

A Provenance Approach to Assess Quality of Geospatial Data

Joana E. G. Malaverri
Institute of Computing
University of Campinas -
UNICAMP
13083-852, Campinas, SP,
Brazil
jmalav09@ic.unicamp.br

Claudia Bauzer Medeiros
Institute of Computing
University of Campinas -
UNICAMP
13083-852, Campinas, SP,
Brazil
cmbm@ic.unicamp.br

Rubens Camargo
Lamparelli
University of Campinas -
UNICAMP
Center for Research in
Agriculture
13083-970, Campinas, SP,
Brazil
rubens@cpa.unicamp.br

ABSTRACT

Geographic information is present in our daily lives. This pervasiveness is also at the origin of several problems, including heterogeneity – given its widespread demand, georeferenced data are collected daily in huge volumes, by very many independent organizations and people, with varying (or even non-existing) quality and reliability criteria. This, in turn, affects the quality and trustworthiness of the results obtained from processing geographic information. Most efforts to improve this situation concentrate on establishing data collection and curation standards, and quality metadata. This paper extends these efforts by presenting an approach to assess quality of geospatial data based on provenance, i.e. the history of the origins and transformation processes applied to a given data source. This approach combines features provided by the Open Provenance Model (OPM) and geographic metadata standards. We illustrate our approach with a case study in agriculture.

Categories and Subject Descriptors

H.2 [Database Management]: Database Applications—*Spatial databases and GIS*

General Terms

MANAGEMENT

Keywords

Data Provenance, Quality, Geospatial metadata standard

1. INTRODUCTION

Scientists from several domains (e.g., environmental sciences, agriculture, social sciences) are using different kinds of techniques and methodologies to capture, produce, process and analyze a huge volume of georeferenced data. This has

given rise to demands that require new computational tools to systematically tackle the challenges posed by distributed and heterogeneous data sources, storage strategies, transfer mechanisms, among others.

Regardless of the application domain, data collected are manipulated by a wide range of users, with distinct research interests, using their own vocabularies, work methodologies, models, and sampling needs. In particular there is a huge effort to improve the means and methodologies to capture, process and disseminate geospatial data. Such data are produced applying several methods, using a variety of devices and from different sources. This information, when adequately described and documented, would help end-users to assess the trustworthiness of an analysis process or a report, and understand the activities associated with in studies involving a given data source [23].

The tracking of historical information concerning a data set is also known as *data provenance* [25]. Data provenance is often seen as a kind of metadata used to describe the derivation history of a data product [30]. It represents the *who, what, when, where, why* and *how* associated with a resource. In the scientific community, *data provenance* has become a basis to determine the authorship, data quality, and to allow the reproducibility of findings [30]. Research on provenance has merited special attention in the context of scientific workflows dealing with a large volume of data and domains like biology, bioinformatics or chemistry [32, 35].

Data provenance in the geospatial domain has been a research topic for a long time - e.g., the work of Lanter [20]. A recent example of this interest is presented by Yue et al. [38], who use metadata to capture data provenance when geospatial data are produced in a service-oriented environment. However, these solutions do not highlight issues of quality in their models. As mentioned by [8] and [30], data provenance can help determine data quality. Our work goes a step further, presenting a model to bridge the gap between provenance and quality of geospatial data. Our solution takes advantage of features provided by the Open Provenance Model (OPM) and FGDC geographic metadata standards [14].

The main contributions of our work are therefore: (1) the proposal of a data quality model based on provenance information to help in the assessment of geospatial data products and (2) the definition of metadata elements which will be used to evaluate the quality of geospatial data sources in agriculture. The rest of this paper is organized as follows. Section 2 presents some challenges of data provenance in the geospatial domain. Section 3 describes our provenance model, giving details of its structure. Section 4 describes a case study in agriculture. Section 5 contrasts our proposal with related work. Finally, section 6 describes conclusions and future work.

2. GEOSPATIAL DATA PROVENANCE

In a broad sense, data provenance can be defined as “the place or source of origin where something was created or collected along with the records of their activities during the course of ownership.”¹ Provenance and *traceability* are therefore intimately associated. Distinct scientific communities instantiated the notion of provenance differently. For instance, database researchers defined it as the description of where a piece of data came from and the process by which it arrived in the database [9]. For the scientific workflow community, data provenance or workflow provenance describes the entire history of the derivation of the final output of a workflow [15]. These two definitions lead to Tan’s [31] characterization of granularities of provenance: coarse-grained and fine-grained. Coarse-grained provenance refers to workflows, and fine-grained provenance is related to a dataset in a database. Process provenance is also often associated with workflow execution [7].

FGDC [14] and ISO [18] standards have included data provenance in the geospatial data quality section. The FGDC metadata standard, for example, uses the field *lineage* to describe the information about the events, parameters, source data and responsible entities that participated in the construction of a data set. Yue and He [39] define data provenance as the processing history of a geospatial product. In our work we follow the same definition.

In order to tackle the problem of representing the provenance of geospatial data, we must consider its peculiarities. Furthermore, even primary data sources undergo cleaning and curation procedures - e.g., a temperature sensor time series may have outliers eliminated, or a satellite image may have undergone filtering processes. Secondary and derived data sources pose even more problems, since they result from specific combinations of merging data, analysis procedures and expert interpretations. For instance, a given erosion map is the result of combining data on soil, vegetation, land use, using algorithms defined by experts in this combination, and then applying weights to distinct areas according to experts’ experience, needs or goals. Soil and land use data themselves may be obtained through analysis of other data, and so on.

These successive analysis, merging and fusion steps may introduce errors that will be propagated and enhanced. Perhaps the most complex issue is the fact that geospatial data are very dependent on the usage context - in particular, its

dependence on space and time. Here, since we use provenance to assess quality, we point out that a common definition of quality is “fitness of use”. So, how can we attack the problem of handling provenance, and its dependence on context?

The representation of data provenance requires the development and implementation of a provenance model. We argue that although a model should be tailored to its own needs, it should follow standards which allow its interoperability with other systems [25]. Furthermore, since we are working in the geospatial domain, this model must take into account the spatial and temporal elements and the relationships among them [37, 38, 39]. The strategy to store provenance information is important to its scalability [30]. Since data provenance can be seen as a kind of metadata, it could be stored in a metadata catalogue. Another possibility is to store data provenance in a provenance repository as proposed in [24]. To disseminate geospatial data provenance we can use services provided by the Open Geospatial Consortium (OGC) [26] as described in [39].

3. PROVENANCE MODEL

Our work combines characteristics provided by the Open Provenance Model (OPM) [25] and quality elements from FGDC’s [14] geospatial metadata standard. Provenance in OPM is represented by provenance graphs. A *provenance graph* (also called causality graph) is a directed acyclic graph which represents causal dependencies among entities. Since OPM does not specify the metadata or internal representation that systems have to adopt to manage provenance, it is up to each domain to adapt the model. In our case, this adaptation comes from the FGDC standard. Although the proposed model is based on the OPM model, we added our own characteristics taking into account the geospatial domain we are working in and that our model highlights the assessment of data quality.

3.1 Model overview

The basic premise of our work is that geographic information needs to have elements which allow to know whether the data is reliable, so that it can be consumed. Our second premise is that, once data provenance can be used to estimate data quality [16], we can use provenance as a means to assess reliability. In the geospatial context, several aspects need to be taken into account when it concerns quality and reliability. To make sure data are useful, we need to provide means to trace the origins of a data source, i.e. to link it with its past. For instance, if the product is a map we need to face qualitative (e.g., mapping methodologies) and quantitative (e.g, resolution) factors. Furthermore, we need to know what is the level of reliability of the entities involved in the data collection and analysis activities.

To alleviate these problems, we designed a model to record the provenance of geospatial data. Our objective is to provide a means to assess and store quality elements. Our research considers the *trustworthiness of source* and *temporality* dimensions of data quality proposed in [21] and also studied in [29]. *Trustworthiness of sources* refers to the degree of confidence of who created or made available the data. *Temporality of data* includes valid and transaction time (i.e., when). Unlike the work described by [29], where

¹Merriam-Webster dictionary

only numeric data values are considered, we also include data such as maps, satellite images, remote data sensing products, data from official providers, among others. Besides *who* (trustworthiness) and *when*, we also need to capture the location where an event has happened, i.e. *where*.

Figure 1 illustrates our provenance data model using the entity relationship notation. The part in bold comes from OPM, the next was added by us. The basic pieces of the model are *Artifact*, *Process* and *Agent*. While the Artifact entity concerns the geospatial data products, the Process entity deals with the processes that generated an Artifact. Finally, the Agent entity is in charge of to execute processes or to provided artifacts. In our model, trust criteria are associated to an artifact and an agent. At this point of this research, we use values ranging from 0 to 1.

Artifact, Process and Agent entities are elements imported from OPM [25]. In the OPM an *artifact* is considered as a physical object (e.g., a book) or a digital representation (e.g., a dataset) in a computer system. A *process* represents activities or actions performed resulting in new artifacts. An *agent* represents entities controlling the process. Examples of artifacts in this work are a remote sensing image or the level of erosion derived from analysis of this image. An Artifact can be provided by an Agent, for example, an official institution like Brazil’s National Geographic Institute (IBGE) [1] or may be the result the execution of a process. A Process is controlled by an Agent and it also might trigger subprocesses.

In our model we consider that at a specific time a process can have several inputs, but can only generate one outcome. In the context of provenance, valid and transaction time are concepts explored by Prat and Madnick [29]. According to them, transaction time can be the timestamp of when a process was run or a time which is attached to a result. This extends the standard definition of transaction time in temporal databases, which is defined by the system clock when a data item is stored [19]. On the other hand, valid time follows the temporal database definition: it is the time (instant or interval) when the fact is true in a specific context [19]. In the geospatial domain, it is important to know how old a data source is because, in some cases, the usefulness and quality of a source decays with age. Therefore, Valid time in the model concerns an Artifact and Transaction time concerns a Process. We follow these definitions to adapt to the geospatial context.

URL Address links an Artifact to its location in a database or directory file. We assume that data related to geographic coordinates or another kind of spatial features are stored in spatial repositories provided by an Agent. *Measure criteria* about data quality have been taken from the FGDC metadata standard [14] and linked to an Artifact.

An Agent uses and applies some methodologies according to the domain where it works. The grade of trust (*Trust Grade*) of an Agent depends on issues such as: is it an official source, what is the reputation of this provider, is it an academic research group, among others. This scenario shows that assigning a confidence value to an agent can be very subjective. Another issue is related to granularity. This

is a real concern that we will face in a future refinement of the model.

3.2 Quality elements

FGDC [14] provides a set of terms to document digital geospatial data. This specification is composed by seven main sections of metadata and three support sections. The main sections contain information about identification, data quality, spatial data organization, spatial reference organization, entity and attribute, distribution information and metadata reference. Support sections contain information on citation, time period of content and contact.

Although there are several metadata criteria related to the quality section in FGDC, a full description using all fields from this section may be too long. Hence, we selected the most relevant parts of this section, taking into account the extensive experience in agricultural planning and monitoring based on processing remote sensing sources (satellite images). These parts are: *positional accuracy*, *logical consistency*, *completeness* and *attribute accuracy*.

- Positional accuracy refers to the accuracy of the positions of spatial objects.
- Logical consistency indicates the fidelity of relationships in the data set and tests used.
- Completeness is information about omissions, selection criteria, generalization, definitions used, and other rules used to derive a data set.
- Attribute accuracy indicates how thoroughly and correctly the features in the data set are described.

It is important to notice that although these are the first metadata elements that we selected to study, we can add other elements (e.g., coverage, horizontal accuracy, etc) that complement them. Each of these criteria must be assigned quantifiers, i.e. a value obtained from the computing of quality of the attributes related to the Artifact. However, tuning these quantifiers is not a trivial work and depends on the usage for which the Artifact is intended. As a first step, we begin by assigning trust values ranging from 1 to 0 to the Agent. This means that the higher the trust value is, the most reliable an Agent is. This is not the first work that uses this kind of values, trustworthiness in an Agent is studied in [29]. Unlike that work, we can get trust values from our domain experts.

4. CASE STUDY

This section presents a case study that concerns the use of our provenance model, showing how it supports the assessment of data quality.

4.1 Problem Overview

The case study concerns the creation of a map (the geospatial product) showing the agricultural mapping for coffee in a given region. The main data sources for this product are remote sensing images. Remote sensing images provide information about the geography and characteristics of an area. They might be acquired from providers like *SPOT*

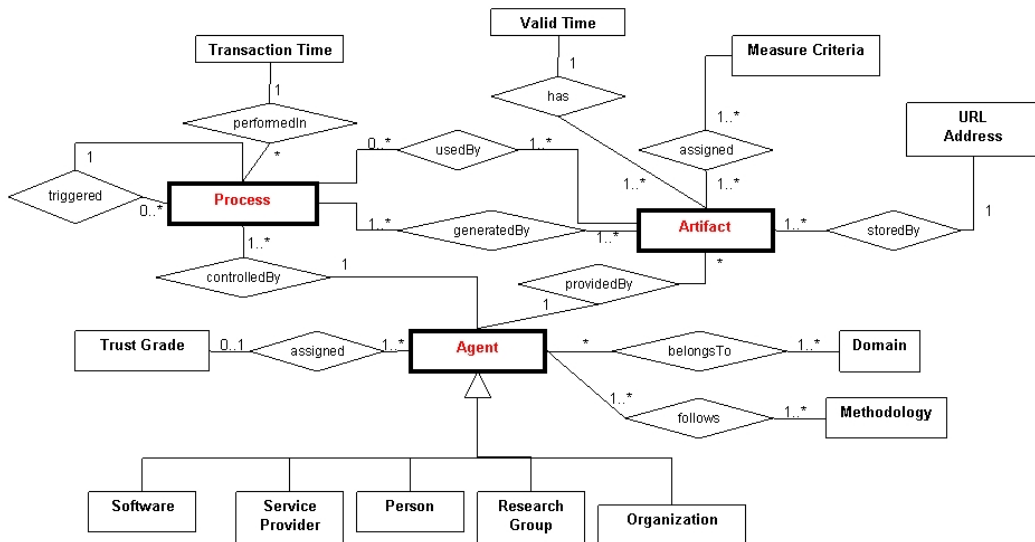


Figure 1: The provenance model

image [4], an official provider who has permission to commercialize SPOT satellite images. Technical documentation and confidence level - e.g., accuracy, scale, and others - related to the image acquisition are described in the manual *Accuracy and Coverage combined*². In order for images to allow an easy identification of coffee areas or others - e.g., forest, roads, etc -, these images are submitted to several processes, for example, to improve their visual appearance or accuracy. Furthermore, a human operator is in charge of image interpretation - e.g., in our case mapping coffee areas - with or without help of special software procedures (e.g., using machine learning techniques) [13]. All these activities are performed by users with a high level of knowledge. At the end, a map with the areas of coffee is obtained. Experts can use this map for crop monitoring, to identify regions with problems, for yield forecast analysis, among others. In some cases, official agencies provide documentation about the procedures and methodologies applied to process the image to obtain the desired product.

4.2 Instantiating the model

Let us now apply our model to this case study. In general, quality issues about an image, the Artifact, are related to textual documentation concerning the activities performed to produce it. Although the documentation is a huge effort to give historical information about how a data product was derived, we believe that this can be restrictive. Information contained only in documents does not facilitate content discovery. Moreover, these documents will not enable to answer user queries in a timely manner. In this context, the only way to know whether a data source is reliable or not is looking for this documentation and reading all the information. Local data providers (e.g., NASA [2], or INPE [3]) the equivalent agency in Brazil) attach quality metadata to im-

ages, thereby partially solving this problem, since this can be queried electronically. However, these metadata fields do not describe the entire process for producing an artifact. Our provenance model attempts to address these gaps.

Figure 2 presents a high level view of the workflow that generates the final map. It shows that the input SPOT image goes through two processing steps, performed by the CEPAGRI agent. The first step concerns corrections and the second the creation of the product.

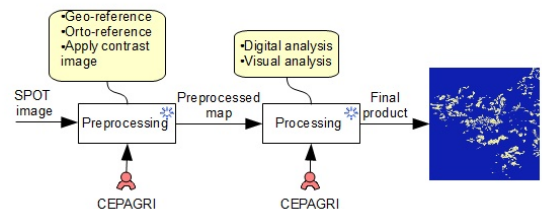


Figure 2: Map process generation

We start by identifying artifacts (an image, but also the final product), processes and agents, and assigning trust values. For instance, in the agriculture domain an Agent can be represented by the Agriculture Research Center at UNICAMP (CEPAGRI)³. Since CEPAGRI has long-standing procedures for processing remote sensing images, the trust value assigned to it should be 1. CEPAGRI is in charge of mapping coffee crop areas. The elaboration of this product involves several activities such as geo-reference, visual and

²<http://brasil.spotimage.com/web/pt/1910-imagens-spot.php>

³<http://www.cpa.unicamp.br/>

digital analysis, among others (see figure 2). These processes can be performed using image analysis tools like ENVI [5]. A process has a transaction time which is the execution time when a process or set of processes were started to generate a data product. In this paper we are only considering processes that were fully executed. Here, the Valid time for a coffee crop map is the range of dates related to a period of time since it was generated.

Accuracy is widely used in the geospatial domain to measure the degree of closeness of measured values - e.g., coordinates - with the real value. Hence, this is a kind of information that can be acquired by the experts. In our context, the identification of coffee areas took into account the accuracy obtained using the Kappa index [10]. This is a statistic index which measures the quality of thematic maps. Therefore, we can use the Kappa index as a reference to assign the “Completeness” measure criterion. The higher the Kappa index is, the higher the value that will be assigned to “Completeness” criterion. Table 1 shows a simplified example of database entries concerning Artifacts with some of its attributes. The attribute *id* identifies an artifact. The artifact’s name is recorded by the attribute *name*. References to the physical location and valid time are identified by *idURL* and *idVTime* respectively. Here, Valid Time is an interval. Attributes of Measure Criteria such as *measure_name* and *measure_value* are shown in table 2. Table 3 represents the allocation of a criterion to the artifact. For instance, it shows that Artifact A2 (coffee_crop_map) is linked to criterion C1. Here, at the end, whoever uses the final map to analyze coffee crop regions will know that its completeness is 0.8, and that it is valid only in an interval time t3-t7.

When an artifact is a data value, for example a productivity index, it is necessary to consider new measure criteria. In order to cover this issue, we follow the approach described in [29], where a value of 1 is assigned when a data value is inside a range of values, and 0 if it is outside.

Table 1: Table Artifact

id	name	idURL	idVTime
A1	SPOT_image	url15	t1-t5
A2	coffee_crop_map	url12	t3-t7

Table 2: Table Measure Criteria

id	measure_name	measure_value
C1	Completeness	0.8

Table 3: Table Artifact_assigned_Criteria

id	idArt	idMCriteria
S1	A2	C1

5. RELATED WORK

Management (and definition) of geospatial data provenance is a not a trivial task. There are many specific concerns re-

lated to user heterogeneity, volume of data and complex processing steps used by systems to produce information. An early study on provenance in GIS is [20]. The author developed a meta-database system for tracking the input-output relationships between map layers and frames from spatial data sets. Another work is Geo-Opera [6], which supports the tracking of geospatial analysis history in a workflow system.

Pastorello et al. [27] describe a framework to support documentation of cooperative environmental planning activities on the Web. Documents generated and processes (workflows) executed during environmental planning are linked to each other and associated with geographical metadata and domain ontology terms. Hence, documents record provenance in three different aspects: what, how and why. Another effort to manage the what, how and why in cooperative work is presented by Voisard et al. [36]. Contrary to [27], these authors use a database supported by a version mechanism to manage those information.

An approach to record the provenance of spatial data together with the analysis of its derivation history is presented in Wang et al. [37]. They use the OPM model to handle spatial data provenance. In our work we also use OPM to develop a provenance model, but contrary to [37] where the main concern are spatial regions, we can cover any geospatial data sources, regardless of spatial extent (e.g., temperature sensor time series).

Other examples of study of geospatial data provenance are described in [34] and [38]. Tilmes and Fleig [34] discuss some general concerns of provenance in the context of Earth data processing systems. Yue et al. [38] have developed an approach to capture geospatial data provenance in the context of geospatial Web services and geo-processing service chains. In addition, the authors use metadata to automatically capture data provenance in the phase of process model instantiation. Contrary to this work, our objective is to capture how a data product (Artifact) from sources (also Artifacts) was derived and to add measure criteria aiming to assess its trustworthiness.

We combine OPM with FGDC standard. There are several metadata standards that were designed to be applied to specific domains. For instance, the Dublin Core Metadata Initiative (DCMI) [12] is a standard for describing digital resources, such as images, services, and physical objects. Examples of elements that can be used to describe provenance information are: *creator*, *contributor*, *source*, *publisher* and *provenance*. In biodiversity, Darwin Core (DwC) is a metadata standard specification to describe information about the occurrence of species and the existence of specimens in collections [33]. It specifies a long list of elements, some of which can help to describe provenance - e.g., *recordedBy*, *associatedMedia*, *identifiedBy*, *identificationReferences*, *identificationQualifier*, among others.

In the geographic information context according to Prado et al. [28] the more widely used metadata standards are FGDC [14] and ISO 19115:2003 [18]. These standards allow the textual description of geographic datasets. Although both of these standards supply elements to assess the quality

of a data source, we chose the FGDC standard, since it is open and easier to understand. We also take into account the knowledge acquired from other research developed in our laboratory using geospatial data [22]. Our solution uses quality elements from FGDC to evaluate the trustworthiness of a geospatial data product produced by several sources. However, we are aware that we need to establish explicit values of trust which allow to assess a data product. This is a topic of discussion of ongoing work.

Work that addresses data quality as associated with provenance issues is presented in [29], [11] and [17]. Prat and Madnick [29] provide a solution to measure and compute data believability based on the provenance of a data value. The computation of believability has been structured into three complex building blocks: metrics for assessing the believability of data sources, metrics for assessing the believability from process execution and global assessment of data believability. Although this is a precise approach to compute believability, the authors only measure the trustworthiness of a numeric data value, which limits its applicability to geospatial data.

Dai et al. [11] describe an approach to determine the trustworthiness of data integrity based on source providers and intermediate agents. Another example is presented by Hartig and Zhao [17] who describe an approach to measure the trustworthiness of data on the web based on the timeliness of data.

To our knowledge, ours is the first work to develop a provenance model to explicitly capture quality of data making use of geographic metadata standards. Hence, it provides the basis for measuring the reliability of geographical data taking into account the providers and the processes that they perform.

6. CONCLUSIONS

Geospatial data are a basis for decision making activities. Common problems are related to how to document and preserve processes that generate products ensuring reproducibility; how to organize and integrate products; how to share the findings and how to assess the quality of the geospatial data results. Most efforts to improve this situation concentrate on establishing documentation about data capture, methodologies, curation standards and quality metadata.

This paper presented and discussed an approach based on data provenance for alleviating this problem. Our provenance model take advantage of features provided by the Open Provenance Model, which are being used by the scientific community to instantiate their solutions. Furthermore, our model integrates concepts from the FGDC metadata standard that we will need for data quality assessment. Taking this into account, we present the first steps to assign weights to the trust criteria. A real case study in agriculture was described, showing how the provenance-based data quality model works.

As future work we intend to investigate techniques to compute and assess the trustworthiness of data. Therefore, we also need to study the best strategies to assign trust values to the measure criteria. Another possibility that we need to

consider is to expand to other kinds of measure criteria. Furthermore, granularity issues are not attacked in this paper, and this is an important problem that we intend to explore further.

Acknowledgment

The authors would like to thank CNPq (process number 142337/2010-2), CAPES and FAPESP for the financial support for this work, as well as the Brazilian Institute on Web Science.

7. REFERENCES

- [1] Instituto Brasileiro de Geografia e Estatística. <http://www.ibge.gov.br/home/>. Accessed in May 2011.
- [2] National Aeronautics and Space Administration. <http://modis.gsfc.nasa.gov/>. Accessed in June 2011.
- [3] Instituto Nacional de Pesquisas Espaciais. <http://www.inpe.br/>, 2006. Accessed in June 2011.
- [4] Spot image. <http://brasil.spotimage.com/>, 2008.
- [5] Environment for visualization images. <http://www.ittvis.com/ProductsServices/>, 2011. Accessed in Jun 2011.
- [6] G. Alonso and C. Hagen. Geo-Opera: Workflow Concepts for Spatial Processes. In *Proceedings of the 5th International Symposium on Advances in Spatial Databases*, London, UK, 1997. Springer-Verlag.
- [7] R. S. Barga and L. A. Digiampietri. Automatic capture and efficient storage of e-Science experiment provenance. *Concurr. Comput. : Pract. Exper.*, 20(5):419–429, 2008.
- [8] P. Buneman, A. Chapman, and J. Cheney. Provenance management in curated databases. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 539–550, New York, NY, USA, 2006. ACM.
- [9] P. Buneman, S. Khanna, and W.-c. Tan. Why and Where: A Characterization of Data Provenance. In *In ICDDT*, pages 316–330. Springer, 2001.
- [10] R. G. Congalton and K. Green. *Assessing the accuracy of remotely sensed data: principles and practices*. Lewis, Boca Raton, FL, 1999.
- [11] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu. An Approach to Evaluate Data Trustworthiness Based on Data Provenance. In *Proceedings of the 5th VLDB workshop on Secure Data Management*, pages 82–98, Berlin, Heidelberg, 2008. Springer-Verlag.
- [12] DCMI. The Dublin Core Metadata Initiative. <http://dublincore.org/>, 2010. Accessed in May 2011.
- [13] J. A. dos Santos, C. D. Ferreira, R. d. S. Torres, M. A. Gonçalves, and R. A. C. Lamparelli. A Relevance Feedback Method based on Genetic Programming for Classification of Remote Sensing Images. *Information Sciences*, 2010.
- [14] FGDC. Content Standard for Digital Geospatial Metadata FGDC-STD-001-1998. Technical report, US Geological Survey, 1998.
- [15] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for Computational Tasks: A Survey. *Computing in Science and Engg.*, 10(3):11–21, 2008.
- [16] C. Goble. Position statement: Musings on provenance,

- workflow and (semantic web) annotations for bioinformatics. In *Workshop on Data Derivation and Provenance*, 2002.
- [17] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *In: Proc. of the Workshop on Semantic Web and Provenance Management at ISWC*, 2009.
- [18] ISO. Geographic information – metadata. iso 19115:2003. <http://www.iso.org/iso/>, 2003. Accessed in May 2011.
- [19] C. S. Jensen, J. Clifford, R. Elmasri, S. K. Gadia, P. J. Hayes, and S. Jajodia. A Consensus Glossary of Temporal Database Concepts. *SIGMOD Record*, 23(1):52–64, 1994.
- [20] D. P. Lanter. Design of a Lineage-Based Meta-Data Base for GIS. *Cartography And Geographic Information Systems*, 18(4):255–261, 1991.
- [21] Y. W. Lee, L. Pipino, J. D. Funk, and R. Y. Wang. Journey to Data Quality. *MIT Press, Cambridge, MA*, 2006.
- [22] C. N. Macário. *Semantic Annotation of Geospatial Data*. PhD thesis, Instituto de Computação - Unicamp, 2009.
- [23] M. Mccann and K. Gomes. Oceanographic Data Provenance Tracking with the Shore Side Data System. In J. Freire, D. Koop, and L. Moreau, editors, *Provenance and Annotation of Data and Processes: Second International Provenance and Annotation Workshop, IPAW 2008*, chapter Oceanographic Data Provenance Tracking with the Shore Side Data System, pages 309–322. Springer-Verlag, Salt Lake City, UT, USA, 2008.
- [24] S. Miles, P. Groth, M. Branco, and L. Moreau. The Requirements of Using Provenance in e-Science Experiments. *Journal of Grid Computing*, 5(1):1–25, 2007.
- [25] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. T. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. G. Stephan, and J. V. den Bussche. The Open Provenance Model core specification (v1.1). *Future Generation Comp. Syst.*, 27(6):743–756, 2011.
- [26] OGC. The open geospatial consortium. <http://www.opengeospatial.org/>, 2011. Accessed in June 2011.
- [27] G. Pastorello, C. B. Medeiros, S. de Resende, and H. da Rocha. Interoperability for GIS Document Management in Environmental Planning. In S. Spaccapietra and E. Zimányi, editors, *Journal on Data Semantics III*, volume 3534, pages 586–587. Springer Berlin / Heidelberg, 2005.
- [28] B. R. Prado, E. H. Hayakawa, T. C. Bertani, G. B. S. Silva, G. Pereira, and Y. E. Shimabukuro. Standards for digital geographic metadata: the iso 19115:2003 model and the fgdc model. *Revista Brasileira de Cartografia*, 62(1):33–41, 2010.
- [29] N. Prat and S. Madnick. Measuring Data Believability: A Provenance Approach. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, volume 0, page 393, Los Alamitos, CA, USA, 2008. IEEE Computer Society.
- [30] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36, 2005.
- [31] W. C. Tan. Provenance in Databases: Past, Current, and Future. *IEEE Data Eng. Bull.*, 30:3–12, 2007.
- [32] Taverna. The Taverna Project. <http://www.taverna.org.uk/>, 2009. Accessed in June 2011.
- [33] TDWG. Darwin Core Task Group. <http://www.tdwg.org/standards/450/>, 2009. Accessed in May 2011.
- [34] C. Tilmes and A. J. Fleig. Provenance Tracking in an Earth Science Data Processing System. In *Proceedings of Provenance and Annotation of Data and Processes: Second International Provenance and Annotation Workshop, IPAW 2008*, pages 221–228, Salt Lake City, UT, USA., 2008. Springer.
- [35] VisTrails. The VisTrails Project. <http://www.vistrails.org>, 2011. Accessed in June 2011.
- [36] A. Voisard, C. B. Medeiros, and G. Jomier. Database Support for Cooperative Work Documentation. In *Proceedings of COOP’2000*, 2000.
- [37] S. Wang, A. Padmanabhan, J. D. Myers, W. Tang, and Y. Liu. Towards provenance-aware geographic information systems. In W. G. Aref, M. F. Mokbel, and M. Schneider, editors, *GIS*. ACM, 2008.
- [38] P. Yue, J. Gong, and L. Di. Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. *Comput. Geosci.*, 36:270–281, 2010.
- [39] P. Yue and L. He. Geospatial data provenance in Cyberinfrastructure. In *Geoinformatics, 2009 17th International Conference on*, 2009.