Estimating the quality of data using provenance: a case study in eScience

Research-in-Progress

Joana E. G. Malaverri
Institute of Computing – UNICAMP
jmalav09@ic.unicamp.br

Matheus Silva Mota
Institute of Computing – UNICAMP
mota@ic.unicamp.br

Claudia Bauzer Medeiros
Institute of Computing – UNICAMP
cmbm@ic.unicamp.br

ABSTRACT

Data quality assessment is a key factor in data-intensive domains. The data deluge is aggravated by an increasing need for interoperability and cooperation across groups and organizations. New alternatives must be found to select the data that best satisfy users' needs in a given context. This paper presents a strategy to provide information to support the evaluation of the quality of data sets. This strategy is based on combining metadata on the provenance of a data set (derived from workflows that generate it) and quality dimensions defined by the set's users, based on the desired context of use. Our solution, validated via a case study, takes advantage of a semantic model to preserve data provenance related to applications in a specific domain.

Keywords

Data provenance, metadata, data quality, evaluation of quality, biodiversity.

INTRODUCTION

Challenges related to the quality of data are common to applications in a variety of domains. Not only can it directly affect decision processes in an organization, but also in a scientific context (e.g., healthcare, environmental sciences, astronomy, etc). With the data deluge generated by groups and organizations around the world, there is a growing demand for new computing solutions to help decision-makers to select the best data that match their needs The same can be extended to a scientific environment: before scientists can take actions to analyze their findings, they need to know the quality of the data sets they are working on.

Problems to be faced include, for instance, data incompleteness, inconsistency, lack of standardization of formats, inaccurate data, among others. Besides that, data of different nature and the variety of information systems hamper the obtention of good quality data (Batini & Scannapieco 2006).

As will be seen, though our case study is in a specific domain (biodiversity), our proposal is generic enough to be applied and extended to any (computational/organizational) environment that requires cooperative work, and that must rely on integration of heterogeneous data sources. The underlying hypothesis is that there are a set of common characteristics in all such environments - such as the need for collaboration among actors with distinct needs and views of the issue at hand, a wide variety of heterogeneous data sources, and the need to coordinate complex data-driven processes.

Depending on the application domain, each of these problems demands different strategies to solve data quality issues. For instance, in the context of database systems, incompleteness of data might be tackled considering the granularity of its elements, i.e., completeness of value, tuple, attribute and relation (Batini & Scannapieco 2006). In the context of Web data, the same problem might be characterized by evolution in time - i.e., the speed at which the data will be completed (Pernici & Scannapieco 2002).

Related work has shown data quality to be a problem that has to be attacked under a multidimensional view (Richard & Diane 1996; Blake & Mangiameli 2011). Quality dimensions can be considered as attributes that allow to represent a particular characteristic of quality (Richard & Diane 1996). In particular, accuracy, completeness, timeliness and consistency have been extensively cited in the literature as some of the most important quality dimensions to information consumers

(Chapman 2005; Parssian 2006; Batini & Scannapieco 2006). These general dimensions can be considered common to both business and scientific domains.

The tracking of historical information concerning the creation of a dataset is also known as *data provenance* (Moreau et al. 2011). Provenance is seen as a kind of metadata that gives information about the what, when, where, how, by whom, and why a dataset, object or artifact was created (Sahoo et al. 2008). Taking these characteristics into account, we explore provenance as a strategy to provide information to evaluate the quality of data.

In some domains and applications, provenance information can involve a complex and scalable relationship network between different resources and processes (Chen et al. 2012; Goodchild & Li 2012; Barga & Digiampietri 2008). In this work we take advantage of the RDF/OWL model flexibility (Lassila & Swick 1999; Davies et al. 2006) and scalability (X. Wang et al. 2005) as a means to represent provenance information and its internal relationships, focusing on the biodiversity domain.

Unlike solutions centered on workflow systems such as (VisTrails 2011; Kepler 2011; Taverna 2009), which aim to provide native support for provenance to reproduce the planning and running of data processing and management steps, our approach can be adopted in different systems to collect domain-specific provenance and use this information to evaluate quality. Although this kind of approach is also investigated in (Sahoo et al. 2008) to allow knowledge discovery, we believe that different considerations need to be taken into account when it is used to analyze how good are the data produced by automated processes.

Our solution also addresses two requirements identified by the international provenance challenge[1] proposed in the context of the Open Provenance Model (Moreau et al. 2011). First, we show the applicability of provenance in the quality context by using it as a key parameter to help determine the quality of data in scientific organizational environment. Second, by making use of ontologies to represent provenance we allow interoperability among groups, enabling them to share and compare the information produced in their work.

The main contributions of this research include: (i) supporting the assessment of quality of scientific data based on its provenance and (ii) the adoption of a semantic model (PROV-O) to represent provenance. The latter extends our earlier work - in which we use a relational model to store provenance. Here, rather than a relational model, we extend the PROV-O semantic model to a new ontology, to consider domain-specific characteristics. We validate our approach through a case study concerning metadata generated in an information-intensive biodiversity experiment.

BACKGROUND OF THE SOLUTION

In our previous work (Malaverri & Medeiros 2013), we presented a conceptual framework to support keeping track of data provenance, in a relational model, to assess data quality. Our framework embeds a Provenance Manager service, a provenance database model, a Data Quality Manager service and a methodology to support the evaluation of the quality of data. Our focus in that work is the development of the data provenance repository and the application of the methodology in the estimation of the quality of data and reports produced in agricultural planning.

In that framework, the database that stores provenance information was designed using the Open Provenance Model (OPM) specification (Moreau et al. 2011). It represents data lineage in terms of **agents** that control **processes** to modify/produce **artifacts**. These elements are associated through five causal relationships within a provenance graph (e.g., an artifact **was generated** by a process). OPM only allows to represent artifacts as *immutable pieces of state*. This means that the state of an artifact cannot be modified after its creation.

Our methodology to support the evaluation of the quality of data in computational processes encompasses three main steps: (i) selection of the quality dimension(s) of interest; (ii) extraction of the information that is necessary to estimate the quality of the target dimensions; and (iii) computation of the score for each dimension. Users might use metrics to estimate the quality score or directly assign the scores based on the provenance information requested. We pointed out that each one of these stages is directly associated with the application domain under study and the activity that will be performed. Users choose the quality dimensions of their interest based on the kind of artifact under study (e.g., a spreadsheet file, a picture, a data statistics graph or a database table).

In this paper we adapt the methodology so that stage 2 should also consider the capture of metadata that will compose the provenance information. The retrieval of this provenance information is part of the activity that can be used to estimate the quality score. Given the fact that OPM does not allow object evolution, and considering that data evolution is a natural state of the world, we decided to change our provenance model. We extended PROV-O, a provenance semantic model, to

[1] *http://twiki.ipaw.info/bin/view/Challenge/WebHome*

represent domain-specific provenance. We adopted this approach because of the flexibility that ontologies provide. Provenance metadata are captured during the execution of operations on data. This can be achieved, as shown by previous work, by progressively storing execution traces, as well as information on data state changes – e.g., see (Kondo et al. 2007)

Earlier studies have investigated the support of provenance management based on domain ontologies (Zhao et al. 2008; Sahoo et al. 2008). The novelty in our work is to support data consumers on the estimation of quality. We can take advantage of several characteristics using an ontology-driven approach to represent and store provenance information. First, semantic modeling improves both interoperability and scalability of systems, since the schema and data can be more easily aligned with other schemas or instances. Second, adopting strategies like linked data (Yeganeh et al. 2011), each item of the schema and each data instance may have unique identifiers, thus enabling alignment with data from other sources that have also been modeled using semantics. This enables interoperability - not only within an organization, but across organizations, or groups.

Our work has been developed in the context of management of data and activities performed by scientists in the animal sound collection of the State University of Campinas, UNICAMP, Brazil (from now on called FNJV[2]). As will be seen, from a high level point of view, these activities are comparable to those executed by people in any information-sensitive organization, to collect, clean and publish their data sets.

PROVENBIO: A PROV-O-BASED ONTOLOGY FOR PROVENANCE INFORMATION FOR THE BIODIVERSITY DOMAIN

Any information-rich environment involves a complex and scalable relationship network between different and distributed resources, processes and users. Distinct organizational scenarios adopt distinct tools, vocabularies and methodologies. In order to represent provenance information and its relationships, we take advantage of the expressiveness that RDF/OWL provide, focusing on the specificities of the biodiversity domain.

PROV-O (W3C 2012) is an ontology based on OWL2 that specifies a data model to express provenance records in different application scenarios. PROV-O is a candidate recommendation in development by the W3C Provenance Working Group. It defines a set of starting point terms which are three core classes: **entity**, **agent** and **activity**. These classes are associated by nine relations such as *wasAttributedTo* and *wasInformedBy*. PROV-O provides additional subclasses and sub-properties that can be used to complement the initial terms and also to add more details among the relations -- and thus specialize it to distinct usage domains. Basically, the datasets that are submitted to a transformation process are instances of the **entity** class and the processes that modify and use the datasets are instances of the **activity** class. The entity responsible for commanding the execution of an activity is modeled as an **agent** class. Agents can also command other agents.

We implemented an instance of PROV-O that we call ProvenBiO -- A Provenance Biodiversity Ontology, available at http://purl.org/provenbio/ontology#. The goal of ProvenBiO is to preserve provenance information related to applications in the biodiversity domain and use this information to support the assessment of quality of data used and/or generated by domain experts. ProvenBiO adopts widely used vocabularies and ontologies (e.g., Dublin core (DCMI 2010), Geospecies (DeVries 2009), Darwin Core (DwC 2009)) aiming at enriching the provenance metadata with terms interesting to the biodiversity context.

Figure 1 illustrates a portion of a set of procedures and elements modeled in a ProvenBiO graph together with their corresponding RDF triples. The figure shows, for example, the properties that we adopted from PROV-O, describing the interaction among them (e.g., entity http://purl.org/fnjv/airtemperature/42 *provo:wasGeneratedBy* the activity http://purl.org/fnjv/activity/exploreEnvVar01). To better distinguish an activity that represents a concept (e.g., provenbio:bioSoftware) from an activity that was performed within a system (e.g., http://purl.org/fnjv/activity/exploreEnvVariable01/), it was necessary to add a new class. Thus, we specialize the class *agent* of PROV-O with a new class called *bioSoftware*. The figure also shows some terms such as *geo:lat* from GeoNames and *dc:identifier* from Dublin Core. In other words, PROV-O can be specialized and modified to meet distinct domain requirements.

CASE STUDY: USING PROVENBIO TO DERIVE DATA QUALITY

Motivating Scenario

The volume and variety of data types, their storage using different formats and through distributed repositories are common problems that hamper the assessment of the quality of data in the domain of biological diversity (Chapman 2005). To

[2] Fonoteca Neotropical Jacques Vieillard, Institute of Biology, UNICAMP

illustrate a scenario, we briefly describe some challenges concerning the management of sound recordings faced by FNJV

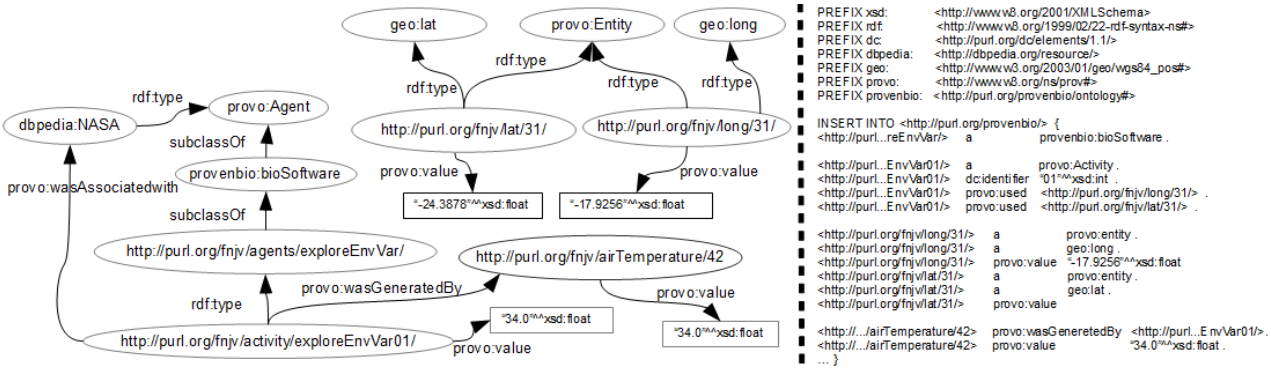

Figure 1 Example of a portion of our ProvenBiO ontology and the corresponding SPARQL query

(Cugler et al. 2012). FNJV maintains the largest collection of animal vocalization recordings in the Neotropics. In order to preserve these recordings, researchers have created a digital repository for them. Metadata are essential to manage recordings, and thus the quality of information provided by metadata has become a crucial issue. Problems found related to such metadata include, for instance, variety of formats, missing data values, abbreviations, misspellings, missing or wrong information about species location. Common data quality problems are related to completeness, accuracy and consistency of data. Our case study investigates the quality of such metadata, in particular after they have been curated and gone through several cleaning processes (and thus, how good are the processes that were run to improve metadata quality).

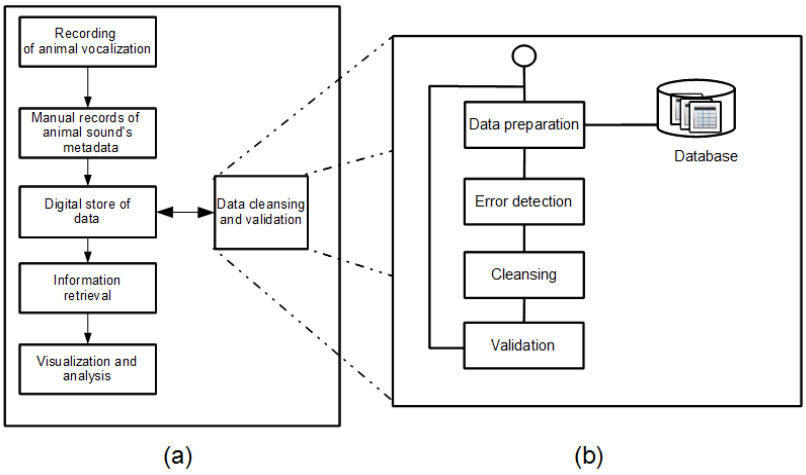

Figure 2 Basic flow concerning processing animal sound recordings - FNJV, inspired on (Cugler et al. 2012)

Figure 2 (a) depicts the basic process concerning the management of the recordings. First, biologists record animal vocalizations using distinct devices. Next, they write metadata in their notebooks (e.g., geographic location, scientific name, weather conditions) concerning the sound recorded and recording environment. Subsequently, all the metadata is stored in a database[3]. A data cleaning process follows this step. Finally, in order to perform scientific analyses, biologists query the database.

(Cugler et al. 2012) faced a subset of these problems by proposing an approach to fill missing metadata fields and derive such information automatically, from external Web sources. However, no evaluation about the quality of the original metadata and the derived datasets was performed. Taking this into account, we focus on the evaluation of data quality when the process to clean and fill missing metadata values is executed. Figure 2 (b) shows the general steps of the data cleansing process.

[3] *using a system developed by* (Cugler et al. 2012)

Figure 3 describes the workflow that is used by (Cugler et al. 2012) to fill missing metadata values. Notice that this is a generic workflow that can be specialized for domain-specific cleansing activities. In the case study, processing starts from the geographic region (usually a location) where the sounds were recorded, from which missing environmental information can be obtained. The location name metadata is used to query the Freebase knowledge base, in order to derive the latitude and longitude of the location informed[4]. Next, the latitude and longitude obtained are combined with stored metadata values "collect time" and "collect date" to be used as input to web services such as NASA's GLDAS and IRIS. These and other services are used to derive metadata on environmental variables at the time and location of the recording. We inserted probes in this workflow to capture provenance information at each stage of the workflow execution. This information is represented as instances of ProvenBiO.

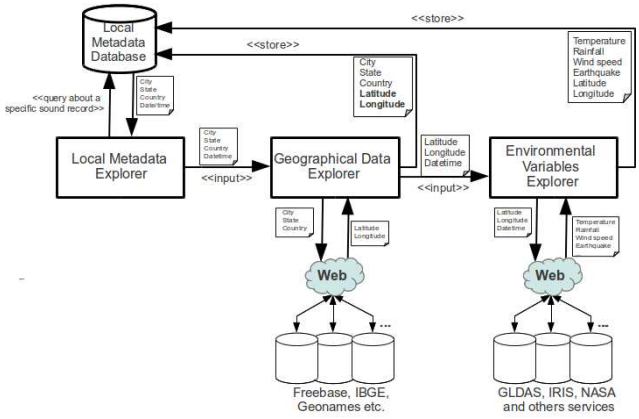

Figure 3 Workflow of the data cleaning activity, based on (Cugler et al. 2012)

ProvenBiO ontology: a running example

In this scenario, provenance plays an important role since biologists need to know how the fields were completed, and track the cleaning processes, users and resources in order to consume the data in their investigations. Typical questions that experts may ask are: "were these metadata fields filled by an expert or a novice user?"; "can I rely on the data collected from this specific source?"; "are the derived metadata complete enough?"

Figure 4 illustrates a portion of ProvenBiO RDF triples. This ontology is also a result of our previous experience in modeling provenance (Malaverri & Medeiros 2013). In the figure, triples correspond to the provenance information collected in one execution of the prototype shown in Figure 3. In the figure, resources and values are nodes and properties are edges. The figure shows, for instance, that there exist OWL classes that represent *Activities* and *bioSoftware* agents. The Activity *http://purl.org/fnjv/activity/exploreGeoData01/* has properties such as *startedAtTime*, *endedAtTime* and *wasAssociatedWith*, which hold the interval when the instance was executed and its associated agents like FreeBase. Furthermore, we also have the data produced by this activity - uniquely identified as *http://purl.org/fnjv/lat/31/* and *http://purl.org/fnjv/long/31/* with their respective values.

Capturing Provenance Information

The Provenance Manager is composed by a set of services that we implemented to allow to capture provenance information. Figure 5 depicts the elements - the Data Provenance Collector and the RDF Serializer services - that compose the Provenance Manager.

The *Data Provenance Collector service* is in charge of capturing the provenance metadata. The goal of this service is the extraction and classification of the metadata that keep track of the activities and entities that participate in the generation of missing values. The figure shows that the Data Provenance Collector service can be plugged in at each processing stage of the data cleaning system being monitored. Information such as people and tools that participated in the generation of a piece of missing data, details of processes, and parameters used by the processes, are collected when the system is executed.

[4] *Most records date back to the 70's, a pre-GPS era. If lat/long is available, this step is bypassed*

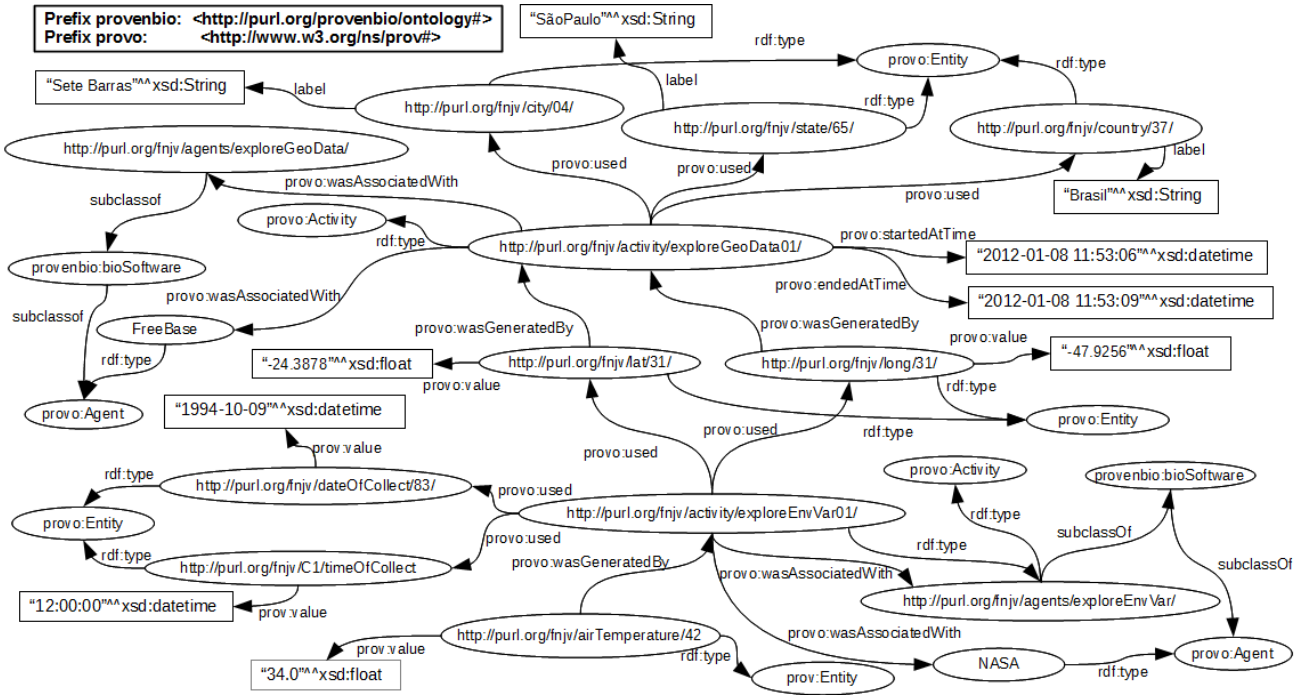

Figure 4 Example of RDF triples of ProvenBiO

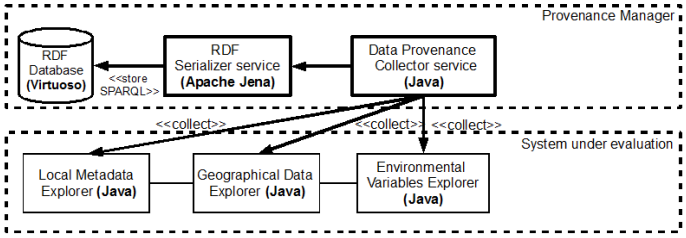

Figure 5 Elements that compose the Provenance Manager

Next, the information collected and classified by the Provenance Collector service is delivered to the *RDF Serializer service*. It takes the provenance information and submits it to a categorization process, where this information is mapped with a corresponding ontology term. We use the set of terms and properties defined in ProvenBiO to represent provenance at this stage. Once all information is instantiated, it is stored into the Provenance Repository in the format of RDF triples. We implemented these services using Java technology. Furthermore, we use the Apache Jena framework to build and write RDF triples which are stored into the Virtuoso database (via SPARQL queries).

Querying Data Provenance to Derive Quality

Let us now regard the workflow of Figure 2 (b) whose general goal is to perform data cleaning and fill in missing metadata values (using the workflow of Figure 3). Consider that an expert wants to know the quality of the datasets that resulted from the data cleansing process, so that (s)he can subsequently use such datasets. Using the strategy described in this work, users can pose queries against the RDF Database (our Provenance Repository), in order to retrieve information to estimate quality. For instance, imagine that the specialists are interested in evaluating the completeness and confidence of the datasets. Examples of queries that we can answer are:

1. *Search for the metadata records that were completed using data sources whose average reputation is higher than "X";*
2. *Retrieve metadata records for which the cleansing activity is over, and which started before a date "D";*

3. *Find newly completed metadata records for species found in tropical countries;*
4. *Retrieve newly completed metadata records that are related to endangered species;*
5. *Retrieve the identifiers of all databases whose reputation is higher than 0.6 and which were used in the workflow of* Figure 3 *to fill missing metadata for Passeriformes species recordings.*

Queries 1 and 2 are simple to solve in a relational database and are supported by other provenance. However queries 3, 4 and 5 are more complex and may involve further information and relationships that we can only solve using ProvenBiO ontology. Also notice that some queries are specific to the domain of our case study, while others can be considered in the context of generic information handling environments. Figure 7 shows the SPARQL query for item 5. The first lines shows that consensual vocabularies and ontologies like Geospecies and Dbpedia-owl were adopted. The second part concerns the query itself.

Once the information is delivered to the specialist, (s)he can apply specific rules to decide whether the data are good enough. We developed a prototype to query the data provenance captured by our Provenance Manager, available at http://purl.org/provenbio/?task=do/querynav. Related work, as discussed in the next section, considers only stored (meta) data. Our approach, on the other hand, allows finding additional information, which is obtained from relationships among stored data and ontologies. Thus, our provenance-based queries can return much more than information restricted to the stored provenance metadata. In other words, the results of these queries are data that can be analyzed by users to evaluate quality according to their criteria. Figure 6 presents a screen copy of our query prototype. It shows some basic queries that specialists can perform in order to evaluate the quality of their data.

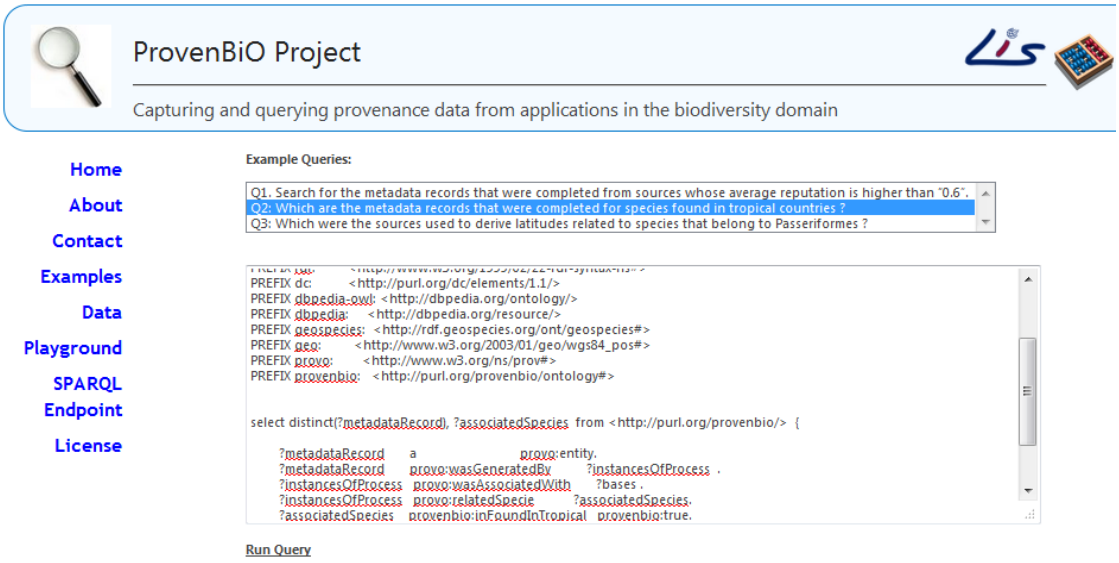

Figure 6 Screen copy of our query prototype. The code for Q2 is partially shown in the window.

RELATED WORK

Data quality is seen as a subjective concept. Frequently data considered good enough for a group of users can be considered bad for others (Simmhan & Plale 2011; Chapman 2005). Thus, the assessment of the quality of data needs to consider the characteristics of a specific context (e.g., e-Business, healthcare, environmental sciences). There are many research initiatives that tackle the assessment of quality by presenting methodologies to measure different data quality dimensions - e.g., (Pipino et al. 2002; Ballou et al. 1998). However relatively little work explores and applies the information produced when a dataset is generated - i.e. its provenance - as a key piece to evaluate the quality of data.

(Simmhan & Plale 2011) describe an approach for personalized quality scoring to rank scientific datasets based on a quality profile. Provenance metadata is used to model a quality function based on weights setting on a user's quality profile. Machine learning techniques are used to construct a quality function to produce a quality score. The main idea behind this solution is to predefine quality scores of the input data to map to the quality score for the derived output data. Although our solution can use the expertise of specialists to annotate quality scores of input data (e.g., confidence of a data source), we believe that this kind of approach can be time consuming - the broader the application domain is, the greater the effort to configure a quality

profile. Rather than relying on (manual) user-assigned scores, our approach tries to automatically get as much information as possible that is produced when a dataset is generated to be used by the specialists in the quality assessment process. Moreover, we do not compute quality scores. Rather, it is up to the user to derive information (s)he considers useful to obtain quality information. Though this is an ad hoc process, on the other hand users are free to investigate any quality criterion of their choice.

In order to compute a quality score that can to be used in the evaluation of the quality of data on the Web, (Hartig & Zhao 2009) describe a solution to annotate provenance metadata (e.g., date of creation) with impact values. The provenance model constructed is directly associated with the timeliness quality dimension. Unlike this work, we do not need to specialize provenance for each dimension of quality. In our case, the specialist can choose the quality dimensions of interest and request for information that can help to assess the dimensions.

Similar to (Hartig & Zhao 2009), (Prat & Madnick 2008) also propose an approach to compute the believability quality dimension 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 measure believability, the authors only measure the believability of a numeric data value, which limits its applicability.

```
PREFIX rdfs:       <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:        <http://www.w8.org/2001/XMLSchema>
PREFIX foaf:       <http://xmlns.com/foaf/0.1/>
PREFIX rdf:        <http://www.w8.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc:         <http://purl.org/dc/elements/1.1/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia:    <http://dbpedia.org/resource/>
PREFIX geospecies:  <http://rdf.geospecies.org/ont/geospecies#>
PREFIX geo:        <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX provo:       <http://www.w8.org/ns/prov#>
PREFIX provenbio:   <http://purl.org/provenbio/ontology#>


select * from <http://purl.org/provenbio/>  {

?databaseAgent          a                           provo:Agent.
?databaseAgent          a                           foaf:Organization.
?databaseAgent          a                           provenbio:publicDataOrganization.
?databaseAgent          provenbio:trustScore        ?trustScore

?swagent                a                           provenbio:swagent.
?swagent                provenbio:isAssociatedWith  <http://purl.org/fnjv/agents/nasa/>

?activityInstance       provenbio:bioSoftware       ?swagent
?activityInstance       a                           provenbio:ActivityInstance
?activityInstance       provo:startedAtTime         ?startDateTime
?activityInstance       provo:wasAssociatedWith     ?databaseAgent.
?activityInstance       provo:used                  ?instanceInputSpecies.

?instanceInputSpecies   a                           geospecies:SpeciesConcept
?instanceInputSpecies   geospecies:inOrder          <http://lod.geospecies.org/orders/hNvZJ>
?instanceInputSpecies   geospecies:hasLocation      ?speciesCountry
?speciesCountry         dbpedia-owl:country         dbpedia:Brazil

FILTER (?trustScore > 0.6)    }
```

Figure 7 SPARQL query corresponding to Item 5

Notice that one singular characteristic of our work is that we generated an instance of a generic ontology for provenance representation. This ontology allows to collect information related to a specific domain and store data provenance that is used to assess quality in a specific context.

CONCLUSIONS AND ONGOING WORK

This paper presented an approach to support specialists in the estimation of quality of datasets based on provenance information, for data-intensive applications. Rather than concentrating our study on standard organizational environments, we analyze environments in which scientific experiments are planned, specified and executed, insofar as they reflect a particular set of procedures and processes to run experiments. In order to provide domain-specific provenance, we generated an ontology instance (ProvenBiO) based on the W3C PROV-O ontology and data model. Besides typical queries focused on

provenance from a system point of view (e.g., processes), this solution enables specialists to investigate relationships among elements within a specific domain. Aiming at the expressiveness of ProvenBiO, we aggregated widely adopted vocabularies such as DwC and Geospecies. This enhances interoperability across distinct groups that want to share and reuse data sets in their processes.

In particular, we use the provenance information to allow experts to perform queries aimed at assessing the quality of data. Distinct members/roles in a given group or organization can be interested in different dimensions of quality, depending on the kind of activity that they are performing. For this reason, the automation of the measurement of quality can be a difficult task, especially if we consider that each dimension of quality may cover other sub-dimensions.

Our solution was validated using a case study concerning recordings of animal sound vocalizations. We implemented a set of services that enable to capture and identify provenance metadata when a system is being executed, and a service that allows to query this information. Future work that we want to investigate is related to the propagation of data provenance among the transformation processes through which a dataset is submitted. Another extension is related to the analysis of provenance as a criterion to adapt a workflow to a specific organizational context. Our queries require knowledge of SPARQL. Future work also involves developing an interface with translation mechanisms to transform user requests into SPARQL queries.

We point out that though our work is concerned with scientific processes and data, it is generic enough to be applicable to other organizational contexts. It suffices to adapt the ontology to contemplate concepts and relations of the domain of interest.

ACKNOWLEDGMENTS

This paper was partially funded by grants from CAPES and CNPq, by the Microsoft Research-FAPESP Virtual Institute NavScales project), CNPq (MuZOO Project and PRONEX-FAPESP), and by the INCT in Web Science (CNPq 557.128/2009-9)

REFERENCES

Ballou, D. et al., 1998. Modeling Information Manufacturing Systems to Determine Information Product Quality. *Manage. Sci.*, 44, pp.462–484.

Barga, R.S. & Digiampietri, L.A., 2008. Automatic capture and efficient storage of e-Science experiment provenance. *Concurr. Comput. : Pract. Exper.*, 20(5), pp.419–429.

Batini, C. & Scannapieco, M., 2006. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*, Springer-Verlag.

Blake, R. & Mangiameli, P., 2011. The Effects and Interactions of Data Quality and Problem Complexity on Classification. *Journal of Data and Information Quality*, 2(2), pp.1–28.

Chapman, A.D., 2005. Principles of Data Quality. *Global Biodiversity Information Facility, Copenhagen*.

Chen, P., Plale, B. & Aktas, M.S., 2012. Temporal Representation for Scientific Data Provenance. In *Proc. 8th IEEE Int. Conf. on eScience 2012*.

Cugler, D.C., Medeiros, C.B. & Toledo, F., 2012. An architecture for retrieval of animal sound recordings based on context variables. *Concurrency and Computation - Practice and Experience*.

Davies, J., Studer, R. & Warren, P., 2006. *Semantic Web Technologies: Trends and Research in Ontology-based Systems*, Wiley.

DCMI, 2010. The Dublin Core Metadata Initiative. Available at: http://dublincore.org/.

DeVries, P.J., 2009. GeoSpecies Ontology. Available at: http://bioportal.bioontology.org/ontologies/1247.

DwC, 2009. Darwin Core Task Group. Available at: http://www.tdwg.org/standards/450/.

Goodchild, M.F. & Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1, pp.110–120.

Hartig, O. & Zhao, J., 2009. Using web data provenance for quality assessment. In *Proc. of the Workshop on Semantic Web and Provenance Management at ISWC*.

Kepler, 2011. The Kepler Project. Available at: https://kepler-project.org/.

Kondo, A.A. et al., 2007. Traceability in Food for Supply Chains. In *Proc. 3rd Int. Conf. on Web Information Systems and Technologies (WEBIST)*. INSTICC, pp. 121–127.

Lassila, O. & Swick, R.R., 1999. Resource Description Framework (RDF) Model and Syntax Specification.

Malaverri, J.E.G. & Medeiros, C.B., 2013. A Provenance-based Approach to Evaluate Data Quality in eScience. *Submitted to Int. J. Metadata, Semantics and Ontology - Special Issue on "Metadata for e-science and e-research."*

Moreau, L. et al., 2011. The Open Provenance Model core specification (v1.1). *Future Generation Comp. Syst.*, 27(6), pp.743–756.

Parssian, A., 2006. Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. *Decis. Support Syst.*, 42, pp.1494–1502.

Pernici, B. & Scannapieco, M., 2002. Data Quality in Web Information Systems. In *Proc. of the 21st Int. Conf. on Conceptual Modeling*. Springer-Verlag, pp. 397–413.

Pipino, L.L., Lee, Y.W. & Wang, R.Y., 2002. Data Quality Assessment. *Commun. ACM*, 45, pp.211–218.

Prat, N. & Madnick, S., 2008. Measuring Data Believability: A Provenance Approach. In *Proc. of the 41st Hawaii Int. Conf. on System Sciences*. p. 393.

Richard, Y. & Diane, M., 1996. Beyond accuracy : What data quality means to data consumers. *Journal of Management*.

Sahoo, S.S., Sheth, A.P. & Henson, C.A., 2008. Semantic Provenance for eScience: Managing the Deluge of Scientific Data. *IEEE Internet Computing*, 12(4), pp.46–54.

Simmhan, Y. & Plale, B., 2011. Using Provenance for Personalized Quality Ranking of Scientific Datasets. *I. J. Comput. Appl.*, 18(3), pp.180–195.

Taverna, 2009. The Taverna Project. Available at: http://www.taverna.org.uk/.

VisTrails, 2011. The VisTrails Project. Available at: http://www.vistrails.org.

W3C, 2012. The PROV Ontology. Available at: http://www.w3.org/TR/prov-o/.

Wang, X., Gorlitsky, R. & Almeida, J.S., 2005. From XML to RDF: how semantic web technologies will change the design of "omic" standards. *Nat Biotech*, 23(9), pp.1099–1103.

Yeganeh, S.H., Hassanzadeh, O. & Miller, R.J., 2011. Linking Semistructured Data on the Web. In *Proc. 14th Int. Workshop on the Web and Databases*.

Zhao, J. et al., 2008. Mining Taverna's semantic web of provenance. *Concurr. Comput. : Pract. Exper.*, 20, pp.463–472.