A provenance-based approach to manage long term preservation of scientific data

Renato Beserra Sousa*, Daniel Cintra Cugler†, Joana Esther Gonzales Malaverri‡, Claudia Bauzer Medeiros§
Institute of Computing
Unicamp
Av. Albert Einstein, 1251, Campinas/SP - Brasil
* renatobeserra@gmail.com
† danielcugler@ic.unicamp.br
‡ jgonzalesm1@gmail.com
§ cmbm@ic.unicamp.br

***Abstract*—Long term preservation of scientific data goes beyond the data, and extends to metadata preservation and curation. While several researchers emphasize curation processes, our work is geared towards assessing the quality of scientific (meta)data. The rationale behind this strategy is that scientific data are often accessible via metadata - and thus ensuring metadata quality is a means to provide long term accessibility. This paper discusses our quality assessment architecture, presenting a case study on animal sound recording metadata. Our case study is an example of the importance of periodically assessing (meta)data quality, since knowledge about the world may evolve, and quality decrease with time, hampering long term preservation.**

I. Introduction

Scientific data is not only a product of research, but a foundation thereof. The relevance of long term preservation of scientific data is fast becoming one of the main concerns of large research initiatives. This can be seen, for instance, in the Status Report concerning preservation of High Energy Physics Data [1]. Though geared towards such data, it also briefly discusses the relevance of preservation of scientific data in other disciplines, namely astrophysics, molecular biology, geosciences, and humanities and social sciences. The text includes a survey that shows that over 70% of over 1000 surveyed physicists consider preservation highly important to crucial to their work.

Curation and preservation are distinct concepts, but are intimately connected. Perhaps the best definition of *curation* that considers these associations is given by the US Council on Library and Information resources (http://www.clir.org/): “the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, *maintain its quality*, add value, and provide for *reuse* over time, and ... include authentication, archiving, management, preservation, retrieval, and representation". This definition presents many reasons for preservation, and *reuse* is the one we will concern ourselves with, especially since reuse implies adequate means to discover and retrieve. We also single out another factor, *quality maintenance*, which we exploit in this paper.

We point out that this definition is stated by research librarians, whose main traditional concern are publications. However, librarians are increasingly concerned with data generated by research, and treating such data as a publication in itself. This point of view is being widely adopted in North America, e.g., as a consequence of NSF’s policy concerning compulsory Data Management Plans. This paper adopts this definition, and considers curation and preservation as activities that cannot be dissociated. Nevertheless, there are two additional issues that must be considered together with this definition:

- What is *scientific data*?
  This paper is concerned with digital scientific data. This covers in principle any kind of digital artifact that is associated with scientific research – e.g., spreadsheets, streams, digitized paper sketches, photos, recordings, and so on. Data concerning an experiment must also include the workflows needed to run that experiment, associated annotations and so on. The documentation of an experiment, including methodology, instrument specifications, and others, should also be considered as part of the “data" associated with the experiment. This paper is restricted to scientific data that are input or output of a given scientific experiment.
- What does *long term preservation* mean?
  According to the definition adopted, long term preservation must also ensure long term curation, e.g., to reflect new knowledge about the world. The definition mentions “data’s life cycle as long as it is of interest to scholarship"; this also poses an interesting challenge. Some data sets may have cyclic scholarly interest, losing and regaining value as research material from time to time. This ultimately means that every kind of scientific data must be curated forever, in case it needs to be reused sometime. Again, we simplify this issue and consider preservation and curation as part of a continuum in which scientists define which data sets to preserve, and the desired preservation period (i.e., with associated lifetime).

These issues notwithstanding, preservation can be seen as a multi-level concern, as for instance illustrated in the classical case study of the Domesday Project [2]. That multimillion pound project, undertaken to create a modern day electronic portrait of Great Britain in the 80’s, was already obsolete at its end, for lack of, among others, appropriate curation and preservation procedures. Analyzing the report on that project, one can see that preservation for usability means preserving not only the data, but also the environment needed to process it (including hardware and software), as well as people trained to use the particular hardware and software. Such problems can be alleviated by initiatives such as adoption of appropriate standards, software design methodologies, choice of adequate metadata, and continuous backing up and porting of data (and software) to new media and devices.

Given this complex scenario, this paper presents an approach to deal with just one such issue, namely *curation of metadata associated with scientific data* to help long term preservation of the associated scientific data. This work is based on our 6 year experience with managing and curating data and metadata associated with biodiversity research, together with scientists from Institute of Biology at the University of Campinas. As will be seen in the discussion of our case study, our procedures for metadata cleaning and curation allowed these scientists to better evaluate the quality of their primary data sets, as well as suppported more sophisticated data analysis procedures. This, in turn, has helped in ensuring longevity of their datasets, in two senses: first, these data sets will be valid and reusable for a longer time span; second, this has prompted tuning up methodologies for data annotation and curation, again intrinsically associated with preservation processes.

In particular, this paper discusses an architecture to support scientists in assessing the quality of their curated datasets. This architecture, partially implemented by [3], is extended in this paper to allow customization of quality dimensions and of the evaluation of metadata quality. The rest of this paper is organized as follows. Section II presents some related work and concepts. Section III presents our architecture. Section IV presents a small case study to show our preliminary results. Finally, section V presents conclusions and ongoing work.

II. Background and Related Work

This paper is centered on handling quality of metadata associated with scientific data, to enable sustainable reuse of scientific data. In this section, we discuss some general approaches for data preservation, and briefly consider related work on data quality. Finally, we give an overview of the kinds of data handled in our experiments, which concern animal sound collections.

A. Data Preservation

Digital preservation is defined in the site of the US Library of Congress as “the active management of digital content over time to ensure ongoing access". In this direction, Conway et al. [4] argue that the data preservation challenge encompasses the ability to “equip future users with the necessary information to reuse the data, and thus include appropriate metadata. They propose a preservation analysis methodology that is organized in the following steps: preliminary investigation of data holdings; stakeholder and archive analysis; defining a preservation objective; defining a designated user community; defining preservation information flows; and analysis of costs, benefits and risks. We point out that the methodology stresses the importance of defining the preservation objective. Indeed, depending on this objective, preservation will encompass distinct activities. This is made clear in the already mentioned DPHEP report [1], which indicates that there are four kinds of preservation models, as stated in Table I, in which level 1 is the least complex to achieve, and level 4 the most complex. This paper concerns level 1.

TABLE I. Preservation models for scientific data, from [1]

Preservation Model	Use Case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Preservation for accessibility is also stressed by [5], as part of an initiative of the Ontario Council of University Libraries [6] to offer a common platform to serve publications and scientific data, including journal papers, books, geospatial data and social science data sets. Examples of other data preservation initiatives are, for instance, Research Data Canada (which also emphasizes the importance of openness and of research data as a public good) [7], and the US Library of Congress Digital Preservation Program (where providing access to current and future generations is stressed) [8]. All such initiatives are concerned with metadata standards and metadata quality as a means of ensuring access and reusability.

Reproducibility also requires apreciation. In this context, Bechhoffer et al. [9] discuss the problem of reproducible research on eScience. They reason that although Linked Data [10] provides an infrastructure for publishing and sharing scientific results, it fails to reflect the research methodology and to respect the rights and reputation of the researcher. They propose considering Research Objects (RO), defined as “semantically rich aggregations of resources that bring together the data, methods and people involved in (scientific) investigations" [9]. They argue that the RO approach can improve the way we conduct scientific research, by providing an infrastructure that enables reproduction of experiments. Their infrastructure relies upon provenance of data to audit data validity and to consider intellectual rights of data providers. We also consider provenance, as a means to help improving (meta)data quality.

B. Data quality assessment

Traditionally, data quality is defined as “fitness for use" and describes it as a multi-dimensional concept. A *quality dimension* can be defined as a set of data quality attributes that allow to represent a particular characteristic of quality [11]. In the literature, accuracy, completeness, timeliness and consistency have been extensively cited as some of the most important quality dimensions for information consumers [12] [13]. However, depending on the application scenario, other quality dimensions may be more adequate. For instance, correctness, reliability and usability are interesting in areas such as simulation processes, as discussed in [14]. Approaches to assess data quality vary from the definition of models to implementation of platforms [3].

Examples of data quality approaches that involve some sort of implementation include the work of Gamble and Goble [15], Chiang and Miller [16] and Lemos [17]. Gamble and Goble [15] focused on giving users support to data quality assessment, regarding scientific data available on the web. It emphasized the separation between the dimensions of trust, quality and utility, defining them as entities for the computation of data quality. Akin to our approach, they defined a mechanism to combine the assessment of the different dimensions by using provenance information. That work uses decision networks to assess data quality. In the network they will typically model the decision of accepting or rejecting the data, according to quality and trust dimensions. The output of the decision network is a utility index that can be used for scoring and ranking data.

Chiang and Miller [16] proposed a data quality paradigm based on the management of rules for data quality. They discuss that data quality rules become invalid as applications evolve and thus it is necessary to repair the rules. Their approach is based on a repair model that recommends a set of attributes to add to a data rule, so that the rule stay consistent within the dataset. That work presented a tool that identifies conditions for which some data rule holds for a subset of data.

Another approach for data quality assessment is provided by the Qbox-Foundation [18], whose goal is to smooth the definition of appropriate metrics for measurement methods, according to the specific uses of an organization. Extending that approach, Lemos [17] proposed a new quality metamodel and developed a service oriented computational infrastructure through which data quality can be measured according to multiple dimensions, given user scores and definitions. The input is based on the definition of quality goals and a set quality metrics, and a set of services that compute these metrics. This platform is an interesting example of how quality can be assessed differently by distinct sets of users, who tailor metrics according to their quality goals.

As will be seen in Section III, our solution is based on assessing quality using provenance information. In this paper, data provenance corresponds to the origins and the transformation processes applied to a dataset until it is (re)used in some experiment. Though provenance attributes can be considered as dimensions that contribute to quality, such attributes are treated apart in quality assessment (e.g., as historical information). In other words, related work either considers provenance to assess quality (which we call provenance-based) or disregards it, considering other attributes (a trend we call attribute based). Under this perspective, our work can be considered as provenance based.

Provenance, in our work, is extracted from workflow execution. A similar research appears in Naim et al [19]. Their paper describes a mechanism for assessment of data quality during workflow execution. There is a component that provides a real-time monitor for data quality. The user can define some thresholds for data quality, so that the monitor shows the level of data quality based on intermediate results. We, instead, instrument the workflow so that it can incorporate data quality information during execution.

C. *Biodiversity Observations and Metadata*

An observation [20] represents an assertion that a particular entity was observed and that the corresponding set of measurements were recorded (as part of the observation). Data in observation databases can be very heterogeneous, and concern observations at multiple spatial and temporal scales, including images, maps, sounds, texts and so on. Observation records may constitute the primary data of a scientific experiment (e.g., when the description of the observed entity is the only fact that can be retained, since the entity itself cannot be stored or preserved). In many cases, however, these records correspond to metadata concerning data generated by the experiment (e.g., in many biodiversity studies, or in high energy physics experiments [1]).

Even though there is no consensual standard on defining metadata fields for observation records, most have a common subset of fields, indicating what was observed, when, where and by whom. Also, depending on the kind of observation conducted, additional metadata fields indicate details on observation methodology (how), or devices used in the observation.

Our case study concerns observations for biodiversity studies, and deals with sound recordings. Here these recordings are the primary stored observation data, and are stored together with metadata concerning such observations. We point out that we have worked with other kinds of biodiversity observations, e.g., animals in museum collections, but this case study, as will be seen, can be used to single out some challenges in long term preservation.

In biodiversity studies, there is growing interest in sound recordings. Several organizations around the world maintain extensive animal sound collections [21], [22], providing information not only about species but also about the environment in which they live. Such collections differ primarily in their number of recordings, the kind of species they have recorded and methods used to obtain recordings. Most of those collections have associated metadata. Such information is widely used in animal habitat prediction, detection of spatial patterns, dynamics of populations, animal conservation, and so on.

Earlier animal recordings were commonly stored in magnetic tapes, requiring special attention to be kept clean, free of humidity and fungus infection. More recently, recordings use devices that save data in a variety of digital formats, such as ATRAC, AIFF, WAV and MP3, contributing to the heterogeneity problem [23]. Once the recordings are stored in a repository in digital format, there begins a new series of data management processes, and associated challenges.

There are two major means of retrieving information from such vocalization databases. One approach is retrieval based on the analysis of acoustic features - e.g., by exploiting the physical properties of sound waves [24]. However, acoustic properties of animal sounds vary widely, hampering this kind of retrieval. Another way is to query metadata, usually posing queries on fields such as species taxonomy, and location where the sound was recorded. Queries on metadata are limited to the stored fields, which are often incomplete or blank. Moreover, there is additional relevant information that is not explicit in the recording metadata and that is part of the context in which the sound was recorded - concerning engineering and biological/environmental variables. Our case study concerns

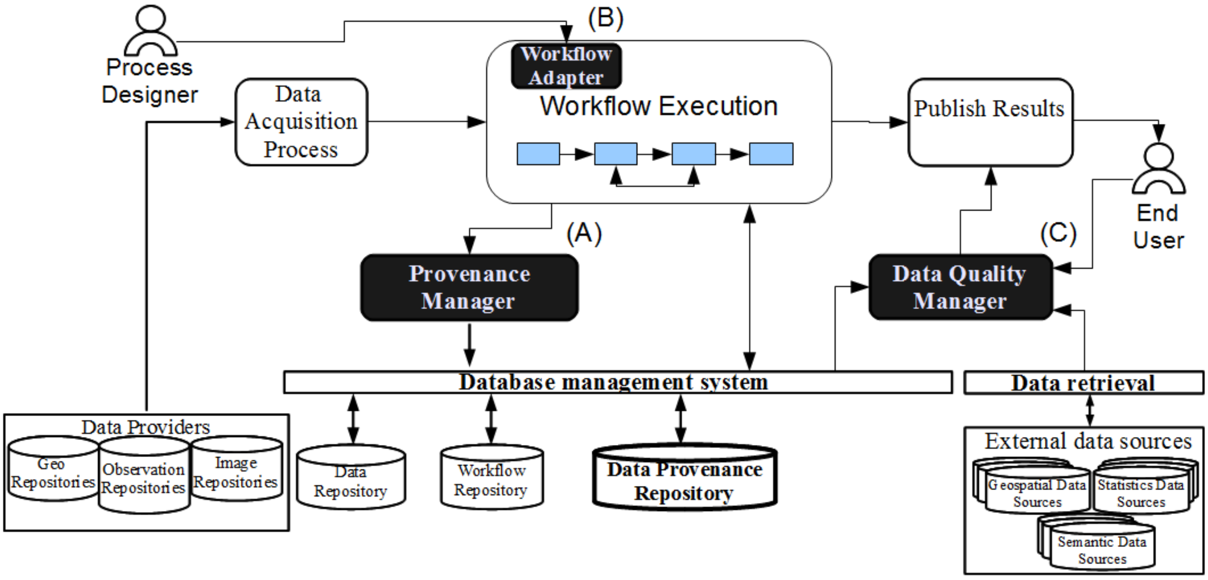


Fig. 1. System architecture

cleaning and curating sound metadata as a means to help long term preservation of the recordings.

III. Deriving data quality using provenance information

The main concept behind this paper is that good quality metadata is necessary for long term preservation. Quality, however, depends on the users and context of use. Thus, metadata must go through some sort of quality evaluation process, in which data curators and consumers define the quality dimensions and metrics of interest, which can also vary as time passes. Figure 1 shows the architecture that enables this process, and which extends the architecture proposed by [3], in particular, the boxes in black.

We point out that this approach can apply to help assess quality of scientific data (and not just metadata). Thus, the boxes in the figure refer to “data", since this can be used on both data and metadata. In the rest of this section we will thus talk about “data" curation and processing, since this is the general idea, but the focus of the paper is metadata quality management.

The main idea is the following: scientific data is acquired and goes through some sort of process (here, assumed to be executed via a scientific workflow). Workflows are made available via a workflow repository. Two factors are needed to help assess the quality of the published results: (a) information about the original metadata and (b) information about the process it went through. I.e., this is a provenance-driven quality assessment process (origins and transformations together constitute *provenance* information).

The Workflow Adapter is a module that allows experts to add quality information to a workflow specification. Quality aware workflows incorporate quality metadata (about data processed and the processes embedded in the workflow). Here, quality information comes from two sources - provenance from running the workflow and tasks incorporated in the workflow that produce additional quality information while the workflow is running (e.g., reputation of a data source or reliability of a process). The Adapter allows experts to integrate metadata quality information, without changing the workflow model.

During workflow processing, the Provenance Manager extracts provenance information from data and workflows, storing such information in the Data Provenance Repository. The database management system provides access to data, workflow and provenance repositories. Such workflows can invoke external and local modules, and web services, each of which may have specific quality properties.

The Data Quality Manager is responsible for assessing data quality, based on expert requirements. This module generates quality information from: (a) the provenance information stored by the Provenance Manager, (b) the quality attributes added to workflows by the Workflow Adapter and (c) external data sources. Quality metrics are computed as defined by end users (scientists).

The Data Quality Manager can also look for information from external semantic data sources to complement the facts provided by the repositories. Semantic data sources are all kinds of sources in which data is stored together with means to attach semantics to it - eg., including a formal description of concepts, terms, and relationships within a given knowledge domain.

We consider two distinct user roles:

(a) **Process Designer**: an expert responsible for specifying some workflow. In particular, (s)he may embed in the workflow additional functionality to allow extraction of quality dimensions during workflow execution. Thus, the workflow produces quality attributes as byproducts of its execution. Here, the designer uses the Workflow Adapter to create such workflows, or to enhance existing workflows with extraction of quality attributes.

(b) **End User**: a scientist who is interested in the data resulting from workflow execution. (S)he will interact with the Data Quality Manager to define the quality dimensions of interest and means to compute quality metrics. Obviously, not all quality dimensions requested by the end user may be available. The results of quality assessment are published in two formats: (i) the workflow trace; and (ii) computed quality attributes.

The work of Malaverri et al. [25] shows a partial implementation of the architecture, without boxes (B) and (C), for an agriculture application, using Java and the Taverna Workflow System. Taverna exports the provenance into the OPM model [26], which is then mapped into a quality model.

IV. Case study

Our work is geared towards supporting metadata-based retrieval, where long term accessibility is associated with long term metadata curation. We exploit two new directions:

1) Improving quality by deriving and/or checking the contents of metadata fields using external authoritative sources;
2) Enhancing preservation by extending the set of metadata attributes, not originally contemplated by the scientists, thereby enhancing the scope of queries that can be supported, and increasing the chances of reuse of the associated data sets.

A. Long term preservation of the animal sound collection at FNJV

Our work is motivated by the challenges faced by the Fonoteca Neotropical Jacques Vielliard (FNJV) at the University of Campinas [27]. FNJV is one of the 10 largest animal sound collections in the world [28]. It has recordings of all vertebrate groups (fishes, amphibians, reptiles, birds and mammals) and some groups of invertebrates (as insects and arachnids). Most of the sounds were recorded by domain experts, who often annotated metadata at recording time.

Our case study addressed the needs of curators of FNJV. Here, researchers have amassed the largest collection of animal sound recordings in the Neotropics [1]. The recordings are offline, and only recently has there been a concentrated effort in publishing the corresponding metadata, and converting them to digital media [30] [27]. Since the core of the collection dates back to the 1960's, it provides several challenges for long term preservation.

Animal sound recordings present interesting preservation and curation challenges, in particular those associated with legacy collections, as is our case. First, they are often made under difficult conditions, with microphones and sensors installed in natural habitats, with background noise. For this reason, many animals appear in a single recording, there being a need to identify individuals (or at least individual species). Even if one assumes that recordings are performed under ideal conditions, vocalizations are very much sensitive to a wide range of contextual variables - e.g., time of the year, geographic distribution or even social aspects [31].

Cleaning and curating metadata is not restricted to syntactic or semantic analysis. These are not, moreover, isolated activities that are performed only once. Ideally, one should periodically revisit the past and compare what was known to what has been discovered since. This new knowledge should be somehow incorporated into metadata, to allow data analysis under original scenarios (i.e., when the data was collected), but also under evolving conditions. Curated (meta)data that in the past was reliable may have its content degraded with time. Degradation is not only physical but new discoveries may invalidate (meta) data. The (bad) news here is that most of the times data has no predictable expiration date, requiring collections to be constantly curated in order to be preserved and kept reliable.

For example, due to new discoveries, species names (observation metadata) can change along time, e.g., species *Elachistocleis ovalis* has had its name changed to *Nomen inquirenda* [32]. This is a common problem faced by biological collections, which hampers their long term preservation. Maintaining such data up to date and consistent is extremely important, since metadata errors regarding a single species can affect the understanding not just of the species, but of wider ecological interactions. Our case study focused on improving long term preservation of metadata species names, repairing names which evolved in time.

Table II shows 22 (out of 51) metadata fields that are present in the FNJV collection. Row 1 gives information to identify the recorded species (what was observed). Row 2 describes observation conditions – when, where and the environment in which the sound was recorded. Row 3 describes the recording features, as well as devices used to record them (how a recording was made).

TABLE II. Subset of metadata fields of the FNJV collection.

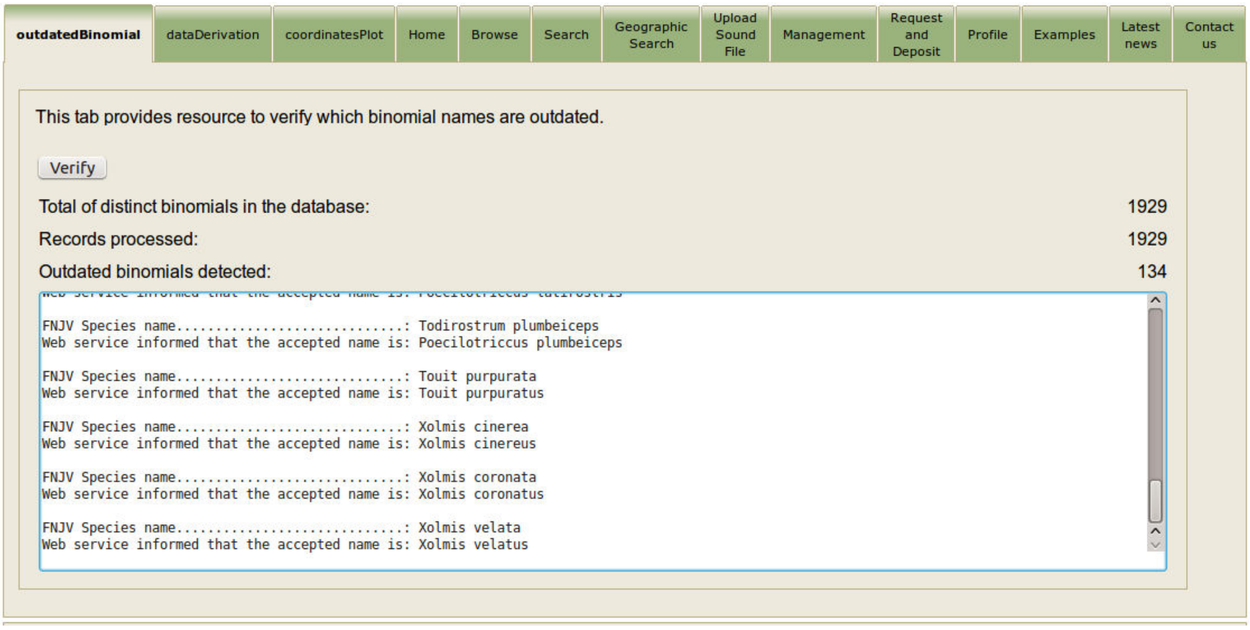
	METADATA FIELD
1	Phylum, Class, Order, Family, Genus, Species, Gender, Number of individuals.
2	Collect time, Collect date, Country, State, City, Location, Habitat, Micro-Habitat, Air temperature (°C), Atmospheric conditions.
3	Recording device, Microphone model, Microphone model, Sound file format, Frequency (kHz).

B. Prototype and Results

Our work towards improving metadata quality started in 2011 and was conducted in two stages, concentrating in fields of rows (1) and (2) of Table II.

The first stage, partially reported in [33], performed three kinds of metadata curation. The first concerned basic metadata cleaning algorithms, e.g., checking attribute domains, and syntactic corrections. Once these tasks were performed, the second curation step was to add geographic coordinates to all metadata records (since most recordings had been made before the advent of GPS). Finally, in the third step, we filled in missing fields whenever possible, in particular those concerning environmental conditions (e.g., humidity or temperature),

[1] The Neotropical region is one of the six biggest biogeographic areas in the world, defined based on its animal life features. It extends from the Southern Mexico Desert into South America [29].



This tab provides resource to verify which binomial names are outdated.

Total of distinct binomials in the database: 1929

Records processed: 1929

Outdated binomials detected: 134

```
Web service informed that the accepted name is: Poecilotriccus latirostris

FNJV Species name............................: Todirostrum plumbeiceps
Web service informed that the accepted name is: Poecilotriccus plumbeiceps

FNJV Species name............................: Touit purpurata
Web service informed that the accepted name is: Touit purpuratus

FNJV Species name............................: Xolmis cinerea
Web service informed that the accepted name is: Xolmis cinereus

FNJV Species name............................: Xolmis coronata
Web service informed that the accepted name is: Xolmis coronatus

FNJV Species name............................: Xolmis velata
Web service informed that the accepted name is: Xolmis velatus
```

Fig. 2. Prototype for detection of outdated species names.

obtained from authoritative sources, once location and date were defined. Every step was validated by human curators, who also helped in disambiguating information whenever our algorithms found problems (for instance, to define coordinates when a location name was too vague).

The second stage, reported in [34], was geared towards using spatial analysis to check errors. Examples of errors found included misidentified species and discovery of possible new species' behavior.

The case study reported in this paper is part of the first stage, and was implemented in the FNJV web site environment [30]. We point out that though the first stage concerns “basic" metadata curation, this is an ongoing (long term) concern. Thus, even though the “first" stage was initially finished in 2011, it was reinitiated in 2013, given preservation requirements. The results reported here correspond to part of a new implementation effort (validated by experts in October 2013). Our goal was to detect outdated species names by contrasting such names with authoritative organizations which publish and maintain official species names lists, in our case the Catalogue of Life [35].

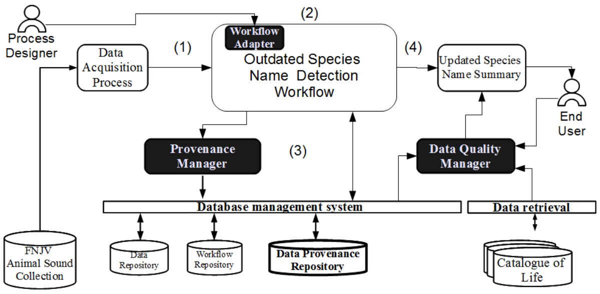


Fig. 3. Architecture instance for the case study

For this kind of metadata curation activity, experts are interested in finding out the accuracy of the original metadata. Figure 2 shows a partial screenshot of the results.

Given a species name, if it is no longer valid, the Catalogue of life web service informs what is the current up to date species name used. For each record in which the species name was detected as outdated in the FNJV database, the prototype persists the updated species name in a separate table and creates a reference between the original metadata record and the species name. This strategy is important in order to maintain the original collection unchanged – given, for instance, that several papers concerning that recording and the outdated name have been written. It also provides a historical log of metadata modifications. Before such names are persisted in the database, they are flagged to be checked by biologists. The prototype shows the progress of name checking, publishing the number of distinct species names in the database, the number of records processed, the number of species names which were detected as outdated and the respective updated names.

These experiments were executed over a total of 11898 records, with 1929 distinct species names analyzed. As shown in Figure 2, 134 distinct species in the collection (7% of the species analyzed) had their scientific names changed along time. The whole process takes a few minutes. Before the implementation of our prototype, such kind of verification was performed manually by biologists, taking from days to months, depending on the species chosen. Now, this verification can be performed frequently, ensuring that the species names will always be up to date, contributing to its long term preservation.

Listing 1. Excerpt from workflow description file

```

<processor>
<name>Catalog_of_life</name>
<annotations>
 ...
```

```
6  <text>
7    Q(reputation): 1;
8    Q(availability): 0.9;
9  </text>
10       </annotationBean>
11       <date>2013-11-12 19:58:09.767 UTC</date>
12       <creators />
13       <curationEventList />
14     </net.sf.taverna.t2.annotation.
         AnnotationAssertionImpl>
15   </annotationAssertions>
16 </annotations>
```

C. Quality assessment using the architecture

We now analyze the case study running in our architecture.

Figure 3 shows the instantiation of our architecture for this case study.

The metadata curation process follows these steps:

1) Experts added quality metadata to the workflow/
2) The Outdated Species Name Detection Workflow receives FNJV sound metadata as input;
3) The workflow checks for outdated names, using the Catalogue of Life external data source;
4) The Provenance Manager stores provenance information from the data source, workflow description and execution logs;
5) The workflow output is a summary of updated species names (see Figure 2).

The workflow (code implemented in Java) was run using the Taverna [36] workflow management system. The Workflow Adapter is being implemented. At the moment, we take advantage of Taverna's Annotation Editor to insert quality annotations for process and data sources before the workflow is executed. Listing 1 shows an excerpt of the annotated workflow specification, where, for instance, the reputation of the Catalogue of Life is 1 (maximum) and its availability is 0.9 (since there are several connection problems) - lines 7 and 8 of the listing.

Taverna exports provenance information using the OPM (Open Provenance Model) model [26]. The Provenance Manager merges this information with Taverna's annotated workflow, and maps the result into the Provenance Repository, which uses Malaverri's model [3].

The Data Quality Manager is also being implemented. As a proof of concept, we implemented a small piece of code to compute the accuracy of species name metadata, defined as a percentage of correct names. Moreover, this code outputs the availability and reputation of the Catalogue of Life. As a result, the end user can see that the original FNJV metadata, compared with an external authoritative source (reputation 1, availability 0.9) is 93% accurate. These results were shown to expert users, helping them to better understand their data. The piece of code accesses both the Provenance Repository and workflow output to provide such quality dimensions. The final version of the Quality Manager will be based on Lemos' proposal [17]. It will provide an interface to end users to specify dimensions and indicate means to compute them - e.g., designating web services or software components.

Ongoing work involves remodelling FNJV metadata database to reflect the history of curation processes (whenever a field is changed, but not syntactic or domain checks).

V. Conclusions

Preservation and curation are indissociable concepts. As such, preservation involves not only ensuring appropriate storage and maintenance procedures, but also accessibility, usability and quality. This paper describes our approach to long term preservation of scientific data, which is specifically concerned with enhancing and maintaining metadata quality, as a means to increase reuse and quality of associated scientific data.

Another associated activity is to provide support to connect curated metadata with Linked Data initiatives. A first prototype, reported in [37], shows how such mechanisms allow crossreferencing scientific papers across distinct research communities, even when they appear to work in seemingly unrelated issues. This will allow breaking down disciplinary boundaries among repositories and enhance reuse (and thus promote curation). Finally, we point out that workflows may also decay - e.g., see Zhao et al. [38]. This reinforces the notion that quality assessment must be a continuous task, as long as users deem the data to be useful - i.e., this task is needed throughout the preservation life cycle.

Acknowledgments

Work partially financed by FAPESP (2011/19284-3), FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), INCT in Web Science, and CNPq. We thank Prof. Omar Boucelma from Univ. Aix-Marseille for his valuable suggestions.

References

[1] Z. A. et al., "Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics," DPHEP - www.dphep.org, Tech. Rep. DPHEP-2012-001, 2012.

[2] "The Domesday Project," 1986, http://www.csa.com/discoveryguides/cyber/overview.php, accessed in November 2013.

[3] J. E. G. Malaverri, "Supporting data quality assessment in escience: a provenance based aproach," Ph.D. dissertation, Universidade Estadual de Campinas, Campinas, SÃčo Paulo, 2013.

[4] E. Conway, D. Giaretta, S. Lambert, and B. Matthews, "Curating Scientific Research Data for the Long Term : A Preservation Analysis Method in Context," *The International Journal of Digital Curation*, vol. 6, no. 2, pp. 38–52, 2011.

[5] L. Trimble and S. Marks, "Supporting Data Access and Reuse in Ontario: Scholars Portals Initiatives," in *Proc. World Social Sciences Forum*, Montreal, Canada, 2013, extended abstract.

[6] "Ontario council of university libraries," 2010, http://http://ocul.on.ca/, accessed in November 2013.

[7] "Research data canada," 2011, http://http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/, accessed in november 2013.

[8] N. D. I. Infrastructure and P. Program, "Digital preservation," 2000, http://www.digitalpreservation.gov/, accessed in november 2013.

[9] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, and C. Goble, “Why linked data is not enough for scientists,” *Future Generation Computer Systems*, vol. 29, no. 2, pp. 599–611, Feb. 2013.

[10] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data-the story so far,” *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009.

[11] R. Y. Wang and D. M. Strong, “Beyond accuracy: What data quality means to data consumers,” *Journal of management information systems*, pp. 5–33, 1996.

[12] Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations,” *Communications of the ACM*, vol. 39, no. 11, pp. 86–95, 1996.

[13] A. Parssian, “Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions,” *Decision Support Systems*, vol. 42, no. 3, pp. 1494–1502, 2006.

[14] H. Scholten and A. J. Udink ten Cate, “Quality assessment of the simulation modeling process,” *Computers and Electronics in Agriculture*, vol. 22, no. 2, pp. 199–208, 1999.

[15] M. Gamble and C. Goble, “Quality, trust, and utility of scientific data on the web: Towards a joint model,” *Web Science Trust*, 2011.

[16] F. Chiang and R. J. Miller, “Active repair of data quality rules,” in *Proceedings of the 16th International Conference on Information Quality (ICIQ)*, 2011.

[17] F. Lemos, “Infrastructure and algorithms for information quality analysis and process discovery,” Ph.D. dissertation, University of Versailles, France, 2013.

[18] L. Etcheverry, V. Peralta, and M. Bouzeghoub, “Qbox-foundation: a metadata platform for quality measurement,” in *proceeding of the 4th Workshop on Data and Knowledge Quality (QDC'2008)*, 2008.

[19] A. Na'im, D. Crawl, M. Indrawan, I. Altintas, and S. Sun, “Monitoring data quality in kepler,” in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*. ACM, 2010, pp. 560–564.

[20] S. Bowers, J. Kudo, H. Cao, and M. P. Schildhauer, “Obsdb: A system for uniformly storing and querying heterogeneous observational data,” in *IEEE Sixth International Conference on e-Science*, 2010, pp. 261–268.

[21] Cornell, “The cornell lab of ornithology,” http://www.allaboutbirds.org, accessed in December 2012.

[22] K.-H. Frommolt, R. Bardeli, F. Kurth, and M. Clausen, “The animal sound archive at the humboldt-university of berlin: Current activities in conservation and improving access for bioacoustic research,” in *Advances in Bioacoustics II*, 2006, pp. 139–144.

[23] M. Cobos and J. Lopez, “Listen up - the present and future of audio signal processing,” *IEEE Potentials*, vol. 29, no. 4, pp. 40 –44, 2010.

[24] R. Bardeli, “Similarity search in animal sound databases,” *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 68–76, 2009.

[25] J. E. Malaverri, A. Santanche, and C. B. Medeiros, “A provenance-based approach to evaluate data quality in escience,” *Int. J. Metadata, Semantics and Ontologies*, 2013, accepted for publication.

[26] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers *et al.*, “The open provenance model core specification (v1. 1),” *Future Generation Computer Systems*, vol. 27, no. 6, pp. 743–756, 2011.

[27] D. C. Cugler, C. B. Medeiros, and L. F. Toledo, “Managing animal sounds - some challenges and research directions,” in *Proceedings V Brazilian eScience Workshop*, July 2011.

[28] R. Ranft, “Natural sound archives: past, present and future,” *Anais da Academia Brasileira de Ciï£¡ncias*, vol. 76, no. 2, pp. 456–460, 2004.

[29] “Encyclopedia britannica - academic edition,” Available: http://www.britannica.com/EBchecked/topic/409141/Neot-ropical-region, accessed in November 2011.

[30] FNJV, “Online animal sound collection - Fonoteca Neotropical Jacques Vielliard,” http://proj.lis.ic.unicamp.br/fnjv, accessed in June 2013.

[31] L. Toledo and C. Haddad, “Reproductive biology of *Scinax fuscomarginatus* (anura, hylidae) in south-eastern Brazil,” *Journal of Natural History*, vol. 39, no. 32, pp. 3029–3037, 2005.

[32] U. Caramaschi, “Notes on the taxonomic status of *Elachistocleis ovalis* (schneider, 1799) and description of five new species of *Elachistocleis Parker*, 1927 (amphibia, anura, microhylidae),” *Boletim do Museu Nacional. Nova Serie, Zoologia*, vol. 527, pp. 1–30, 2010.

[33] D. C. Cugler, C. B. Medeiros, and L. F. Toledo, “An architecture for retrieval of animal sound recordings based on context variables,” *Concurrency and Computation: Practice and Experience*, June 2012.

[34] D. Cugler, C. B. Medeiros, S. Shekhar, and F. Toledo, “A Geographical Approach for Metadata Quality Improvement in Biological Observation Databases,” in *Proc. 9th IEEE International e-Science Conference*, 2013.

[35] “Catalogue of life,” http://www.catalogueoflife.org, accessed in October 2013.

[36] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn, “Taverna: a tool for building and running workflows of services,” *Nucleic acids research*, vol. 34, no. suppl 2, pp. W729–W732, 2006.

[37] M. S. Mota and C. B. Medeiros, “ Introducing Shadows: Flexible Document Representation and Annotation on the Web,” in *Proc International Workshop on Data Engineering Meets the Semantic Web (DESWEB)*, Brisbane, 2013, co-located with 29th ICDE conference.

[38] J. Zhao, J. M. Gomez-Perez, K. Belhajjame, G. Klyne, E. Garcia-Cuesta, A. Garrido, K. Hettne, M. Roos, D. De Roure, and C. Goble, “Why workflows breakâĂŤUnderstanding and combating decay in Taverna workflows,” in *E-Science (e-Science), 2012 IEEE 8th International Conference on*. IEEE, 2012, pp. 1–9.