A Provenance Approach to Assess the Quality of Geospatial Data

Joana E. Gonzales Malaverri[1], Claudia Bauzer Medeiros[1], Rubens Camargo Lamparelli[2]

[1]Institute of Computing – State University of Campinas (UNICAMP)
13083-852 – Campinas – SP – Brasil

[2]Center for Research in Agriculture – State University of Campinas (UNICAMP)
13083-970 – Campinas – SP – Brasil

{jmalav09, cmbm}@ic.unicamp.br, 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 and trustworthiness – of the data sources, of the data providers, and of the data products derived from the original sources. 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.*

1. Introduction

We use geospatial data everyday and everywhere. 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. 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 [Buneman et al. 2006].

The tracking of historical information concerning a data set is also known as *data provenance*. In the scientific community, *data provenance* has become a basis to determine authorship, data quality, and to allow the reproducibility of findings [Simmhan et al. 2005]. In real life situations, provenance information of geospatial data is used to decide pre-processing procedures, storage policies and even data cleaning strategies – with direct impact on data analysis and synthesis policies.

Trust and quality go hand-in-hand. Taking this into account, our work describe a geospatial data provenance model to help to determine whether (and how much) users can trust data sources and data providers, and to assess data quality. Our solution takes advantage of features provided by the Open Provenance Model (OPM) [Moreau et al. 2011] and FGDC geographic metadata standards [FGDC 1998].

2. Model overview

The basic premise of our work is that, given its importance, geographic information needs to have elements which allow to know whether the data are reliable, so that it can be

consumed. Our second premise is that, once data provenance can be used to estimate data quality, we can use provenance as a means to assess trustworthiness. For instance, if the data to consider is a map, we need to face qualitative (e.g., mapping methodologies) and quantitative (e.g, resolution) factors. Furthermore, we need to know the level of reliability of the entities involved in the data collection (e.g., providers) and analysis activities used to produce the map.

Our research considers the *trustworthiness of source* and *temporality* dimensions of data quality of [Prat and Madnick 2008]. *Trustworthiness of sources* (who) refers to the degree of confidence of who created or made available the data. *Temporality of data* (when) includes valid and transaction time. Besides *who* and *when*, we also need to capture the location where a event has happened, i.e *where*.

Figure 1 illustrates the main elements of our provenance data model using the entity relationship notation. The part in bold comes from OPM, the rest was added by us. The basic pieces of the model are *Artifact, Process* and *Agent*. While the Artifact entity concerns geospatial data products, the Process entity deals with the processes that generated an Artifact. Finally, the Agent entity is in charge of executing processes or providing artifacts. In our model, trust criteria are associated to an Artifact and an Agent and have normalized values ranging from 0 to 1.

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 NASA or Brazil's National Geographic Institute (IBGE), or may be the result the execution of a process. A Process is controlled by an Agent and it also might trigger subprocesses.

Our model considers that at a specific time a process can have several inputs, but can only generate one outcome. In the geospatial domain, in some cases, the trustworthiness and quality of a source decay with age. Therefore, Valid time concerns an Artifact and Transaction time concerns a Process. *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 [FGDC 1998] 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, the reputation of this provider, is it an academic research group. This scenario shows that assigning a confidence value to an agent can be very subjective.

3. Quality elements

FGDC [FGDC 1998] provides a set of terms to document digital geospatial data, with several metadata criteria. We selected the most relevant criteria, taking into account our experience in agricultural planning and monitoring based on processing remote sensing sources (satellite images). These parts are:

- *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.

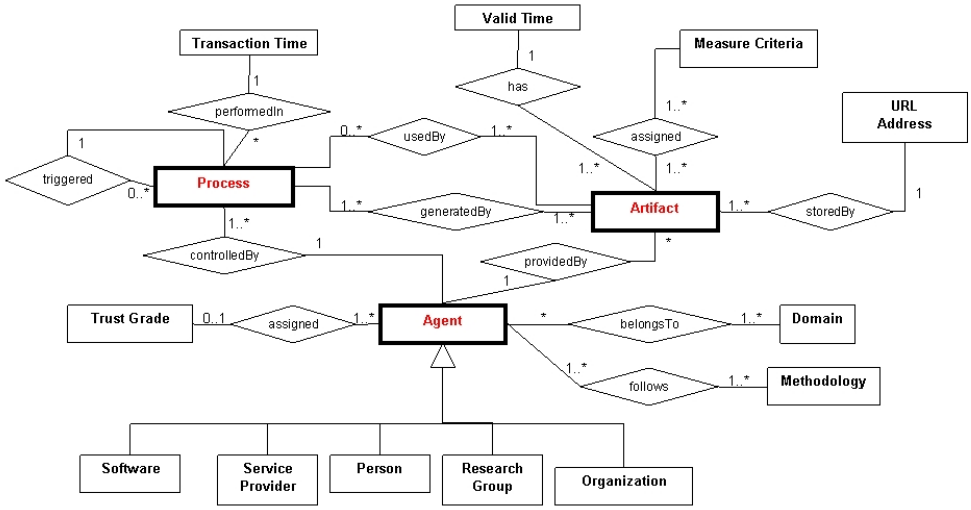

Figura 1. Our provenance model

- *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.

Though these are the basic metadata elements that we selected, we can add other elements (e.g., coverage, horizontal accuracy) that complement them. Each of these criteria must be assigned quantifiers, i.e. a value obtained from computing the 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.

We are conducting case studies in agriculture to validate our model, storing quality information in database tables. This database is created from the ER diagram in figure 1. In such examples, input data concerns satellite images, crop information and others. Outputs include maps and reports, produced after several manual and automatic processing steps. All these are taken into consideration in provenance and quality evaluation.

4. Conclusions

Geospatial data are a basis for decision making activities that affect our daily lives. The trustworthiness of these data (and recommendations based on analyses thereof) is becoming increasingly important. This is complicated by the fact that the processing of geospatial data is essentially a cooperative, distributed, effort, which hampers determining its reliability. Most efforts to improve this situation concentrate on establishing documentation about data capture, methodologies, curation standards and quality metadata.

This paper presented a novel approach based on data provenance for alleviating this problem. Our provenance model takes advantage of features provided by the Open Provenance Model, which are being used by the scientific community to instantiate their

solutions. The model integrates concepts from the FGDC metadata standard needed for assessment of data quality.

Acknowledgments: Support from CNPq project, CAPES, FAPESP, and the Brazilian Institute on Web Science.

Referências

Buneman, P., Chapman, A., and Cheney, J. (2006). Provenance management in curated databases. In *Proc. of the 2006 ACM SIGMOD Conf.*, pages 539–550.

FGDC (1998). Content Standard for Digital Geospatial Metadata FGDC-STD-001-1998. Technical report, US Geological Survey.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P. T., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. G., and den Bussche, J. V. (2011). The Open Provenance Model core specification (v1.1). *Fut. Gen. Comp. Syst.*, 27(6):743–756.

Prat, N. and Madnick, S. (2008). Measuring Data Believability: A Provenance Approach. In *Proc. 41st Hawaii Int. Conf. on System Sciences*, volume 0, page 393. IEEE Computer Society.

Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36.