biblio.ugent.be

The UGent Institutional Repository is the electronic archiving and dissemination platform for all UGent research publications. Ghent University has implemented a mandate stipulating that all academic publications of UGent researchers should be deposited and archived in this repository. Except for items where current copyright restrictions apply, these papers are available in Open Access.

This item is the archived peer-reviewed author-version of:

A Model of Provenance Applied to Biodiversity Datasets

Flor Mamani Amanqui, Tom De Nies, Anastasia Dimou, Ruben Verborgh, Erik Mannens, Rik Van de Walle, and Dilvan Moreira

In: 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 235 - 240, 2016.

http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7536468&tag=1

To refer to or to cite this work, please use the citation to the published version:

Mamani Amanqui, F., De Nies, T., Dimou, A., Verborgh, R., Mannens, E., Van de Walle, R., and Moreira, D. (2016). A Model of Provenance Applied to Biodiversity Datasets. *2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* 235 - 240. 10.1109/WETICE.2016.59

A Model of Provenance Applied to Biodiversity Datasets

Flor K. Amanqui^{1,2}, Tom De Nies¹, Anastasia Dimou¹, Erik Mannens¹,
Rik Van de Walle¹, Dilvan Moreira²

¹Ghent University - iMinds - Data Science Lab, Belgium

²University of São Paulo, ICMC, São Carlos, Brazil

Email: {flork, dilvan}@icmc.usp.br, {tom.denies, anastasia.dimou, erik.mannens, rik.vandewalle}@ugent.be

Abstract—Nowadays, the Web has become one of the main sources of biodiversity information. More and more biodiversity research institutions add new specimens and their related information to their biological collections and make this information available on the Web. However, mechanisms which are currently available provide insufficient provenance of biodiversity information. In this paper, we propose a new biodiversity provenance model extending the W3C PROV Data Model. Biodiversity data is mapped to terms from relevant ontologies, such as Dublin Core and GeoSPARQL, stored in triple stores and queried using SPARQL endpoints. Additionally, we provide a use case using our provenance model to enrich collection data.

Index Terms—Semantic Web; Linked Data; Provenance; Biodiversity.

I. INTRODUCTION

Biological diversity is essential to life on Earth and motivates many efforts to collect data about species [1]. That gives rise to large amounts of data. These data are collected in different places and published in different formats. Collecting data in the field is expensive, difficult, and sometimes dangerous. Not only does it require close interaction with organisms, but it also requires close collaboration with different people [2].

Several research institutions are setting up biological collection programs as part of their scientific strategic plan. Some of these research institutions are: the Global Biodiversity Information Facility¹ (GBIF), the Biodiversity Database Collection of the National Research Institute for the Amazon² (INPA), the Large-Scale Biosphere-Atmosphere Experiment in Amazonia³ (LBA), the Reference Center on Environmental Information⁴ (CRIA), and the New York Botanical Garden⁵ (NYBG). While most researches in biodiversity pay much attention to the generation of biodiversity datasets and Web access, information that specifies how/where these data are derived and who owns/publishes the data is often ignored.

In this paper, we propose a conceptual model for provenance in biodiversity data for species identification. This model is based on PROV [3]. The PROV specification provides the concepts and supporting definitions to enable the interoperable interchange of provenance information in heterogeneous environments such as the Web [4]. This conforms to the principles of Linked Open Data (LOD), which encourages a set of best practices for publishing and connecting structured data on the Web [5].

To further clarify the relation between the biodiversity domain and provenance, we provide an example in Figure 1.

Figure 1: Example of Provenance in Biodiversity Data

Due to new discoveries, species names could change over time. As Figure 1 shows, this is the case for the *Batrachospermum Alga*. This species was collected and was registered in a csv file by the Collector. After a cataloguing process, this species was determined as *Batrachospermum Helminthosum* by the Cataloguer. After 15 years, a User needs to determine the genetic name of this species. Subsequent to molecular studies, this species had its name changed to *Batrachospermum viride-brasilense*. This is a common problem faced by biodiversity collections. The user needs to answer the following questions: Who was the cataloguer of the species? Who was the collector of the species? When was the data collected? Why was the data collected? Which institution can provide the data? The user needs to know the history (provenance) of the species. This means that the trustworthiness of the cataloguer, the person who determined the species and the user involved

¹<http://www.gbif.org>

²<http://colecoes.inpa.gov.br>

³<http://lba.inpa.gov.br>

⁴<http://www.cria.org.br>

⁵<http://www.nybg.org>

should be judged, since they participate in the identification and modification of the species name.

As our main contribution in this paper, we present an extended provenance model applied to biodiversity datasets. We mapped a set of representative data about biodiversity (206,000 records) from the Botanical Institute (IBt/SP). This data was downloaded from the SpeciesLink web site⁶. SpeciesLink is a distributed information system that integrates primary data from biological collections from many research institutions from Brazil and abroad. It is also a popular online tool to search for biodiversity data. We also use the GeoSPARQL language⁷ (an extension to the SPARQL language that allows queries based on spatial relations, such as being inside a polygon, etc.) to answer spatially complex queries.

The remainder of this article proceeds as follows: Section II discusses related work. Section III shows the provenance model for biodiversity datasets. In Section IV, we present a use case, where we model all 217,829 records of the SpeciesLink website using our proposed approach and Section V concludes by summarizing our results and describing future work.

II. RELATED WORK

Biodiversity data is an assortment of different types of organisms that co-occur in time and space (geospatial) [1]. We investigated existing works related with the geospatial and biodiversity domain that use a provenance model.

Zhao et al. [6] propose a method for recording the provenance data into biological dataset. This method helps scientists to obtain the information about particular biological terms. The authors use the Dublin Core Metadata Initiative (DCMI) [7] and named RDF graphs to represent the aspects of data provenance. The authors only considered two types of links between a pair of genes, i.e., either they are same as or different from each other. Provenance information about these links is needed to provide reliable and accurate services to researchers.

Beserra et al. [8] propose a provenance-based approach to manage long term preservation of scientific data. This approach uses a case study related to the long term preservation of the animal sound collection at the Fonoteca Neotropical Jacques Viellard (FNJV)⁸. Their approach is based on the Open Provenance Model (OPM) [9]. However, this approach does not provide support to connect curated metadata with Linked Open Data. This will allow breaking down disciplinary boundaries among repositories and enhance reuse.

There are a number of studies, which have used provenance in the geospatial domain [10], [11], [12], [13]. For example, Wang et al. [10] propose a provenance-aware architecture to record the lineage of spatial data in Geographic Information Systems (GIS). Their architecture is based on the OPM model and organize spatial provenance as objects in spatial data store, while a spatial process provenance is represented as graphs and stored using Semantic Web technologies based on the RDF, and is backed with standard storage technologies

(e.g. database) and RDF stores (e.g. Sesame). However, the authors do not consider the variable spacetime to enhance the geospatial capabilities of either provenance-aware GIS.

Yuan et al. [11] propose a Linked Data approach for geospatial data provenance. The authors defined a geospatial data provenance ontology based on the Provenir ontology⁹, published geospatial data provenance as Linked Data, and analyzed queries of linked geospatial data provenance. Their approach is based on the Registry Information Model (ebRIM)¹⁰ and the DCMI model. However, this approach does not achieve geospatial reasoning based on the linked geospatial data provenance.

Magnuson et al. [1] explain that biodiversity research will often have specific taxonomic or ecosystem interests, but the primary keys that link all of these things are related to space (geographic location) and time (when the observation was made). However, most of the available biological collections are not consistently georeferenced making use of a coordinate system. The authors explain that there is still a fundamental lack to answer complex queries, i.e., queries that need logical inference that use spatial and temporal relations (e.g., plantations within a protected area in Manaus, Brazil between 2005 and 2011).

A critical look at the available literature indicates that a number of techniques have been developed for using provenance models, such as OPM and DCMI, in different scientific domains (biological, biodiversity and geospatial). Despite the variety of models, there is currently no unified, conceptual model for biodiversity information that can be applied to different datasets and setups, while remaining both expressive and generic enough to cover many use cases.

The PROV specification [3] defines a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or thing in the world. However, there is a lack of expressiveness using this generic W3C recommendation to model the different types of organisms that co-occur in time and space (geospatial relations).

III. ARCHITECTURE OF PUBLISHING LINKED BIODIVERSITY DATA

This section presents our architecture for publishing linked biodiversity data (as illustrated by Figure 2). Our architecture uses Linked Data and Semantic Web standards (Resource Description Framework (RDF) [14] and the Web Ontology Language (OWL) [15]) to represent biodiversity data.

SPARQL is a W3C standard language for querying RDF data (triples) [16]. Figure 3 shows a RDF triple, it is comprised of three pieces of information: Subject (S), Predicate (P), and Object (O). Where S and O are nodes and P is the property or aspect that relates the subject to the object. In Figure 3, the butterfly#1 (S) was collected in the stage of life (P) larva (O).

⁶<http://splink.cria.org.br/>

⁷<http://www.opengeospatial.org/standards/geosparql>

⁸<http://www2.ib.unicamp.br/fnjv/>

⁹http://wiki.knoesis.org/index.php/Provenir_Ontology

¹⁰<http://docs.oasis-open.org/registregrep/registregrep-core/v4.0/registregrep-core-rim-v4.0.html>

Figure 2: Architecture of publishing linked biodiversity data

Figure 3: RDF example, the `butterfly#1` (S) was collected in the stage of life (P) larva (O)

SPARQL syntax and the way it queries data are based on the RDF triple scheme (the basis for RDF data representation). That makes it possible to create searches that seek not only based on instances, but also on the relationships between them. SPARQL Endpoints are portals to data that a provider makes available for querying using SPARQL. They are usually implemented using triple stores. Triple store is the common name given to a database management system for RDF triples. They provide data management and data access by way of APIs and query languages to RDF triples (such as SPARQL).

When biodiversity data is collected and catalogued by third parties (*Cataloguer* and *Collector*), it is registered and stored in commercial spreadsheets (e.g., Microsoft Excel) or databases (e.g., DBase, PostgreSQL) or files associated with statistical programs (e.g., R or SPSS) [1]. A mapping component converts the biodiversity data to RDF. The RML language [17] is used to map biodiversity data to RDF triples. The RDF triples are integrated with our provenance model for biodiversity data. These RDF triples are stored in the triple stores. After that, users can retrieve biodiversity information through SPARQL queries.

In the next subsections, we explain more details about each component of our architecture of biodiversity data provenance.

A. Provenance Model for Biodiversity Data (BioProv)

In order to create our provenance model, five biodiversity scientists were interviewed to categorize important information from biodiversity data (e.g. collecting process, genus, family, species, description of location). These interviews helped us to understand more about their work and to form a common ground for discussions. A list of our interviews are available at <http://java.icmc.usp.br:1100/provWeb/U.html>.

Using the information gathered through these interviews, we created our provenance model for biodiversity data, as shown in Figure 4. This model is based on the W3C PROV model [3]. It defines a set of starting point terms which are three classes: Entity, Agent and Activity. These classes are associated by relations such as *prov:wasAttributedTo*, *prov:wasInformedBy*, etc. The entity responsible for commanding the execution of an activity is modeled as an *Agent*.

In our model, a species is denominated *Collection Object* (CO). The CO was generated by an activity denominated *Collecting* (in Figure 4, *prov:wasGeneratedBy*). A *Collector* agent was associated with this activity (in Figure 4, *prov:wasAssociatedWith*). We trace the date this activity was executed with the property *prov:atTime*.

After a *Collecting* process, the *Collection Object* needs to be identified through an activity denominated *Cataloguing* (in Figure 4, *prov:wasGeneratedBy*). The *Cataloguer* agent assigns a unique identifier to each collection item using the taxonomic classification. The *Cataloguer* uses the reference work to indicate the published material in which the collection object is mentioned (in Figure 4, *Reference Work*).

The *Collection Object* has a relationship to the *Reference Work*, indicating the published material in which the *Collection Object* is mentioned (in Figure 4, *prov:wasDerivedFrom*). The *Location* describes where the data was collected (in Figure 4, *prov:atLocation*). It defines locality, named place, habitat and spatial information.

The Agents describe persons and organizations, which deal with the biological collection information and interacts with all activities and all information that is updated in the model. The *Curator*, *Collector*, *Cataloguer* and *User* agents are members of an *Organization Agent* (in Figure 4, *prov:actedOnBehalfOf*).

The PROV data model provides extensibility points that allow designers to specialize it for specific applications or domains (*subtypes*, *roles*, and *Attribute-value lists*) [18]. In order to model the species names that are defined by the agents, we propose the following extensions that are *subtypes* of *prov:Entity*:

- *bioprov:OriginalSpeciesName*: denotes an original species name that is not derived from any other name and the collector who emitted it is the initiator of the identification process.

Figure 4: Provenance model for biodiversity data.

- *bioprov:CataloguerSpeciesName*: denotes a species name which is based on another name that has been published in the past. The *Cataloguer* agent emitted this name based on the original species name.
- *bioprov:MolecularSpeciesName*: denotes a species name that is produced by modifying an existing species name. It is possible that the species name is altered.

With these three extensions we have covered the main case of the example illustrated in Section 1 (species names could change over time). Our model is available at <http://java.icmc.usp.br:1100/provWeb/bioprov.ttl>.

B. Mapping Provenance and Biodiversity Data to RDF

The Mapping component of our architecture of publishing linked biodiversity data loads the domain ontologies, provenance model, taxonomic information and the collection database and transforms them in a set of RDF triples. We used the RDF Mapping Language (RDF) [17] to represent the mapping between rows of data tables (in csv files) and properties and objects in RDF.

The RML is a mapping language defined to express customized mapping rules from heterogeneous data structures and serializations to the RDF data model. RML is defined as a superset of the W3C-standardized mapping language (R2RML) [17].

A Triples Map defines how triples of the form (subject, predicate, object) will be generated. A Triples Map consists of three main parts: the Logical Source, the Subject Map and zero or more Predicate-Object Maps. In the following, we show an example of a triple map:

```

@prefix rr:<http://www.w3.org/ns/r2rml#>.
@prefix rml:<http://semweb.mmlab.be/ns/rml#>.
@prefix geobio:<http://geobio.lod.usp.br/>.
@prefix prov:<http://www.w3.org/ns/prov#>.
@prefix dterms:<http://purl.org/dc/terms/>.
@prefix adms:<http://www.w3.org/ns/adms#>.
@prefix bioprov:<http://geobio.lod.usp.br/bioprov/>.
@prefix skos:<http://www.w3.org/2004/02/skos/core#>.
@prefix xsd:<http://www.w3.org/2001/XMLSchema#>.
@prefix foaf:<http://xmlns.com/foaf/0.1/>.
<#BioMapping>
rml:logicalSource [rml:source "
    speciesLink_all_28128_20160203101852.csv";
rml:referenceFormulation ql:CSV];
rr:subjectMap [
rr:template "http://geobio.lod.usp.br/ibt/id/{code}";
rr:class adms:Identifier];
rr:predicateObjectMap [rr:predicate skos:notation;
    rr:objectMap [rml:reference "code"];];
rr:predicateObjectMap [rr:predicate prov:agent;
    rr:objectMap [rml:reference "institutioncode"];];
rr:predicateObjectMap [rr:predicate prov:generatedAtTime;
rr:objectMap [
rml:reference rml:reference "datelastmodified";
rr:datatype xsd:dateTime ];];
  
```

The Logical Source represents the source to be mapped. This can be a pointer to any dataset (Line 12-13). The Subject Map (Line 14-16) defines how unique identifiers (URIs) are generated for the mapped resources and is used as the subject of all RDF triples generated from this Triples Map. A Predicate-Object Map (Line 17-24) consists of Predicate Maps, which define the rule that generates the triple's predicate and Object Maps or Referencing Object Maps (Line 18, 20, 22), which define how the triple's object is generated. The Subject Map, the Predicate Map and the Object Map are Term Maps, namely rules that generate an RDF term (an IRI, a blank node

or a literal).

IV. USE CASE

In order to validate our provenance model, biodiversity scientists were interviewed to define use cases with features and scenarios to identify the various user tasks. A list of our use cases are available at <http://java.icmc.usp.br:1100/provWeb/U.html>. In this article, we present one of these use cases:

USE CASE 01: Molecular Identification of *Cladophora delicatula* Alga

USER: Monica Paiano, 32 years-old, Collector and Cataloguer of the Laboratory BETA, UNESP, Brazil and Phycology Research Group, Ghent University, Belgium.

GOAL: To determine the scientific name of *Cladophora delicatula* alga thought a genetic identification.

MOTIVATION: Due to new discoveries, species names of *Cladophora delicatula* alga could change over time. Keeping such data up to date and consistent is extremely important because the presence or not of some species of this alga can serve as biological markers (bio indicators) that indicate the degree of conservation or degradation in a aquatic habitat.

TASKS

1. Retrieve all information about *Cladophora delicatula* alga. For example, when it was collected, who collected it, all the specific characteristics;
2. Store all the information in csv, text file or in a biodiversity database;
3. Start the molecular studies of the *Cladophora delicatula* alga;
4. Review the reproduction isolation of *Cladophora delicatula* alga;
4. Identify the species names in a flexible way: using the broader taxonomic level (phylum or genus) without having to worry about whether the original collection used this particular classification level.

NECESSARY TOOL FEATURES

1. Retrieve all specifications of the bio-marker species using the species name or any higher taxonomic level, like phylum, genus or family.

After studying our use case, we mapped the corresponding biodiversity provenance records to RDF. We used the RML language to convert all IBT's records from the SpeciesLink web site (217,829 records) to RDF triples. This RDF data was stored in our Virtuoso Triple Store and can be explored using SPARQL queries. The biodiversity datasets are available at <http://java.icmc.usp.br:1100/provWeb/ProvGeoIBT.rdf>.

In order to show how the previous use case was implemented, in the following subsections we explain the more important activities of our provenance model: *Collecting* and *Cataloguing Activity*.

A. Collecting Activity

For this activity (Figure 4, *Collecting*), it is important to keep track of (1) When was the species collected (2) Who was the collector of the species, (3) Where was the species collected, and (4) Which institution can provide the species. This activity is crucial for capturing the origin of the biodiversity data, as it is only at this step that information is known. In the following, we show an example of our

provenance model applied to the collecting process of the *Cladophora Delicatula* species.

In Example 1, we used the subclass *bio-prov:OriginalSpeciesName* to define the species name (Figure 4, *bioprov:OriginalSpeciesName*).

Example 1:

```
:AgentCollector #Who was the collector of the
species
a prov:Agent;
foaf:givenName "M.C.Marino & R.Marino";
prov:actedOnBehalfOf :IBT
.
:IBT #Who institution can provide the species
a foaf:Organization, prov:Agent;
foaf:name "Botanical Institute (IBT/SP)"
.
:Collecting #When was the species collected
a prov:Activity;
prov:wasAssociatedWith :AgentCollector;
prov:atTime "1966-11-11T01:01:01Z"^^xsd:dateTime;
bioprov:OriginalSpeciesName "Cladophor Delicatula"
.
```

In Example 2, we used the properties *prov:activity*, *prov:atTime* and *prov:atLocation* to define the spatiotemporal location of our species collected. To deal with this, we used the GeoSPARQL language and the Well-Known Text (WKT), a pattern defined by the Open Geospatial Consortium (OGC) for defining coordinates in the form of points, lines and polygons.

Example 2:

```
:Location #Where was the species collected
a prov:Entity;
foaf:name "Ponta do Gil Lake";
prov:qualifiedGeneration [
a prov:Generation;
prov:activity geo:feature; #Link to GeoSPARQL
prov:atTime "1966-11-11T01:01:01Z"^^xsd:dateTime;
prov:atLocation "POINT((-46.7175
-23.653056))"^^geo:wktLiteral; //Latitude and
Longitude
].
```

B. Cataloguing Activity

The *Cataloguing* activity (Figure 4, *Cataloguing*) permits the taxonomic identification of the biodiversity data. The taxonomic identification information contains the identifiers of the biological classification as Order, Family, Genus, Superspecies, Species and nearly all of them were used. In the following, we show an example:

Example 3:

```
:AgentCataloguer // Who was the cataloguer of the
species
a prov:Agent;
foaf:givenName "D.P. Santos"; //Name of the
Cataloguer
prov:actedOnBehalfOf :IBT
.
:Cataloguing // When was the species catalogued
a prov:Activity;
prov:wasAssociatedWith :AgentCataloguer;
prov:atTime "1982-01-01T01:01:01Z"^^xsd:dateTime;
bioprov:CataloguerSpeciesName "Cladophor
Delicatula"
.
```


We use the ProvValidator and ProvTranslator tools¹¹ to validate and translate PROV representations about the collecting and cataloguing activities of our biodiversity datasets, making them fully interoperable. The complete representations of our use case are available at <http://java.icmc.usp.br:1100/provWeb/ExampleBioProv.html>.

To integrate the biodiversity data in RDF to the wider LOD community on the Web, we set up a SPARQL endpoint (<http://java.icmc.usp.br:1100/strabonendpoint/>). Our endpoint allows third-party programs to query our knowledge base, via the SPARQL language, and reuse it in their applications.

C. Querying Linked Biodiversity Data Provenance

For the previous use case, a *User* wants to identify all information about *Cladophora delicatula alga*. One of the big advantages of having the biodiversity data in RDF is to be able to connect it to other sources. We created a SPARQL query for integrating different triples stores. The following example is provided to show the SPARQL query used to obtain the provenance information of a specific dataset.

```
SELECT ?Species ?agent ?activity ?date ?
      pontowkt
WHERE {
  ?Species prov:wasAttributedTo ?agent.
  ?Species prov:wasGeneratedBy ?activity.
  ?activity prov:atTime ?date.
  ?Species geo:hasGeometry ?Geometry.
  ?Geometry geo:asWKT ?pontowkt.
  ?Species bioprov:CataloguerSpeciesName "
    Cladophora delicatula"^^<http://www.w3.org
    /2001/XMLSchema#string>.
  FILTER(?date > "1980-01-01T01:01:01Z"^^<http
    ://www.w3.org/2001/XMLSchema#dateTime>).
  geof:sfWithin(?pontowkt, "Polygon((-1 -58, -7
    -58, -7 -69, -1 -69, -1 -58))"^^sf:
    wktLiteral).
}
```

Using this query, we could retrieve the lineage of the *Cladophora delicatula* species. In our provenance model, we reused the GeoSPARQL ontology terms to describe georeferenced data. This implementation permits to answer complex queries such as: Locate all occurrences containing *Cladophora delicatula* alga samples inside of a Polygon (-1 -58, -7 -58, -7 -69, -1 -69, -1 -58)

V. CONCLUSION

In this work we presented a model for biodiversity data provenance (BioProv). This model is based on the W3C PROV ontology and data model. BioProv enables applications that analyze biodiversity to incorporate provenance data in their information. We defined a mapping document for the biodiversity data from IBT to generate RDF triples. We also reused the GeoSPARQL ontology terms to describe georeferenced data. We use the provenance information to allow experts in biodiversity to perform queries and answer scientific questions.

As future work, we also intend to extend our current implementation with more advanced structured queries, in partnership with biodiversity researchers.

ACKNOWLEDGMENTS

The research activities described in this paper were funded by Ghent University, iMinds, the IWT-Flanders, the FWO-Flanders, and the European Union, and the FINCyT Science and Technology Program from Perú.

REFERENCES

- [1] W. Magnusson, *Biodiversity and Integrated Environmental Monitoring*. Program for Planned Biodiversity and Ecosystem Research (PPBio), 2013.
- [2] J. L. C. dos Santos, "A biodiversity information system in an open data/metadata architecture," Ph.D. dissertation, Enschede, 2003. [Online]. Available: <http://doc.utwente.nl/41436/>
- [3] W3C, "The prov ontology (prov-o)," 2013. [Online]. Available: <http://www.w3.org/TR/prov-o/>
- [4] I. Taxisidou, T. De Nies, R. Verborgh, P. Fischer, E. Mannens, and R. Van de Walle, "Modeling information diffusion in social media as provenance with w3c prov," in *Proceedings of the 6th International Workshop on Modeling Social Media*, May 2015. [Online]. Available: <http://www.w3.org/2015.it/documents/proceedings/companion/p819.pdf>
- [5] W. Kuhn, T. Kauppinen, and K. Janowicz, "Linked data - a paradigm shift for geographic information science," in *Geographic Information Science*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8728, pp. 173–186. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11593-1_12
- [6] J. Zhao, G. Klyne, and D. Shotton, "Provenance and linked data in biological data webs," in *Proceedings of the WWW2008 Workshop on Linked Data on the Web (LDOW 2008)*, C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, Eds., 2008. [Online]. Available: <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/paper07.pdf>
- [7] S. L. Weibel and T. Koch, "Dublin core metadata initiative (dcmi)," 2000.
- [8] R. Beserra Sousa, D. Cintra Cugler, J. Gonzales Malaverri, and C. Bauzer Medeiros, "A provenance-based approach to manage long term preservation of scientific data," in *Data Engineering Workshops (ICDEW)*, 2014 *IEEE 30th International Conference on*, March 2014, pp. 162–173.
- [9] L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson, "The open provenance model: An overview," in *Provenance and Annotation of Data and Processes*. Springer, 2008, pp. 323–326.
- [10] S. Wang, A. Padmanabhan, J. D. Myers, W. Tang, and Y. Liu, "Towards provenance-aware geographic information systems," in *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '08. New York, NY, USA: ACM, 2008, pp. 70:1–70:4. [Online]. Available: <http://doi.acm.org/10.1145/1463434.1463515>
- [11] J. Yuan, P. Yue, J. Gong, and M. Zhang, "A linked data approach for geospatial data provenance," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 51, no. 11, pp. 5105–5112, Nov 2013.
- [12] J. E. G. Malaverri, C. B. Medeiros, and R. C. Lamparelli, "A provenance approach to assess the quality of geospatial data," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, ser. SAC '12. New York, NY, USA: ACM, 2012, pp. 2043–2044. [Online]. Available: <http://doi.acm.org/10.1145/2245276.2232116>
- [13] I. Celino, "Human computation vgi provenance: Semantic web-based representation and publishing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 51, no. 11, pp. 5137–5144, Nov 2013.
- [14] O. Lassila, R. R. Swick, W. Wide, and W. Consortium, "Resource description framework (rdf) model and syntax specification," 1998.
- [15] G. Antoniou, G. Antoniou, G. Antoniou, F. V. Harmelen, and F. V. Harmelen, "Web ontology language: Owl," in *Handbook on Ontologies in Information Systems*. Springer, 2003, pp. 67–92.
- [16] E. Prud'hommeaux and A. Seaborne, "SPARQL Query Language for RDF," W3C, Tech. Rep., 2006. [Online]. Available: <http://www.w3.org/TR/rdf-sparql-query/>
- [17] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle, "RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data," in *Proceedings of the 7th Workshop on Linked Data on the Web*, Apr. 2014. [Online]. Available: http://events.lindatdata.org/ldow2014/papers/ldow2014_paper_01.pdf
- [18] L. Moreau, P. Missier (Eds.), and W3C Provenance Working Group, "PROV-DM: The PROV Data Model. W3C," 2013.

¹¹<https://provenance.ecs.soton.ac.uk/>