Data Provenance Architecture Supporting Environmental Monitoring Processes

Daniel Lins da Silva[1], André Batista[2], Pedro Luiz Pizzigatti Corrêa[3]
University of Sao Paulo
Brazil
daniellins@usp.br
andrefmb@usp.br
pedro.correa@usp.br

ABSTRACT: *Long-term research and environmental monitoring are essential for the improved management of ecosystems and natural resources. However, to reuse this data for new experiments, decision-making processes, and integrate these data with other long-term initiatives, scientists need more information related to data creation and its evolution, intellectual property rights, and technical information in order to evaluate the use of this data. Provenance metadata emerges as a way to evaluate the quality and reliability of data, audit processes and the data versioning, while enabling the data reuse and the reproducibility of experiments and analysis. However, most solutions for the capture and management of provenance metadata are based on specific tools, restricted scopes, and they are difficult to apply in distributed and heterogeneous environments. In this paper, we present an approach for capturing, managing, and publishing the provenance metadata generated in the environmental monitoring processes. Our computational architecture comprises three main components: (1) a data model based in PROV-DM and Dublin Core; (2) a repository of RDF Graphs; and (3) a Web API that provides services for collecting, storing, and querying provenance metadata. We demonstrate the application of our approach and show its practical usefulness by evaluating this architecture to manage provenance metadata generated during an environmental monitoring simulation. The results show that our approach is effective in collecting and storing provenance metadata and allows the query of an entire provenance of datasets and data products, thus enabling reuse, discovery, and visualization of raw data, processes, and scientists involved in its generation and evolution.*

Keywords: Data Provenance; Computational Architecture; Data Management; Environmental Monitoring

Received: 19 June 2016, Revised 24 July 2016, Accepted 19 July 2016

1. Introduction

The long-term research, which often includes monitoring, has a vital scientific role by revealing long-term trends that can lead to new knowledge and understanding. It is a necessary component of environmental science and policy design [1], [2]. The monitoring process consists of measuring physical, chemical, and/or biological variables over time, to provide information on ecosystem change. With this information, scientists can evaluate the health of our natural resources and make science-based management decisions [2], [3].

Environmental monitoring programs can vary significantly in the scale of their spatial and temporal boundaries, and in scope,

ranging from community-based monitoring to collaborative global monitoring programs [1]. Furthermore, monitoring processes uses a variety of equipment, technologies and techniques depending on the goals of the monitoring.

In recent years, the Internet of Things (IoT) initiatives has offered a number of new tools (sensors, cameras, microphones, etc.) to improve the environmental and biological monitoring [4]–[6]. These innovations provide cost-effective ways to collect monitoring data at large spatial scales over long survey windows, thereby increasing the statistical power of these survey efforts [4].

However, the rise of cheap and powerful sensors has created an ever-increasing data volume, creating more challenges to handling and managing the large amount of data generated by these devices. To be useful for experiments and decision-making processes, these data must be accompanied by context on how they are captured, processed, analyzed, validated, and interpreted [7]. Scientists need information regarding how the data were created and updated, intellectual property rights, the original source object from which this data object derives, and technical information [8], [9]. Information that describes the data origin, people, institutions, and processes involved in data generation and evolution are called provenance metadata [10].

1.1 Data Provenance

Data provenance can be used for various purposes, such as evaluating the quality and reliability of data, audit processes, data versioning, reproducibility of experiments, the establishment of property data, and discovery of new data [11]. However, to obtain the benefits of data provenance, scientists need to register provenance metadata during the various stages of the data lifecycle.

Currently in the environmental science community, the most used tools for the documentation of scientific processes and data analysis activities are the provenance components of a Workflow Management Systems (WfMS).

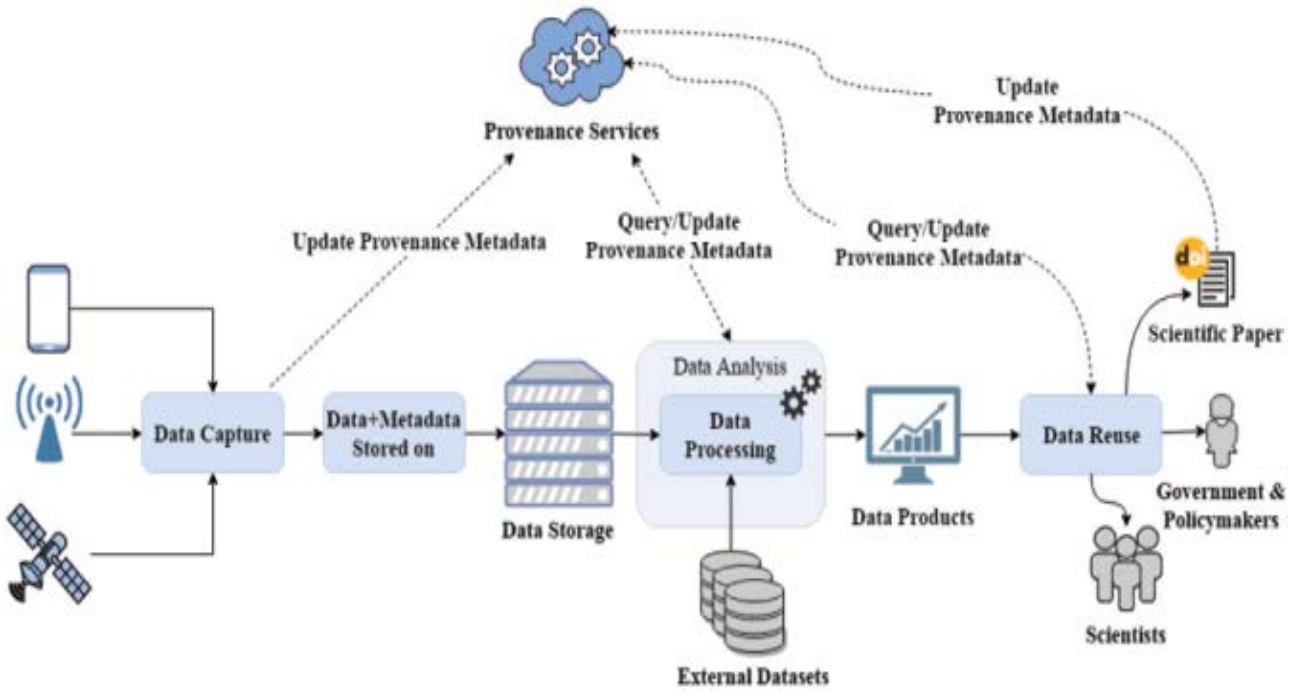

Figure 1. Proposed Approach For Managing Provenance Metadata In The Environmental Monitoring Process

Several studies have proposed standards, techniques, and tools to improve the capture of provenance data in WfMSs [12]–[14]. The advantages of these mechanisms are the quick configuration to capture provenance metadata and the possibility of automatic reproducibility of workflows. The main drawback is the mandatory use of these tools. Often the use of WfMS is not possible, due to the characteristics of the processes and techniques applied. The WfMSs also are not compatible with other

tools, making it difficult to maintain the provenance metadata in distributed and heterogeneous environments. In addition, the provenance metadata generated by some WfMSs are stored locally in files or internal databases, which makes difficult to discover and reuse these data by other scientists.

In this paper, we propose an approach to the management of data provenance from environmental monitoring through a computational architecture based on services and web standards. Fig. 1 shows our approach, where the data provenance are recorded during all phases of data capture, data processing, and data derivation. These metadata are recorded in a centralized provenance repository and are linked to the data through a Uniform Resource Identifier (URI), associated to the data object. Using this identifier, scientists and machines can access provenance metadata through our provenance services on the Web.

The main contribution of this work is the computational approach used to capture, manage, and publish the provenance metadata in distributed and heterogeneous environments.

2. Provenance Computational Architecture

To support capturing, recording, querying, and managing the provenance metadata, we have specified a provenance computational architecture. This architecture is comprised of three main components:

• A **provenance data model** to organize and structure the provenance metadata;

• A **provenance services** that provide a Web API for recording, querying, and visualizing provenance metadata in distributed environments;

• A **provenance repository** to store provenance metadata and provide capabilities to query and analyze these data.

This computational architecture is presented in Fig. 2. We will describe the main components of this architecture in the next sections.

2.1 Provenance Data Model

We defined a provenance data model based on the W3C PROV-DM to describe detailed information about the data and processes involved in the lifecycle of data objects.

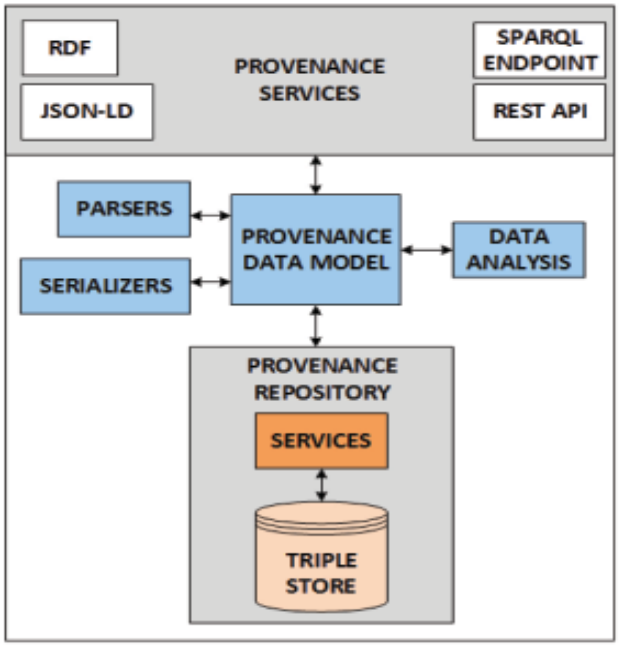

Figure 2. Over View of The Computational Architecture For Managing Provenance Metadata

The PROV-DM was designed to describe people, entities, and activities involved in producing a piece of data or any object [10]. The concepts found in the core of PROV-DM are described below [15]:

• **Entity:** Entity is something we want to describe. They can be physical, digital, or other types of objects, e.g. records, datasets, papers, and models;

• **Activity:** Activity is something that occurs over a period of time and acts upon or with entities. They are described by duration and their correlation with entities. The specified activities are consuming, processing, transforming, modifying, relocating, using, and generating entities;

• **Agent:** Agent is something that bears some form of responsibility. They can be responsible for activities that have taken place, for the existence of entities, and for the activities of other agents, e.g. scientists, and institutions;

• **Relations:** Relation is a property that describes a binary assertion among the previous PROV types (entity, activity, agent);

• **Extended Structures:** In addition to the three core PROV types and relations, there are some extended structures like subtypes, identifications, and expanded relations.

Using these types and relations, we can describe all relationships about the data objects and the activities performed during the data processes. These provenance metadata take the shape of a directed graph, considering its nodes (vertices) to represent entities, activities, and agents (data objects, people, institutions, and processes), and its edges (arrows), to represent relations between nodes (how these nodes relate to each other).

Figure 3. shows a directed graph that represents the data provenance of a dataset generated by a hypothetical process of environmental monitoring.

Based on this information, scientists can understand the entire lifecycle of this dataset. First, in the data capture process, we can analyze equipment, settings, and the institution involved. This process generated a raw dataset (version 1). Next, another institution processed this dataset through a software tool and generated a new version (version 2) of this dataset. About this data processing activity, it is possible to analyze the algorithms and input parameters considered. All dataset versions were connected and can be discovered.

An additional concept called Bundle is also considered in our data model. A Bundle is a named set of provenance metadata, and is itself a PROV Entity, so allowing provenance of provenance to be expressed [15]. Using Bundle, we can describe metadata about the recording process of data provenance, such as the responsible scientist, the recording tools, or additional information about this process. In the management of provenance repository, the Bundle information is essential to control the traceability and the security of metadata.

Furthermore, to describe additional metadata in the provenance graph, we can define an application profile based on existing Resource Description Framework (RDF) vocabularies. Using existing vocabularies and properties, when appropriate, requires less effort and increases the interoperability of metadata. If these properties are not already available, it is possible to create a new vocabulary.

An application profile is a generic schema to design metadata records that meet specific application requirements, providing semantic interoperability with other applications based on globally defined vocabularies and models [16]. In the definition of our application profile we included the Dublin Core, Friend of Friend (FOAF), RDF Schema (RDFS), and PROV Ontology (PROV-O), along with domain-specific dictionaries, such as the Darwin Core and the EML.

The NASA Earth Science Data Preservation Content Specification [17] presents the required metadata to describe the data provenance and context of long-term scientific research. Table 1 shows these metadata grouped by categories. Using our application profile, scientists can describe these metadata about the data and the processes involved in a long-term research.

2.2 Provenance Services

We defined a provenance Web API, following the Representational State Transfer (REST) style [18], to provide services for recording, querying, and visualizing provenance metadata.

Resources are the fundamental concept of a REST API. A resource is an object with a type, associated data, relationships with other resources, and a set of methods that define its operations.

Category	Content
Preflight / pre-operations calibration	Instrument description; Preflight / Pre-operational calibration data
Data products	Raw data; derived data products; metadata
Data product documentation	Product team members; product requirements and designs; processing and algorithm version history; product generation algorithms; product quality; product application
Data calibration	Calibration method; calibration data
Data product software	Source code; software documentation; programming considerations; exceptions; test datasets; test plans; test results
Data algorithm inputs	Algorithm input documentation; algorithm input datasets
Data product validation	Validation record; validation datasets
Data software tools	Software readers and display tools

Table 1. Required Metadata To Describe The Provenance And Context of Data In Long – Term Scientific Research[17]

These resources are grouped in collections and identified by an URL. The default URL schema for accessing the provenance services is described below.

http://**[domain]**/api/**[resources]/[rId]?bundle=[bId]**

Where **[domain]** is the domain name of the Web API. **[resources]** corresponds to the resources described in Table 2. **[rId]** corresponds to the URI that uniquely identifies each resource. For Bundle association, we can use the parameter **bundle**, in the query string of URL. The use of bundle is optional.

Table 2 shows all resources provided by the provenance Web API, and their available methods.

For representation of data handled by the provenance services, we use the JSON-LD format [19]. We consider the use of JSON-LD to ensure the contextualization of metadata elements used in the provenance data model, without the need to represent them in RDF. Thus, we do not add the complexity of RDF handling and new requirements to the devices and applications (client applications of the provenance services), which mostly are compatible with REST and JSON, but not compatible with RDF.

JSON-LD are compatible with JSON format and adds semantic context to a JSON document using the "@context" element. The @context element can be directly embedded into the JSON document or be an external file.

In our approach, we created an external file called prov.jsonld1, which contextualizes all definitions related to our provenance data model based on the ontologies and vocabularies considered.

Resource	Description	Methods
/bundles/	The provenance of provenance. Metadata that describes the metadata capturing process.	GET POST PUT DEL
/entities/	The described data, such as datasets, records, files, scripts, and parameters.	GET POST PUT DEL
/activities/	Events that occur over a period of time and act upon or with Entities.	GET POST PUT DEL
/agents/	Responsible for an Entity or the execution of an Activity.	GET POST PUT DEL
/provenance/	Get the complete metadata about an Entity or Bundle.	GET
/bundles/<id>/data/	Get the complete metadata associated with this bundle <ID>.	GET
/sparql/	SPARQL Endpoint	GET
/entity/<id>/files/	Manage files associated with an Entity.	GET POST PUT DEL
/rdf/	Import RDF documents to the repository.	POST

Table 2. Description The Resources That Compose The Provenance Web API

To use the provenance context (prov.jsonld) with JSON documents during the web services calls, we only have to include a reference to this context file, which can be held in the message body or in the message header of the requests [19].

2.3 Provenance Repository

Although we consider the JSON-LD in provenance services, we decided to use a triple store to manage data in RDF. The ain eason for our decision was to ensure interoperability with the various tools and technologies compatible with RDF, such as SPARQL. For this reason, we implemented parsers and serializers to perform the conversion of JSON-LD documents to RDF and vice versa.

[1] https://goo.gl/oPcNMD

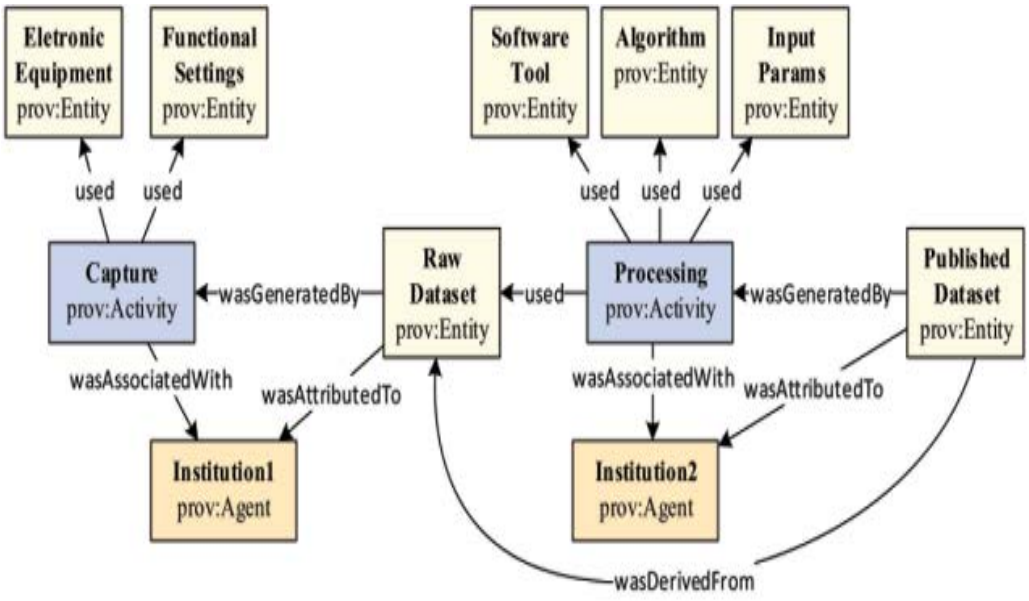

Figure 3. Representation of The Proposed Data Model, Based In Prov-DM Core Structures, To Describe A Environmental Monitoring Process

The RDF is a framework for representing information on the Web. The core structure of RDF is the RDF triple. Each triple consists of a subject, a predicate, and an object. A set of such triples is called an RDF Graph [20].

The provenance metadata represented for RDF graphs are stored in a triple store. The triple store is a database specialized in storing and retrieving RDF graphs. This type of database is compatible with the SPARQL, a RDF query language.

The provenance component, presented in our computational architecture, can access the triple store through a Web API. Many triple stores already provide Web APIs for data access. These Web APIs facilitate the standardization of systems, allowing the replacement of database vendor without major impacts on the application. Moreover, it avoids problems with the triple store native drivers, often outdated and limited to a few programming languages.

3. Case Study

To validate our approach, we developed a Web API Application called "BioProv Server"[2], based on the presented computational architecture, to manage the provenance metadata generated during the environmental monitoring and analysis processes. Furthermore, we implemented an R package, called "BioProv Client"[3], which provides a mechanism for the creation and storage of provenance metadata through the communication with the "Bioprov Server". Therefore, in our simulation, we used a suite of R functions provided by the "Bioprov Client" to record the provenance metadata during the execution of the simulated data process, represented by the R script.

Considering this approach, any device or application with internet access and compatible with HTTP protocol can query or insert provenance metadata using the provenance services provided by the "BioProv Server". Using the same Bundle URI, many devices or applications in distributed environments can store provenance metadata related to the same process, ensuring the description of processes running in a distributed manner.

The provenance metadata generated during our simulation was stored in the provenance repository and are available on the Web through the provenance services. "The BioProv Client" also enables the generation of provenance reports through the

[2] https://goo.gl/PLk3Ie
[3] https://goo.gl/ioiqCk

functions *getProvenanceOfBundle()*, that uses the Bundle URI as input parameter, and *getProvenanceOfEntity()*, that uses the URI associated to the data object. These reports can be generated in different formats: PROV-N, Turtle (RDF serialization), PNG, and SVG. In our simulation, the dataset generated is identified by the URI http://ib.usp.br/data_products/150.

The simulation script and its provenance reports, in PNG and RDF, are available through the links below:

- Simulation Script: https://goo.gl/dfKf95;
- Provenance Report (PNG): https://goo.gl/ffnFOP;
- Provenance Report (RDF): https://goo.gl/zQSVQ0.

To query the provenance metadata about this dataset, we can also do a GET request to the resource /provenance/ of the provenance services, using the dataset URI as an input parameter:

http://[domain]/api/provenance/**http://ib.usp.br/data_products/150**.

4. Conclusion And Future Works

In this paper, we presented a provenance-based approach for managing the provenance metadata of environmental monitoring processes, supported by a computational architecture designed for distributed and heterogeneous environments.

To validate our approach, an implementation of our computational architecture was used to manage the provenance metadata generated during the execution of an environmental monitoring simulation. After the simulation, the provenance services allowed us and other scientists to query the provenance metadata about the generated data, its generation process, and additional relevant information.

The results show that our approach is effective in collecting and storing provenance metadata and allows the query of an entire provenance of data products, thus enabling discovery and visualization of raw data, processes, and scientists involved in its generation and evolution.

We intend to carry out more case studies, capturing data provenance using other mechanisms and real devices. Moreover, we will continue the definition of a domain-specific application profile for environmental monitoring, including dictionaries of the environmental sciences community.

Another future work is the analysis of a strategy to the management of security and privacy of provenance metadata with sensitive information. Some initiatives need to restrict the public sharing of information about their processes. Therefore, a management approach to sensitive metadata is necessary.

With these efforts, we will seek to create a standardized approach that can be used for a variety of long-term researches and environmental monitoring programs, ensuring reliability of data, interoperability, reusability, and the maintenance of the data origin and its context.

References

[1] Lovett, G. M., Burns, D. A., Driscoll, C. T., Jenkins, J. C., Mitchell, M. J., Rustad, L., Shanley, J. B., Likens, G. E., and Haeuber, R., "Who needs environmental monitoring?," *Front. Ecol. Environ.*, vol. 5, no. 5, pp. 253–260, Jun. 2007.

[2] Lindenmayer D. B., and Likens, G. E., "Adaptive monitoring: a new paradigm for long-term research and monitoring," *Trends Ecol. Evol.*, vol. 24, no. 9, pp. 482–486, Sep. 2009.

[3] Frank, N. H., Mazzotti, J., "Why Do We Need Environmental Monitoring for Everglades Restoration?," 11-Dec-2013.

[4] Klein, D. J., McKown, M. W., and Tershy, B. R., "Deep Learning for Large Scale Biodiversity Monitoring."

[5] Baker, E., "Open source data logger for low-cost environmental monitoring," *Biodivers. Data J.*, vol. 2, p. e1059, Feb. 2014.

[6] Lehning, M., Dawes, N., Bavay, M., Parlange, M., Nath, S., and Zhao, F., "Instrumenting the earth: next-generation sensor

networks and environmental science, *Microsoft Res.*, Oct. 2009.

[7] Hills, D., Downs, R. R., Duerr, R., Goldstein, J., Parsons, M., Ramapriyan, H. (2015). The Importance of Data Set Provenance for Science, *Eos*, vol. 96, Dec. 2015.

[8] National Information Standards Organization (U.S.) (2004). *Understanding metadata.* Bethesda, MD: NISO Press.

[9] Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., and Gemeinholzer, B. (2012). The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —, *Ecol. Inform.*, vol. 11, p. 25–33, Setember 2012.

[10] Moreau, L., Groth, P., Cheney, J., Lebo, T., Miles, S. (2015). The rationale of PROV, *Web Semant. Sci. Serv. Agents World Wide Web*, V. 35, p. 235–257, Dec. 2015.

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

[12] Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., de la Hidalga, A. N., Vargas, M. P. B., Sufi, S., Goble, C. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud, *Nucleic Acids Res.*, 41 (W1). W557–W561, Jul. 2013.

[13] Altintas, I., Barney, O., Jaeger-Frank, E., (2006). Provenance Collection Support in the Kepler Scientific Workflow System, *In*: *Provenance and Annotation of Data*, L. Moreau and I. Foster, Eds. Springer Berlin Heidelberg, 2006, p. 118–132.

[14] Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., Vo, H. T. (2006).VisTrails: Visualization Meets Data Management, *In*: *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2006, p. 745–747.

[15] Belhajjame, K., Far, R. B., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., Tilmes, C., Moreau, L. (2013). PROV-DM: The PROV Data Model. *W3C Recommendation*, 30-Apr-2013.

[16] Baker, T., Dekkers, M., Heery, R., Patel, M., Salokhe, G. (2001). What Terms Does Your Metadata Use? Application Profiles as Machine-Understandable Narratives, *Int. Conf. Dublin Core Metadata Appl.*, p. 151–159, Oct. 2001.

[17] Ramapriyan H. K., Moses, J. F. (2013). NASA Earth Science Data Preservation Content Specification, National Aeronautics and Space Administration (NASA), Goddard Space Flight Center, Greenbelt, Maryland, 423-NaN-1, 2013.

[18] Fielding, R. T. (2000). Architectural styles and the design of network-based software architectures, University of California, Irvine.

[19] Sporny, Manu., Longley, Dave., Kellogg, Gregg., Lanthaler, Markus., Lindstrom, Niklas (2014). JSON-LD 1.0: A JSON-based Serialization for Linked Data, 16-Jan-2014. [Online]. Available: https://www.w3.org/TR/json-ld/.

[20] Cyganiak, R., Wood, D., Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax, *W3C Recomm.*, Feb. 2014.