Provenance Information in Geodata Infrastructures

Christin Henzen, Stephan Mäs and Lars Bernard

Abstract When it comes to usability evaluation of geodata information about its provenance or lineage are vital. Nevertheless, in practice the corresponding metadata elements are often neglected. Even if available, the tabular or listed metadata representations in current metadata catalogue user interfaces do not sufficiently support the users browsing and comparing metadata. This chapter proposes an interactive application for data provenance visualization called MetaViz. As a foundation for the MetaViz design the chapter provides a detailed analysis on modeling aspects, available standards and specifications for data provenance and presents possible design and implementation choices. A scientific geodata infrastructure that supports researchers sharing results of numerical simulations of different environmental phenomena serves as the underlying use case.

1 Introduction

In geodata infrastructures (GDI) metadata is meant to support (1) discovery, (2) evaluation and (3) integration of heterogeneous geodata sources. However, most of today's geocatalogue and geoportal developments primarily focus only on discovery aspects. Once also data evaluation gets into the focus data provenance becomes of major interest: Learning about a dataset's history, its origin, its

C. Henzen (✉) · S. Mäs · L. Bernard
Professorship of Geoinformation Systems, Technische Universität Dresden Department of Geosciences, Helmholtzstraße 10, 01069 Dresden, Germany
e-mail: Christin.Henzen@tu-dresden.de

S. Mäs
e-mail: Stephan.Maes@tu-dresden.de

L. Bernard
e-mail: Lars.Bernard@tu-dresden.de

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_8,

previous treatments and potentially experiences in using it are crucial aspects in assessing whether and how a considered dataset might fit for an application (Di and Yue 2011; Simmhan et al. 2005; Moreau 2010). Besides supporting usability assessments information about data provenance facilitates transparency, maintenance documentation and might even ensure reproducibility (Di and Yue 2011; Glavic and Dittrich 2007; Osterweil et al. 2010).

In current geoinformation metadata standards (ISO 19115, INSPIRE) provenance descriptions are defined using elements for a textual lineage description and if more detailed by providing references and free text documentations of data sources and data creation processes. There is not only a lack in harmonised vocabularies on describing data provenance; it also shows in practice, that creation of these metadata elements is often neglected.

Yet another issue in supporting metadata based evaluation of geodata is the way metadata is presented in geocatalogues. Problems such as the absence of customizable detail levels as well as the lack of effective communication methods for metadata contents are obvious (cp. Bowers 2012; Malaverri et al. 2012; Kindermann et al. 2007). Further, geocatalogues and geoportals mostly do not offer suitable and compact representations of the evaluable metadata. Metadata is typically presented in user interfaces consisting of long lists or tables that do not support user-friendly navigation, browsing or guidance through the metadata or even interactive analysis and comparison of metadata sets (cp. Fisher et al. 2009; Bowers 2012; Zargar 2009). Convincing (visual) inspection tools, showing how to make use of provenance information for geodata and thus motivating the provision of such metadata can hardly be found.

Focussing on data provenance this chapter proposes an interactive application for metadata visualization called MetaViz. The presented solutions and scenarios stem from the development of a Scientific GDI to support researchers in sharing input and results of numerical simulations of different environmental phenomena. The chapter provides a detailed analysis on the current state in describing data provenance as a basis for the design of MetaViz. Then implementation details and functionality of MetaViz and its integration in the GDI environment are presented. A discussion of the achieved results will help to identify future research and development needs.

2 Aspects of Provenance

Generally provenance metadata informs the user about the history of a data product, source data and processes (Moreau 2010; Simmhan 2005; Di and Yue 2011). It offers the possibility to document the data origin by references but without a need to publish all interim results or any potentially access restricted input data itself. The term provenance is often used synonymously with the terms lineage or pedigree. Additional data usage documentations describe concrete applications of the data (e.g. visualizations or analysis) or link to further processed data products.

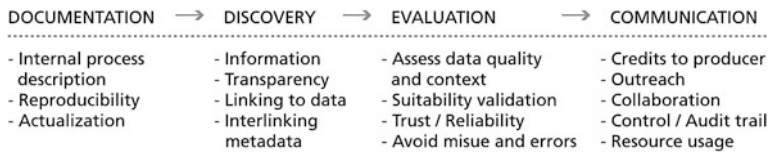

Fig. 1 Purposes of provenance information

The latter can be seen as a different view on lineage information focusing on the derived data products instead of the data history. Therefore, usage information can be deduced from provenance information and is typically described in the same schema. In this chapter usage is conceived as derivation of data products, leaving out the concrete applications of data.

The main purposes of provenance information can be categorized into documentation, discovery, evaluation and communication (Fig. 1). These categories also mirror the process steps from metadata acquisition to usage and communication of results. From the data producers perspective provenance information documents internal processes and might even facilitate the reproducibility. Discovery and evaluation correspond to the general purposes of metadata, that is enabling the user to find data and to assess whether the dataset suits the requirements of an application or not.

One well-known issue in assessing data is trust. Naturally trust strongly relates to data producers and their trustworthiness or reputation (Malaverri et al. 2012; Di and Yue 2011). Furthermore, trust can be raised by applying more formal methods such as data validation or by enabling a reproduction of data product on providing all required provenance information. Additional provenance data also helps to avoid misuse or misinterpretation of data and thus ideally helps in preventing from incorrect data usages (Devillers et al. 2005; Malaverri et al. 2012).

Another important purpose of provenance is the communication of credits to contributors of source data, and derivation algorithms or processes, outreach documentation and indicator in audit trails. The derivation of the usage is not only interesting for further data users and controller but also for the data producer.

The following subchapters provide an in-depth analysis of the different characteristics in modelling, presenting and standardizing provenance data. These characteristics will then be used to (1) to classify related work and (2) to describe the design and implementation of MetaViz.

2.1 Modelling Provenance

Describing provenance of information resources is a well-researched topic that can be examined from different perspectives such as several modelling or classification concepts, architectures and standards as well as user requirements and interfaces.

Table 1 Modeling aspects of provenance

Subjects/entities	Data		Processes	
Granularity/level of detail	Coarse (e.g. dataset level)	Fine (class or attribute level)	Coarse (e.g. only one process between datasets, without sub processes)	Fine (e.g. detailed workflows with sub processes)
Representation	Directed to provenance	Directed to usage	Bidirectional	
	Successive process steps		Complete process sequence	

Due to the application domain and specific user requirements provenance information can be modelled data- or process-oriented. Data provenance describes the history of a data product on a fine-grained level using classes and relationships (Table 1). Spéry et al. (2001) developed such a fine-grained data provenance model that is used to capture manipulations on the feature-level of spatio-temporal data. Vert et al. (2002) and Pastorello et al. (2005) analyse web-based file and document management of GIS data and define coarse-grained models for data provenance that use the document as highest granularity stored together with adapted FDGC metadata in a database or managed via services.

Process provenance, sometimes called service provenance, focuses on detailed information about the workflow and corresponding sub processes facilitating the reproduction (cp. Osterweil et al. 2010). In some cases data provenance can be deduced from process provenance, e.g. by omitting the process information and only showing data derivations (cp. Simmhan et al. 2005).

It is also possible to derive a coarse-grained provenance model from a fine-grained one, which induces different views of the data model Visualizing provenance information on different levels of granularity allows the user to get a brief summary of provenance or a very detailed view, for example on attribute level. Provenance can generally be represented either as separate successive processes or as the complete sequence of all processes:

- If represented in separated process steps each entity only contains information about the processing of the prior dataset (direct predecessor). Thus lineage information is stored step-wise in several linked metadata sets.
- Provenance can be modelled as complete provenance with fine or coarse granularity. In contrast to the successive provenance representation, the complete provenance contains information about the whole lineage process. Within a chain of processed data, provenance descriptions are stored redundant in the metadata of the data and in the metadata of its successor.

Another aspect of provenance modelling is the representation of direction or navigation links: Some provenance models only provide backward links to origin processes and source data. Others focus on usage and link only to derived data and the respective processes. Usually the missing direction can be deduced. In a

Table 2 Options for system design of provenance GUI

System implementation and architecture			
Application domain	Web	Desktop	
Storage	Tightly coupled with data	As part of the metadata	Separated storage systems
Data interchange	Standard interface	Proprietary interface	
Infrastructure	Distributed environment (service-based)	Standalone application	

bidirectional representation no further processing is needed, but the metadata storage might be redundant.

2.2 Exploring Provenance Data

Evaluating the fitness for use with the help of provenance information does not only depend on the underlying metadata model but also on the information representation techniques. The basic options for the design of a GUI representing provenance data can be discriminated in being either visual representations or being part of the query structures. Provenance visualizations such as trees or directed acyclic graphs are often used to illustrate processing workflows (Anand et al. 2010; Cheung and Hunter 2006) or linked data whereas textual descriptions are typically used in metadata catalogue systems such as GeoNetwork.[1] In such systems queries are usually formulated in textual form, like search terms or keywords. Other more formal querying methods in provenance user interfaces are textual as well as graphical query languages, such as Query Language for Provenance (QLP) (Anand et al. 2010) or Little-JIL (Cass et al. 2000).

When analysing approaches based on the implementation characteristic distinctions are the provenance storage, interchange format and infrastructure. Storage of provenance data can be either coupled with data or with metadata holdings (Di and Yue 2011) (Table 2). The latter can for example be done in geocatalogues following the Catalogue Service for the Web (CSW) interface (OGC 2007) to provide standardized access to geometadata storage systems.

Managing provenance in service-oriented architectures is a pressing challenge (Wang et al. 2008; Yue et al. 2011; Kindermann et al. 2007; Di and Yue 2011) and approached in different ways. Wang et al. developed a three-tier architecture with a web service layer that handles storing, searching and browsing requests, a logic layer and a repository layer that contains the spatial data store and the separated semantic repository. Yue et al. (2011) extended a geospatial metadata catalogue to manage data and service provenance. Automatic metadata generation during data

[1] http://geonetwork-opensource.org

production or process execution, actualization and exchange has been discussed as an additional aspect of provenance (Di and Yue 2011).

Further approaches on geoinformation provenance management such as GeoOpera, GOOSE, ESSW and Geolineus are reviewed by Bose and Frew (2005). Research in provenance of geoinformation mostly addresses modelling and implementation aspects of workflow management or origin and processing of sensor observations. Only a few approaches also address the graphical representation of provenance. Lanter (1991), for instance, introduced a graphical language and user interface for layer-based geographic data. The interface displays an interactive flow diagram and focusses data provenance but does not address the processing steps. Current investigations on provenance of geodata do either focus on interactive visualizations or on using provenance standards.

Outside the geoinformation domain several visualisation methods for provenance data can be found. There are provenance clients such as Provenance Browser (Anand et al. 2010) or Provenance Explorer (Cheung and Hunter 2006) as well as graphical languages, e.g. Little-JIL (Fig. 2a) (Cass et al. 2000). These approaches suggest context-sensitive provenance views, as a data-dependency view or an invocation graph and also present different granularity levels for provenance graphs.

2.3 Standards and Specifications

A provenance information model should follow a standard, but should also be tailored to the specific context (Malaverri et al. 2012; Di and Yue 2011). Geosciences mostly lack appropriate provenance metadata as well as suitable standards (Tilmes 2008; Di and Yue 2011; Yue et al. 2011).

A generic and thematically independent provenance model is the Open Provenance Model (OPM) (Moreau et al. 2011), which specifies a provenance model in a technology-agnostic manner. The three basic elements of this specification are (1) artefacts that describe entities e.g. datasets, (2) agents as a kind of controlling units and (3) processes. Moreover OPM realizes a role as well as a view concept dealing with hierarchical and overlapping accounts.

Another commonly used provenance model is the qualified Dublin Core.[2] Being very compact and producer focussing it is typically used to model web resources. Standards used to describe lineage information in GDI are ISO 19115 part 2 (ISO 2005) and the FGDC Content Standard for Digital Geospatial Metadata (CSDGM).[3] Both define basically the same entities to describe data provenance and only differ in naming conventions. While some researchers argue about too static views and technical names (Fisher et al. 2009; Zargar 2009) these standards

[2] http://dublincore.org/documents/dcmi-terms/

[3] http://www.fgdc.gov/metadata/geospatial-metadata-standards#csdgm

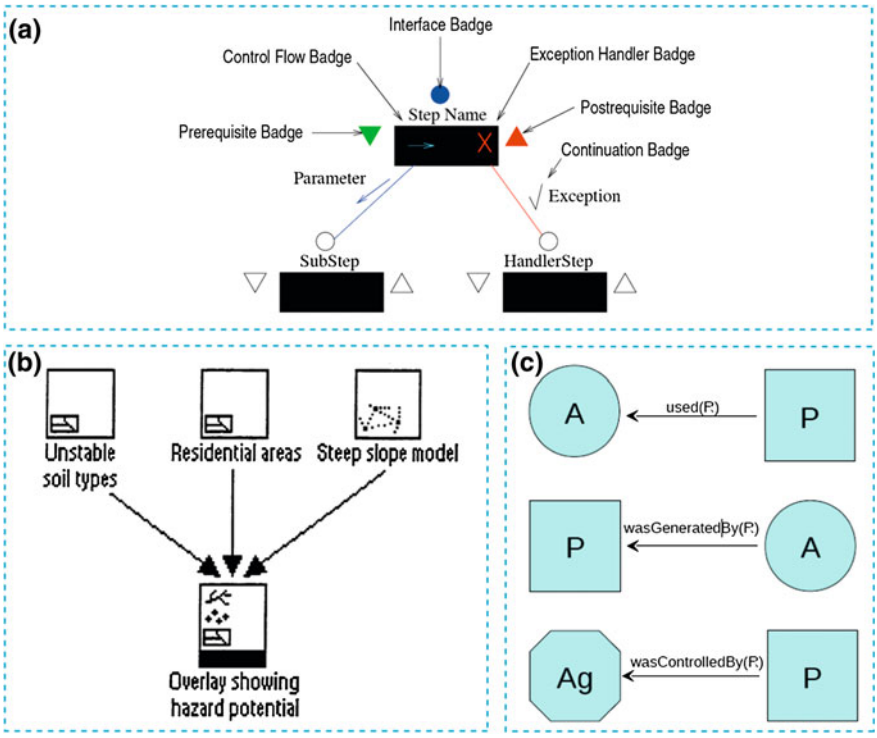

Fig. 2 Provenance visualizations. **a** Legend of the language Little-JIL (Cass et al. 2000). **b** Flow diagram with intermediate and product layer (Lanter 1991). **c** Provenance graph in OPM-Notation (Moreau 2010)

are often applied and need only a few adaptations to support the concepts of OPM, because they are defined more precisely focussing on data provenance of geodata.

All specifications, apart from Dublin Core, represent provenance information at fine or coarse granularity (Table 3)[4] and allow different views on the provenance.

To harmonize these specifications their main elements have been identified (Table 4). Thus, a process step is described by inputs and outputs, a process or model description and the data producer. It is sometimes controlled or facilitated by agents (Moreau et al. 2011) and further explained in (separate) documentations. The specifications do not always allow a one-to-one mapping (Nogueras-Iso et al. 2005), but a transformation between them is possible.

Sometimes specifications get combined: Malaverri et al. (2012) introduce a coarse-grained data provenance model based on a combination of OPM and FGDC illustrated by a use case on map generation. They analyse several quality indicators, such as timeliness of data or reputation of data provider with regard to trustworthiness.

[4] http://dublincore.org/documents/dcmi-terms/#terms-provenance

Table 3 Data provenance concepts in metadata and provenance standards

Provenance standard/ specification	ISO 19115-2 Lineage subclasses	FGDC CSDGM	OPM	Qualified Dublin Core
Concept definition	"Specify lineage of imagery and gridded data datasets" (ISO2005)	"Description of the source material […] and the methods of derivation […] (for the) digital files" (FGDC 2000)	"Represent provenance for "any" thing" (Moreau et al. 2011)	"A statement of any changes in ownership and custody of the resource […]"
Subjects	Data provenance	Data provenance	Data or process provenance	Data provenance
Granularity/ level of detail	Mainly coarse (dataset level)	Mainly coarse (dataset level)	Fine or coarse	Coarse
Representation	Directed to provenance	Directed to provenance	Directed to provenance or usage or bidirectional	Directed to provenance

Table 4 Mapping of provenance elements of metadata and provenance standards

Provenance element	ISO 19115-2 Lineage subclasses	FGDC CSDGM	OPM	Qualified Dublin Core
Input	Source	Source information	Artifact	Source
Output	Source	Source information	Artifact	Source
Process/Model	Process step	Process step	Process	Provenance
Agents	Source, Process step	Source information, Process step	Agent	Provenance
Documentation of processes	Documentation	Source used citation abbreviation	Annotation	Provenance
Data producer	Processor	Process Contact	Agent	Creator, Contributor, Publisher

3 MetaViz: An Interactive Provenance Visualization Client

The need for an application visualizing data provenance arose during the GLUES Project, which is the coordination project of the international interdisciplinary research program 'Sustainable Land Management'[5] of the German Ministry of Education and Research. Within this funding measure several so called regional collaborative projects are researching the impacts of climate and socio-economic changes and a corresponding optimization of the use of land and natural resources

[5] http://modul-a.nachhaltiges-landmanagement.de

in different regions. Since this interdisciplinary research is policy-oriented the projects cooperate with regional scientists and stakeholders. The major aims of GLUES are to support 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 (Eppink et al. 2012; Mäs et al. 2011).

Technically, the access to the results of GLUES and the regional projects will be provided by means of a scientific GDI.[6] The provided data can be used by other scientists as input into their simulation models. To avoid misinterpretations and misapplication of data a major focus of the GDI implementation is on acquisition and representation of meaningful metadata and, in particular, provenance data. Due to the complexity and the high amount of metadata, a visual illustration of the interrelationships between different datasets and models is essential. Therewith, scientists can, for example, get a comprehensive view which models provide data for a certain scenario or whether an input data also served into other models. Beside the scientific work, such provenance visualization can also be of interest for research assessment and outreach analysis, since it represents the data exchange and collaboration between different research institutions.

There are several restrictions on provenance modelling and visualization based on the properties of the data and the user requirements within the project. For instance, detailed model descriptions (i.e. detailed descriptions of subprocesses) are not available in the metadata and would possibly not be feasible due to the models' complexity.

However, the importance of linking a model and its scientific publications has been pointed out on a GLUES workshop on models and consistent datasets. The following list characterizes models and data in the GLUES context:

- A model is represented as a single process step
- A model is described by a short summary, several scientific publications and a reference to the modeler or scientific institution
- A model can have several inputs and outputs, but is not directly connected to another model
- A dataset can be input of one model and simultaneously output of another model
- Pre-processing steps, such as cleaning the data, will only be collected as textual description of a process, but not as further process steps
- The provenance of a model is not considered.

The users in this research project are as wide-ranging as its thematic fields. However, the scientific community is just one of the four identified user groups, namely policy makers, stakeholder, society and scientist. The main objective of all user groups is the discovery of data to get a general overview of existing data or to find relevant data for a specific problem (Table 5). Data provenance information can support the assessment of the data quality and evaluation of the fitness for use.

[6] http://geoportal.glues.geo.tu-dresden.de

Table 5 User groups and their objectives within the research project

User groups	Objectives and activities in the project
Scientific Community	(Internal) documentation
	Discover relevant data
	Communicate (scientific) results
	Evaluate data
	Use results
Policy maker	Get overview
	Communicate
	Transfer (scientific) results
Stakeholder	Get information
	Implement results
	Communicate
	Transfer results
Society	Get general information

Descriptions of numerical models and their output data are complex and a quality evaluation requires detailed knowledge about the model and its initialisation, basic assumptions and research goals. For scientists having this background knowledge the provenance visualisation shows dependencies among data sets and it can indicate how model assumptions, restrictions and even errors propagate. Further, data provenance illustrates the data exchange and collaboration between the scientists.

The technical basis for this interdisciplinary work is a geoportal as an entry point to the GDI that provides a common metadata pool for the documentation of global long and midterm scenarios, its resulting datasets and synthesis results. The integrated metadata catalogue supports for the manual or scripted acquisition of ISO 19115 metadata including lineage information about source data and processes.

Figure 3 shows an extract of the catalogue user interface. It displays a part of the lineage information for a dataset that is generated by a model, named CAPRI. The user interface shows information about the model, such as a documentation link and brief summary. The model has at least two input sources, listed below the model information. Although this extract does not contain all lineage information, it shows that the table-like and non-interactive information representation is complex and difficult to apprehend for users.

3.1 Requirements for an Interactive Provenance User Interface

The interpretation of provenance information gets a significant support through the design of an easy-to-use and comprehensive user interface. Design dimensions are the provenance information model, information and interaction design as well as

Process information:	**Identifier:**	CAPRI	
	Software reference:		
	Procedure description:	The CAPRI model is a comparative static global partial equilibrium model for the agricultural sector. It endogenously determins market balances, area use and yields and many other variables for agricultural raw products and a number of processed products. CAPRI has been developed within EU Framework projects and is been used widely for policy impact analysis.	
	Documentation:	**Title:**	CAPRI model documentation
		Date:	2011-01-01
		Datatype:	publication
		Identifier:	
		Other citation details:	http://www.capri-model.org /docs/capri_documentation_2011.pdf
Source:	**Description:**	FAOSTAT data	
	Source citation:	**Title:**	FAOSTAT data
		Date:	2012-08-07
		Datatype:	publication
		Identifier:	faostatdatabasedomains (Codespace: urn:glues-ext:fao:metadata:dataset)
		Other citation details:	
	Description:	AGLINK-COSIMO (OECD,FAO)	
	Source citation:	**Title:**	AGLINK-COSIMO (OECD,FAO)
		Date:	2012-08-07
		Datatype:	publication
		Identifier:	oecdfaoagriculturaloutlook209-2018 (Codespace: urn:glues-

Fig. 3 Extract of the provenance information of a metadata record in the catalogue

the technical design. Although the successive data provenance (Table 1) information model based on ISO 19115 is quite static it is used here to support the integration in the existing GDI environment and connected to the Open Geospatial Consortium Web Catalogue Service (CSW) services and the existing metadata. The information design shall enable to answer the following questions, which summarize user requirements as defined by Kunde et al. (2008) and adapted to the introduced use case:

- Which data was used for the generation of a dataset?
- Which data was generated using a given dataset?
- Which actors (organizations, tools…) have been involved?
- Which resources from other models have been used in the generation of a dataset?
- In which stage of a processing chain is a given dataset?
- Did the model the dataset is part of reach a satisfactory conclusion by some given regulations or criteria?

Therefore the user interface should support an objective data representation of who, what, why, when and how the data is generated. The way of representing this information should be efficient using visualizations instead of long textual descriptions. The visualization should also display relationships and qualify the

user to position the dataset or model in space and time (Edwards et al. 2010; Bowers 2012; Zargar 2009).

The views should be adaptive and allow the typical scientific iterative data exploration (cp. Wang et al. 2008). In addition, all user groups require efficient and user-friendly navigations through lineage and usage information and dataset hierarchies as well as linked context-sensitive data representations such as the visualization of data in a map client. The information should be presented in a way understandable for users who are not familiar with the application and are going to use it rather seldom such as citizens or politicians. At the same time the presentation has to fit the purposes of scientists searching for detailed model or data descriptions and the corresponding publications. Conclusively, the interaction design should allow the navigation among related metadata, datasets or its visualization (Di and Yue 2011) and supports guidance through the data instead of complex querying.

Technically, the logic of the user interface has to offer possibilities to request and process ISO compliant metadata. This indicates that the application has to deal with the successive provenance steps of ISO 19115-2 deducing dataset inheritance and hierarchies based on scripted ID-matching. As shown in Table 4, a mapping from ISO to other specifications, especially FGDC, can be made easily to use the user interface with different data schemes.

Finally, a brokering mechanism that generates parameterized application links such as a link to a metadata's detail page of the catalogue has to be included.

3.2 Architecture and Implementation

The application MetaViz[7] is an interactive web-client consisting of a Java-based backend that contains the application logic and a user interface realized with HTML and JavaScript. Since MetaViz uses the standardized CSW 2.0.2 interface it does not need further storage systems, but directly requests the configured metadata catalogue for the ISO 19115 compliant metadata with provenance information (Table 6). The catalogue's response is processed and transformed into an intern and condensed JSON model with respect to the specific requirements of provenance visualization, such as sorting and pre-selection of required lineage and usage information.

Pre-processing of metadata is a computing expensive process. This is particularly because the ISO metadata schema does not directly fit the necessary requests that answer the user's requirements:

[7] A lineage example for the dataset PROMET shown in MetaViz application: http://geoportal.glues.geo.tu-dresden.de:8080/MetaViz/detail.jsp?id=glues:lmu:metadata:dataset:promet

Information about MetaViz is summarized in a factsheet: http://geoportal.glues.geo.tu-dresden.de/geoportal/documents/Fact_Sheet_MetaViz.pdf

Table 6 Characteristics of MetaViz

Property	Realization in MetaViz
Subject/entities	Data
Granularity/Level of detail	Coarse
Representation	Bidirectional
	Successive provenance steps
Visual representation	Graph
Querying	Visual query interface
Application domain	Web
Storage	Tightly coupled with metadata
Interfaces and Data Formats	Standard interface using ISO 19115-2, CSW 2.0.2
Infrastructure	Distributed environment (service-based)

- Lineage and usage information
- Parent–child-relations between datasets and data series
- Connected view services (or other services supporting further exploration)

ISO metadata stores lineage as sequential provenance steps instead of a complete provenance graph in one metadata entry. Due to this, several catalogue requests have to be made to compose the whole lineage of a dataset. Furthermore, usage information has to be deduced from the lineage descriptions, as the 'usage of a dataset' is only stored as a lineage of another dataset. Usage is considered here only in terms of processes that lead to new data products, leaving out direct applications like data visualization).

Parent–child-relations are also not stored bidirectional: metadata sets contain links to parents, but not to children (cp. Nogueras-Iso et al. 2005, p. 37). Links of datasets and their connected view services are likewise stored within the metadata of the service instead of the metadata for the dataset. Thus all data offered by the CSW has to be analyzed to get the children, the usage or the linked view services of a dataset. This pre-processing is quite computing expensive. To increase the query performance MetaViz can be switched between a direct database mode or a CSW mode. In the database mode the metadata is requested directly from the underlying database of a metadata catalogue, resulting in much better response times. Using the standardized CSW interface in the CSW mode lacks in performance but allows more flexibility in being less tightly coupled to the database scheme of the used catalogue.

As illustrated in Fig. 4, MetaViz can be linked with other clients being used in a GDI like the geocatalogue GUI or geovisualization clients. This allows for a continuous user interaction. MetaViz is not only requesting data from a metadata catalogue but also linking back to a catalogue's detail page, which shows the entire list of metadata elements in the traditional manner. Furthermore the application is coupled with a map client to visualize the data if the metadata contains a reference to a Web Map Service (WMS). By calling MetaViz parameterized with a dataset id, it can be embedded into other websites or applications.

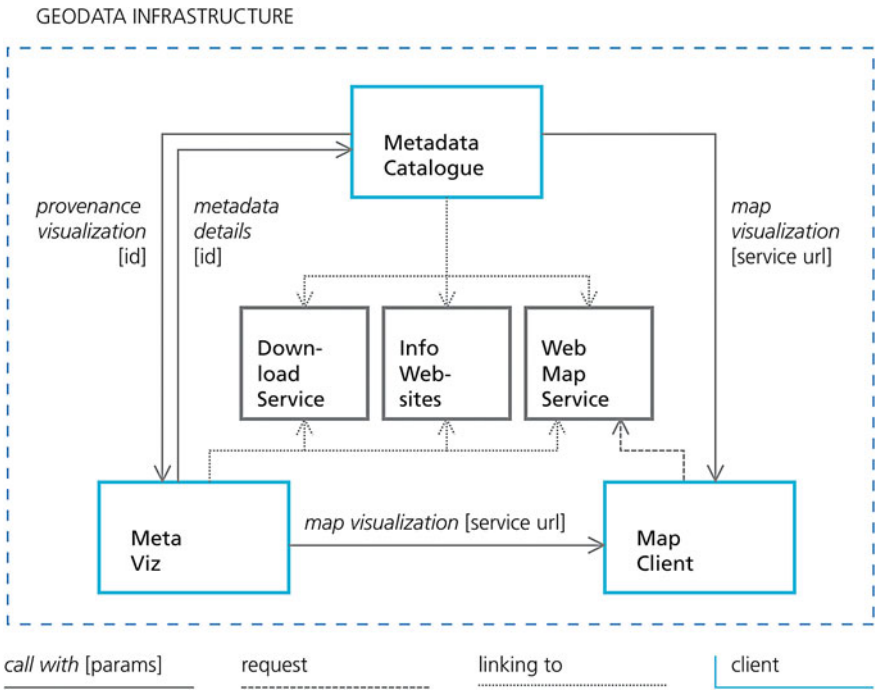

Fig. 4 Linkage and information exchange between MetaViz and other GDI clients

3.3 User Interface

MetaViz focuses on a user-friendly and compact visual description of lineage and usage of datasets within a GDI. The main element of the application is a tree-like interactive lineage and usage graph (Fig. 5)[8] showing the provenance of a dataset with its name displayed above the graph on the left. Next to the graph some general information such as name, temporal and spatial extent, tagged keywords and (interactive) relations to parent or child datasets are listed.

Below the graph extended provenance information is displayed. Process descriptions and publications are separated visually to arrange information in a well-structured and easy-readable way. The process description contains free-text about rationale of the process step and process parameters such as software reference, processor and time of process execution. Pre studies with scientists in the GLUES project lead to a design which does not display all elements of the ISO 19115-2, such as detailed runtime parameters, to keep a rather simple and quickly to grasp user interface.

[8] The example shown in the screenshot is available in the web: http://geoportal.glues.geo.tu-dresden.de:8080/MetaViz/detail.jsp?id=f872b5b8-bb23-4df5-a906-0b396c99cc22

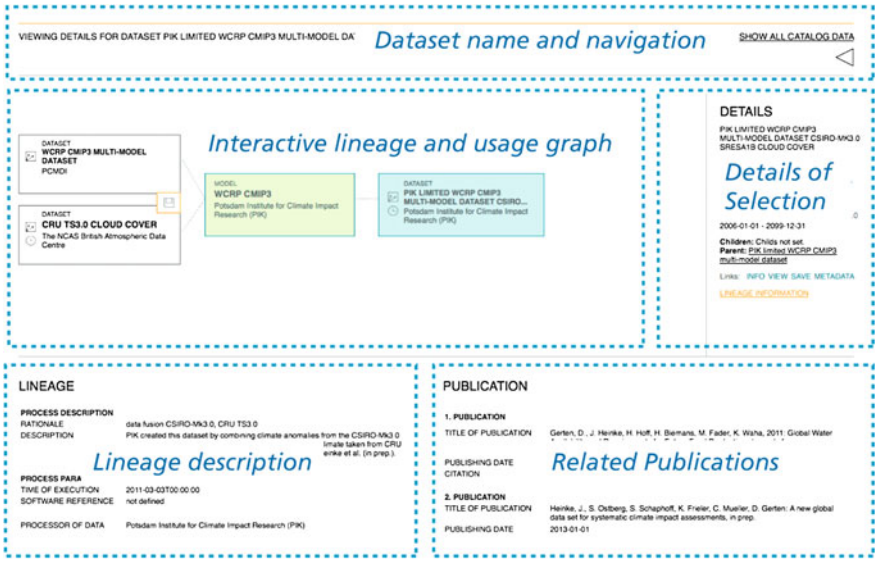

Fig. 5 Overview of areas in MetaViz application GUI showing the lineage of a dataset

The publication information displayed on the right lists a BibTex-like reference, the date of publication as well as a link to the publication, if available.

The lineage graph shows the dataset derivation. Explored from left to right, it shows lineage information on the left and usage information, if available, on the right. It connects the focused dataset (blue box) to the process (green box) where it originates from as well as the source datasets (white boxes) of these processes.

Each graph contains a maximum of one lineage and one usage step to keep the presentations comprehensible. It is possible to focus a presentation either on lineage (Fig. 6)[9] or on usage (Fig. 7).

The application does not only visualize the lineage as a graph but also displays relevant information to assess data quality. This information like data provider, data type or time-variant are displayed in the visualization as texts or symbolized via icons (Fig. 8)[10]. All icons are explained with short tooltips texts to facilitate the application handling.

The general navigation concept of the application is as simple as the visualization. The users do not have to formulate complex queries. Navigating through the lineage graph or to the context-sensitively linked applications is done by clicking on an icon, link or button.

[9] The example shown in the screenshot is available in the web: http://geoportal.glues.geo.tu-dresden.de:8080/MetaViz/detail.jsp?id=glues:lmu:metadata:dataset:promet

[10] The example shown in the screenshot is available in the web: http://geoportal.glues.geo.tu-dresden.de:8080/MetaViz/detail.jsp?id=glues:pik:metadata:dataset:csiro-mk3.0sresa1bcloudcover

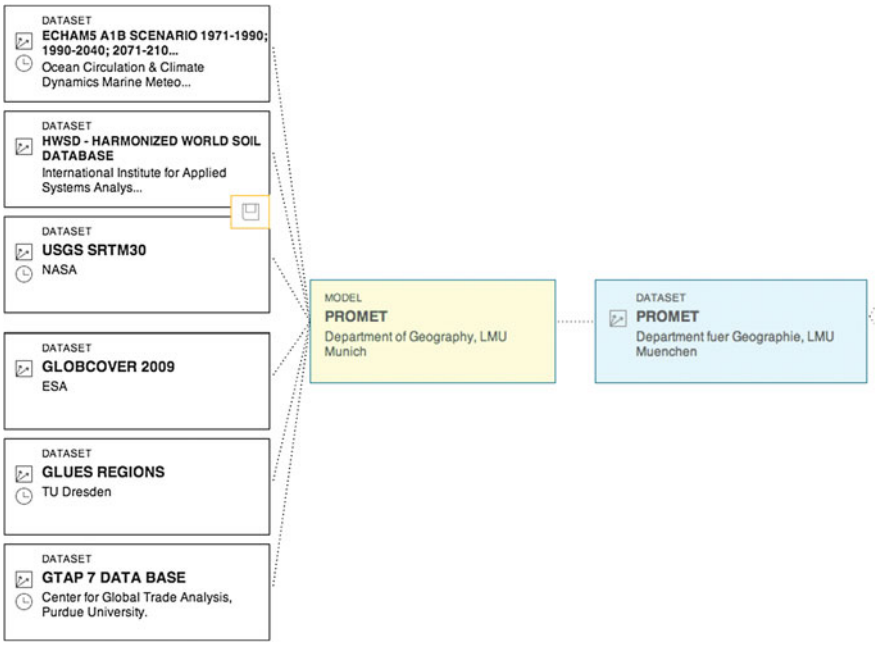

Fig. 6 Lineage graph of the dataset PROMET visualized in MetaViz

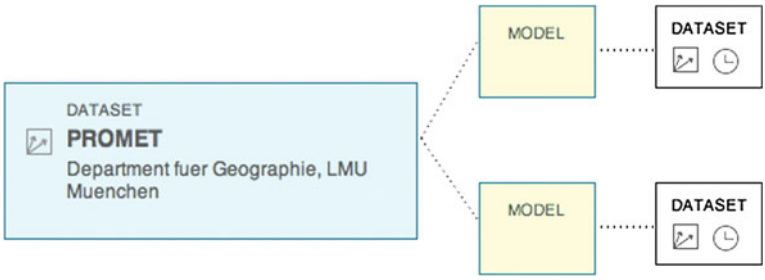

Fig. 7 Simplified usage graph of the dataset PROMET

MetaViz can be used as standalone application or integrated in other website with a parameterized call as well as called from the catalogue GUI as one view of the standardized metadata. This offers the different user groups, such as the data modellers, the possibility to integrate the application in their own research website, link to it from scientific publications or use a screenshot of the graphical presentation in their publication.

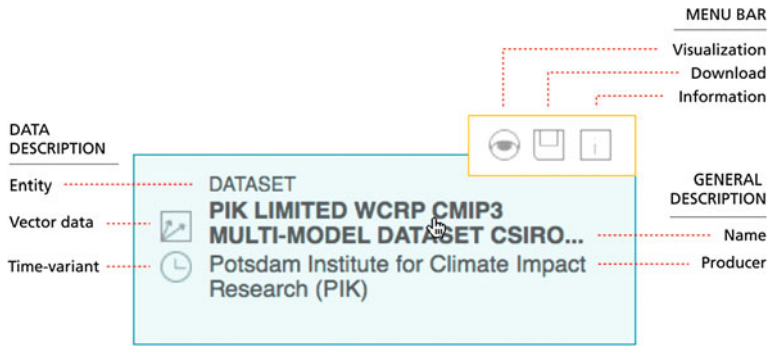

Fig. 8 Visual representation of dataset in MetaViz and context-sensitive menu

4 Conclusions and Future Work

Management and representation of provenance information in GDI has not gained much attention so far. The introduced provenance visualization client illustrates how the presentation of metadata in GDI can be enriched by interactive, intuitive and user-friendly interfaces. Such visualization supports the communication of data quality, enhances interpretation and prevents misinterpretation or misapplication of geodata. Moreover, it is felt that intuitive and convincing metadata applications—as intended with MetaViz—can further stimulate the willingness to generate and maintain metadata. The representation of the processing steps is understandable, even for non-expert users. In scientific GDI usage information can play a major role for the evaluation of scientific outputs, comparable to the way citations are used to rank scientific publications.

The current metadata standards, including ISO 19115, do not entirely fit the requirements of our use case. The description of numerical models and their output data would require data elements explicitly representing information about model initialisation, scenarios, drivers and basic assumptions of the model. In particular scenarios, that define a projection of a potential future based on a coherent set of assumptions (Nakićenović et al. 2000), would be useful to classify and compare datasets. Statistical analyses, like in spatial econometrics, require provenance information for analysis workflows and the applied (spatial) weights (Anselin and Rey 2012). At least for scientists working with these data such information is indispensable for evaluation.

Another issue are parent–child-relations among datasets, which are typical for the machine generated data in our use case. Although very useful for structuring data, these relations are hardly represented and navigable in current GDI catalogue user interfaces. In particular for datasets with a high number of child datasets it is not sufficient to store the relation without an explicit description of the concrete commonalities and differences of the sub datasets (Nogueras-Iso et al. 2005). The navigation and illustration of these relations is a possible future extension of MetaViz.

Scientists, but also data producers in general, are clearly not passionate in collecting detailed metadata descriptions. Therefore, the application MetaViz requires only a minimal set of lineage attributes, being evaluated and approved by domain modelling experts within the GLUES project, and integrates existing descriptions, such as publications. Nevertheless, automatic metadata derivation and acquisition remains a big issue for future research.

So far, MetaViz considers usage only in terms of processes that lead to new data products. To also include direct applications of the data (e.g. visualizations or analysis) future extensions for direct user feedback and a rating system are planned. Additionally, a usability evaluation of the current GUI shall help in improving the design. Further, the suitability of the application for other use cases, such as metadata created by web processing services will be analyzed.

References

Anand MK, Bowers S, Ludäscher B (2010) Provenance browser: displaying and querying scientific workflow provenance graphs, ICDE, 2010

Anselin L, Rey SJ (2012) Spatial econometrics in an age of CyberGIScience. Int J Geogr Inf Sci 26(12):2211–2226

Bose R, Frew J (2005) Lineage retrieval for scientific data processing: a survey. ACM Comput Surv 37(1):1–28

Bowers S (2012) Scientific workflow, provenance, and data modeling challenges and approaches. J Data Semant 1(1):19–30

Cass AG, Lerner E, McCall K, Osterweil LJ, Sutton SM, Wise E (2000) Little-JIL/Juliette: a process definition language and interpreter. International Conference on Software Engineering, 2000

Cheung K, Hunter J (2006) Provenance explorer—customized provenance views using semantic inferencing. ISWC 2006, (LNCS), vol 4273. pp 215–227

Devillers R, Bédard Y, Jeansoulin R (2005) Multidimensional management of geospatial data quality information for its dynamic use within GIS. Photogram Eng Remote Sens 71(2):205–215

Di L, Yue P (2011) Provenance in earth science cyberinfrastructure. A White Paper for NSF EarthCube, 2011

Edwards P, Pignotti E, Reid R (2010) Weaving a provenance fabric to support next generation science. IEEE internet computing special issue on provenance in web applications, 2010

Eppink F, Werntze A, Mäs S, Popp A, Seppelt R (2012) Land management and ecosystem services: how collaborative research programmes can support better policies. GAIA: Ecol Perspect Sci Soc 21(1):55–63

FGDC (2000) Content standard for digital geospatial metadata workbook (For use with FGDC-STD-001-1998), Version 2.0. Federal geographic data committee, May 1 2000

Fisher P, Comber AJ, Wadsworth R (2009) What's in a name? Semantics, standards and data quality. In: Devillers R, Goodchild H (eds) Spatial data quality: form process to decisions. CRC Press, Boca Raton, pp 3–16

Glavic B, Dittrich K (2007) Data provenance: a categorization of existing approaches, BTW'07, pp 227–241

ISO 19115-2 (2005) International standard on geographic information—Part 2: metadata for imagery and gridded data

Kindermann S, Stockhause M, Ronneberger K (2007) Intelligent data networking for the earth system science community, German e-Science

Kunde M, Bergmeyer H, Schrieber A (2008) Requirements for a provenance visualization component, Provenance and annotation of data and processes (Lecture notes in computer science), vol 5272. pp 241–252

Lanter DP (1991) User-centered graphical user interface design for GIS. National center for geographic information and analysis, report. pp 91–96

Malaverri JEG, Medeiros CB, Camargo R (2012) A provenance approach to assess quality of geospatial data. 27th symposium on applied computing (SAC)

Mäs S, Müller M, Henzen C, Bernard L (2011) Linking the outcomes of scientific research: requirements from the perspective of geosciences. Proceedings of the first international workshop on linked science 2011 (LISC2011), CEUR Workshop Proceedings, vol 783

Moreau L (2010) The foundations for provenance on the web. Found Trends Web Sci 2(2–3):99–241

Moreau L, Clifford B, Freire J, Gil Y, Groth P, Futrelle J, Kwasnikowska N, Miles S, Missier P, Myers J, Simmhan Y, Stephan E, Van den Bussche J (2011) The open provenance model—core specification (v1. 1). Future generation computer systems

Nakićenović N, Alcamo J, Davis G (2000) IPCC special report on emissions scenarios (SRES) Cambridge. Cambridge University Press, NY

Nogueras-Iso J, Zaragaza-Soria FJ, Muro-Medrano PR (2005) Geographic information metadata for spatial data infrastructures: resources, interoperability and information retrieval. Springer, Berlin 2005

OGC (2007) OpenGIS catalogue services specification. Version 2.0.2, OGC 07-006r1

Osterweil L, Clarke L, Ellison A, Boose E, Podorozhny R, Wise A (2010) Clear and precise specification of ecological data management processes and dataset provenance. IEEE Trans Autom Sci Eng 7(1):189–195

Pastorello G, Medeiros C, Resende S, Rocha H (2005) Interoperability for GIS document management in environmental planning. J Data Semanti III (LNCS), vol 3534. pp 100–124

Simmhan Y, Plale B, Gannon D (2005) A survey of data provenance in e-science. SIGMOD record 34:31–36

Spéry L, Claramunt C, Libourel T (2001) A spatio-temporal model for the manipulation of lineage metadata. Geoinformatica 5(1):51–70

Tilmes C, Fleig A (2008) Provenance tracking in an earth science data processing system. Lect Notes Comput Sci 5272:221–228

Vert G, Stock M, Jankowski P, Gessler P (2002) An architecture for the management of GIS data files. Trans GIS 6(3):259–275

Wang S, Padmanabhan A, Myers J, Tang W, Liu Y (2008) Towards provenance-aware geographic information systems, GIS. ACM, NY

Yue P, Wei Y, Di L, He L, Gong J, Zhang L (2011) Sharing geospatial provenance in a service-oriented environment. Comput Environ Urban Syst 35:333–343

Zargar A (2009) An operation-based approach to the communication of spatial data quality in GIS