-vulnerabilities/models/stars	rearranges observed time series with respect to a given linear trend for a selected variable
SWIM (Soil and Water Integrated Model) https://www.pik-potsdam.de/research/climate-impacts-and-vulnerabilities/models/swim	to investigate climate and land use change impacts at the regional scale, where the impacts are manifested and adaptation measures take place
WRF (Weather Research & Forecasting Model) http://www.wrf-model.org	mesoscale numerical weather prediction system designed for both atmospheric research and operational forecasting needs
JAMS/J2K http://jams.uni-jena.de	distributed and process oriented distributed hydrological model for hydrological simulations of meso-and macro-scale catchments
MIKE FLOOD, 11, HYDRO Basin https://www.mikepoweredbydhi.com	global water simulation models focusing flood resp. coast and sea