Handling Provenance in Biodiversity

Joana E. Gonzales Malaverri[1], **Claudia Bauzer Medeiros**[1]

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

{jmalav09, cmbm}@ic.unicamp.br

Abstract. *One of the concerns in eScience research is the design and development of novel solutions to support distributed collaboration. In this context, regardless of the scientific domain, an important problem is the reproducibility of the results from scientific activities, considering the heterogeneous data involved and the specific research context. This paper presents a proposal to help solve this problem, proposing a software architecture to handle provenance issues.*

1. Our approach for provenance and quality

The goal of this work is to provide means to collect and store provenance information related to the scientific processes performed by researchers in biodiversity studies, in the context of eScience on the Web. We are concerned with two kinds of provenance information: data provenance (who, what, where, when and how associated to a given data set) and process provenance (related to the execution of scientific workflows that receive and produce such data sets). Our approach complements other studies, whose main focus is to enable the reproducibility of the data derivation processes.

To become useful, provenance information needs to be digitally discoverable, accessible, comprehensible, and provide necessary context information to reproduce data analysis results [Myers et al. 2009]. For this reason, we have decided to explore semantic annotations as a means to record provenance information. Our semantic annotations are formally described in [Pastorello Jr. et al. 2008] as a set of metadata fields that are associated to a term from a domain ontology. Since we are concerned with biodiversity domain we will start with the DwC metadata standard [TDWG 2009] and extend it to reflect provenance needs. We will need to define the domain ontologies, to represent the concepts and relationships described by the metadata. Metadata fields must be accompanied by elements that help in the assessing of the quality of the data (e.g., accuracy, precision).

Figure 1 outlines our architecture, where each box denotes different data access and manipulation levels. The two boxes outlined are the main focus of this research. The Data Acquisition Process can rely on some data acquisition software, which works as a mediator to data sources, or specialized spreadsheets used by biologists to insert data directly. In more detail, raw data are processed (1) and stored in the Data Repository (2). Data are used by processes run as scientific workflows (3) that are retrieved from the Workflow Repository (4). Results are published (5) by specific processes and, again, stored in a Data Repository. At all these steps, the Provenance Extraction Process (6) extracts metadata information from data and processes, storing such metadata in the corresponding Provenance Metadata Repository (7), where metadata fields point to ontology

terms from the Ontology Repository (8) - the semantic annotation. The Ontology Repository contains the concepts related to a specific application domain. In the end, specific processes may be invoked to assess result quality, based on provenance information.

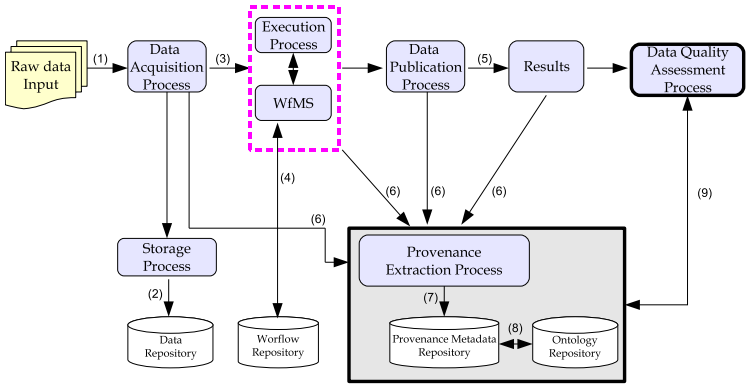

Figura 1. Supporting provenance extraction and management

Our implementation takes advantage of the annotation infrastructure of [Macário 2009]. Ontology management and ontology repository are provide by Aondê, an ontology web service responsible for wide range of operations to manipulate ontologies. Furthermore, we have a set of modules that allow to manage and query biodiversity records.

There are many challenges to be met in this work, including the fact that we need to consider the provenance of manual activities, different Web information sources, and the assessment of the results based on data inputs and in the processes executed to generate them. Since most data sources (and even processes) are on the Web, this also involves challenges in Web Science. We need furthermore consider how to acquire the metadata that we use to describe provenance and what actually is considered by the researchers to be data provenance. We will also need to define the methodologies to estimate the trustworthiness of data.

Acknowledgments This work was partially financed by CNPq (BioCORE project), INCT in Web Science (CNPq 557.128/2009-9) and CAPES.

Referências

Macário, C. N. (2009). *Semantic Annotation of Geospatial Data.* PhD thesis, Instituto de Computação - Unicamp.

Myers, J. D., Futrelle, J., Gaynor, J., Plutchak, J., Bajcsy, P., Kastner, J., Kotwani, K., Lee, J. S., Marini, L., Kooper, R., McGrath, R., McLaren, T., Rodriguez, A., and Liu, Y. (2009). Embedding Data within Knowledge Spaces. *CoRR*, abs/0902.0744.

Pastorello Jr., G. Z., Daltio, J., and Medeiros, C. B. (2008). Multimedia Semantic Annotation Propagation. In *ISM '08: Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia*, pages 509–514, Washington, DC, USA. IEEE Computer Society.

TDWG (2009). Darwin Core. http://www.tdwg.org/standards/450/. Accessed in June 2010.