A 3-Layer Model for Metadata

Keith G Jeffery
Consultant, UK
keith.jeffery@
keithgjefferyconsultants.co.uk

Anne Asserson
University of Bergen,
Norway
anne.asserson@fa.uib.no

Nikos Houssos
EKT, Greece
nhoussos@ekt.gr

Brigitte Jörg
Consultant, UK
brigitte.joerg@gmail.com

Abstract

We present a 3-layer model for metadata of which the key component is CERIF in the middle, contextual, layer. CERIF forms the lowest, most detailed level of metadata information that is common across research objects such as datasets. Its richness of representation makes it a superset over many other metadata formats allowing their congruent generation from CERIF. CERIF is used in42 countries and is an EU Recommendation to member States.

Keywords: research information; metadata; dataset; discovery; contextual; detailed; schema; syntax; semantics

1. Research and Research Information

It is generally accepted that research leads to wealth creation and improvement in the quality of life. In order for this to be effective information about the research needs to be collected, made available, communicated and curated. It can then be used by researchers for managing their CV, bibliography, generating web pages and finding collaborators. It can be used by research managers for evaluation and benchmarking as well as managing intellectual property of an organisation. It can be used by innovators to take research ideas through to products and services. It can be used by the media to communicate 'research stories' and by citizens interested in research and in 'citizen science'.

One research product is research datasets – and associated software. Information about the dataset (metadata) is needed for discovery and use. To provide the end-user with some assurance there is a need to understand the context of the dataset – why it was collected, by whom, under what conditions and using what equipment at which organisation. Also it is useful to know how the dataset relates to the purposes of the project, the funding and related scholarly publications (both white and grey). All of this contextual information assists the end-user in judging the applicability and quality of the dataset for their (re-)purposing.

2. Metadata

Sometimes unhelpfully described as 'data about data'[1] in fact metadata is data. Consider an electronic library catalog card system: for the researcher the library card is metadata for discovering the book or article of interest. For the librarian the card system can be used as data to analyse the relative completeness of the collections by subject, by publisher, by year etc.

Metadata has a long history especially in libraries and museums. It has been used extensively in information processing and especially in database technology since the 1960s but enjoyed prominence with the widespread adoption of WWW.

[1] http://dublincore.org/metadata-basics/

There are many classifications of metadata; the NISO (National Information Standards Organisation in the US) classification into descriptive, structural and administrative conflicts with a classification into linear, planar and hierarchic while OAIS (Open Archival Information System) has a reference model giving priority to digital preservation. In fact metadata encompasses also aspects of terminology (vocabularies, dictionaries, thesauri, ontologies), ensuring data integrity by constraints (rules) and managing rights such as copyright or access rights.

3. The Problems with currently used Metadata Formats

The currently most-used metadata formats for research datasets are individual, personal and non-interoperable. Where an attempt has been made to use a recognised format the most common are DC (Dublin Core), DCAT (Data Catalog Vocabulary) and – for geospatial data – INSPIRE. In the open government data domain CKAN (Comprehensive Knowledge Archive Network) is becoming more widely used (but still representing only about 4% of metadata in European data.gov sites.)

All these metadata standards are linear or flat and based more-or-less on DC. This leads to many problems. Although over the years DC has moved in syntax terms from text through HTML to XML and now RDF very few implementations use the richer syntax. Similarly and in parallel the semantics have moved from none to namespaces and on the ontologies linked with RDF syntax. Even in the richest DC there are problems. For example the element <source> is in fact one restricted example of <relationship>. Similarly <creator> and <contributor> can be ambiguous, and may be persons, organisations or services. Many elements are free text and so while computer readable are not computer understandable. However, the most serious problems are:

1) they violate basic principles of information integrity; they have elements which do not depend functionally on the uniquely identified metadata record. For example if <creator> in DC is a person then that person exists independently of whether they create a publication or not, and the person will have multiple other relationships in the research domain (maybe reviewer, project leader, employee…);
2) they store event flags or dates in the metadata e.g. 'date of publication'. In fact this is better (and more accurately) represented as a relationship between the publication and the organization (publisher) with date/time information to avoid the user failing to update the value of the date field every time an event occurs in the associated workflow;
3) they do not handle well multilinguality and multiple linguistic versions of the same text field;
4) they do not manage well versioning and provenance – this requires time-stamped relationships between one research information entity and another e.g. between person in role author and publication or between organization and organization (recursively) if there is a change of name, status or other parameter or between successive versions of a research dataset;
5) they do not allow multiple classification schemes for the same entity or – more generally – multiple terminology schemes for the same attribute of an entity;
6) they do not provide mechanisms for crosswalking between different vocabularies;
7) they do not provide extension mechanisms that preserve interoperability;

4. CERIF

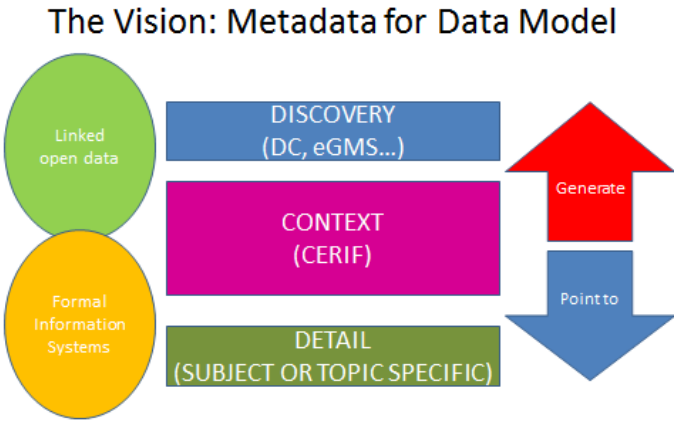
CERIF (Common European research Information Format) is an EU recommendation to Member states and used in 42 countries. It is maintained, developed and promoted (at the request of the

EC) by the euroCRIS community (www.euroCRIS.org). CERIF is contextual metadata. Its key features are:

1. it separates base entities (e.g. project, person, organisation, publication) from linking entities which link together instances of 2 base entities with a role (author, employee, project leader) and temporal interval of validity. This is much more advanced in semantics and integrity than hypermedia models, the use of XLINK or LOD (Linked Open Data);
2. it has formal syntax and declared semantics: it separates all terms into a semantic layer referenced from the syntax (so in link entities the role is a pointer to the semantic layer and in base entities list-restricted attribute values such as country code are in the semantic layer). This ensures consistency and integrity;
3. the linking mechanism also applies in the semantic layer so terminology schemes and the terms within them can be related with role and temporal duration. This allows semantic crosswalking for interoperability;
4. the richness of CERIF means it can act as a superset interoperation hub for other metadata formats, generating them congruently from the CERIF format;

5. A 3-Layer Model for Metadata

As indicated above the vast majority of metadata formats are particular to a domain or project or even a dataset. This means it is not possible to find a general metadata format. Thus the requirement is to find the lowest (most detailed) format that is valid across all datasets. euroCRIS has proposed CERIF - as contextual data – for that role and using CERIF to generate discovery data (in DC, DCAT or other such formats) and in turn to point to detailed (specific, schema level) metadata associated with each dataset.



This architecture has been used in several projects and appears to work well. It also provides a mechanism for bringing together the worlds of LOD/semantic web and conventional information processing combining the benefits of easy browsing and linking with formality and precision.

FIG. 1. A 3-Layer Model for Metadata

Cross-Domain Metadata Interoperability: Lessons Learnt in INSPIRE

Andrea Perego Michael Lutz Massimo Craglia Silvia Dalla Costa

European Commission – Joint Research Centre, Italy
{firstname.last-name}@jrc.ec.europa.eu

Abstract

Since 2007, EU Member States have been involved in creating an infrastructure for spatial information in Europe (INSPIRE), based on a legal and technical interoperability framework. This paper presents some of the lessons learnt during the implementation of this infrastructure (which started in 2009) and during work on data and service interoperability coordinated with European and international initiatives. We describe a number of critical interoperability issues affecting both scientific and government data and metadata, and propose how these problems could be effectively addressed by a closer collaboration of the government and scientific communities, by taking advantage of their complementary competencies, and by influencing the development and adoption of standards.

Keywords: INSPIRE; metadata; interoperability; multilingualism; licensing; data quality; provenance; lineage; versioning; persistent identifiers; harmonisation; standardisation.

1. Background

INSPIRE is a Directive (OJ, 2007)[1] of the European Parliament and of the Council aiming to establish a EU-wide spatial data infrastructure to give cross-border access to information that can be used to support EU environmental policies, as well as other policies or activities having an impact on the environment. The actual scope of this information corresponds to 34 environmental themes[2], covering also areas having cross-sector relevance – e.g., addresses, buildings, population distribution and demography.

In order to ensure cross-border interoperability of data infrastructures operated by EU Member States, INSPIRE sets out a framework based on common specifications for metadata, data, network services, data and service sharing, monitoring and reporting. Such specifications consist of a set of implementing rules (i.e., legally binding legislation), along with the corresponding technical guidelines.

The datasets, dataset series and services that make up the INSPIRE infrastructure can be discovered based on harmonised metadata and catalogue services (called "discovery service" in INSPIRE) giving access to this metadata. The INSPIRE metadata schema is defined in the INSPIRE Metadata Regulation (OJ, 2008), and it includes a number of elements relevant for resource discovery. Further metadata elements for evaluation and use are defined in the INSPIRE Regulation on interoperability of spatial data sets and services (OJ, 2010). Some of the key features include support to cross-language and spatial search, and semantic annotations based on controlled vocabularies and thesauri.

Following the INSPIRE implementation roadmap[3], since December 2010 EU Member States are making available INSPIRE metadata, which since November 2011 have to be published through INSPIRE discovery services. These discovery services are used by the INSPIRE

[1] The full list of the legal and technical documentation concerning INSPIRE is available from the INSPIRE Web site: http://inspire.ec.europa.eu/
[2] For the list of INSPIRE themes, see: http://inspire.ec.europa.eu/index.cfm/pageid/2/list/7
[3] See: http://inspire.ec.europa.eu/index.cfm/pageid/44

Geoportal[4], operated by the European Commission, to harvest and index metadata, thus providing a single access point to discovery INSPIRE data and services from EU Member States.

It is worth noting that, although INSPIRE focuses on environmental data, the majority of INSPIRE metadata elements are generic enough to describe also other types of data and services (a summary of the elements of the INSPIRE metadata schema is provided in Figure 1). Moreover, INSPIRE metadata are currently being used for describing data and services of both public administrations and research organisations.

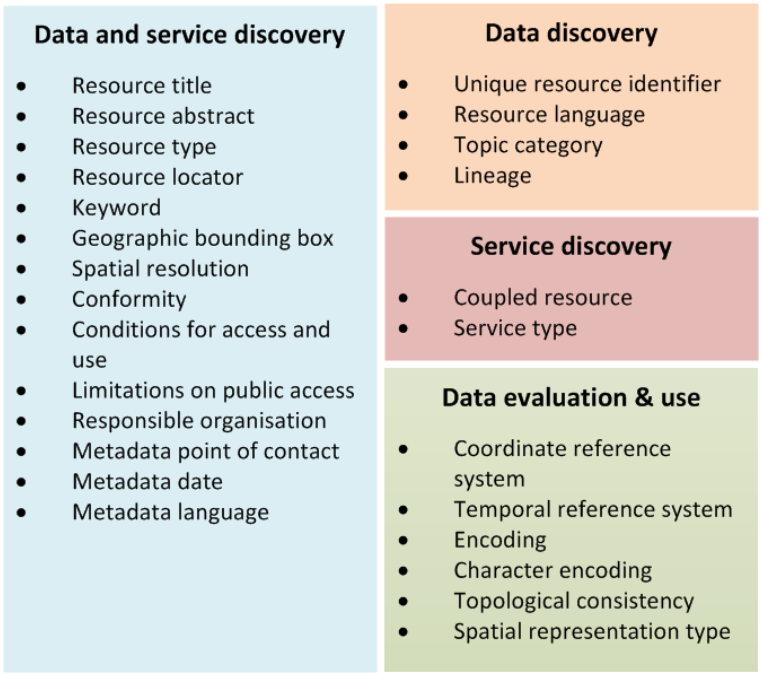


FIG. 1. INSPIRE metadata elements at a glance

For these reasons, the Joint Research Centre of the European Commission (JRC), as technical coordinator of the implementation of INSPIRE, is collaborating in a number of European and international initiatives concerning best practices and cross-domain interoperability of government and scientific metadata (for a general overview, see Perego et al., 2012). These initiatives include a number of working groups chartered in the framework of Research Data Alliance[5], and work in the European Commission on the definition of core vocabularies for public administrations[6], a vocabulary and framework for sharing semantic interoperability assets[7], and of a common metadata interchange format for European data portals, based on the Data Catalog Vocabulary of W3C[8]. At the same time, JRC is investigating possible extensions to INSPIRE metadata in order to address domain-specific requirements for scientific data, in the scope of a number of activities concerning EU institutions and Member States.

2. Outstanding interoperability issues

One of the main lessons learnt in INSPIRE and in the initiatives JRC is involved in, is that government and research data are not two separate worlds. Although they may have different and

[4] The INSPIRE Geoportal is available at: http://inspire-geoportal.ec.europa.eu/
[5] See: https://rd-alliance.org/
[6] See: http://joinup.ec.europa.eu/community/core_vocabularies/description
[7] See: http://joinup.ec.europa.eu/asset/adms/description
[8] See: http://joinup.ec.europa.eu/asset/dcat_application_profile/description

domain-specific requirements, their scopes are overlapping. Also, government data are commonly used as a basis to create scientific data, and vice-versa. Consequently, it is fundamental to adopt a consistent approach to address interoperability issues shared by both government and scientific data.

In particular:

- Controlled vocabularies are a key component to support semantic and cross-language data discovery across domains. However, what is missing is a framework for the (collaborative) maintenance and publication of controlled vocabularies, providing support to multilingualism (cross-language indexing and search), versioning (backward / forward interoperability) and mapping - both between terms in the same vocabulary and in different ones (cross-domain interoperability).
- The lack of a consistent approach to licensing is one of the key issues that prevents effective data re-use. Possible solutions include transparent and machine-readable representations of licences for data and services, and a common protocol to digital rights management, for resource discovery, access, and use.
- The ability to verify the quality of data is fundamental both in a government and scientific context. This may be addressed by adopting transparent and machine readable representations of data provenance, lineage, and use. Other features may include support to users' feedback and quality rating (possibly also through third-party quality certification).
- Data may be subject to changes and updates, therefore support to data versioning would grant access to historical data. This is important, for instance, when decisions or predictions based on given data result to be incorrect. In such cases, the ability to restore the original data is fundamental. The widespread use of persistent identifiers may play an important role also to address this issue.

Notably, the government and research communities have specific competencies that can be re-used to address some of such issues. An example is the long and consolidated tradition concerning data archiving and curation typical of the scientific community, which can be adapted to government data, without re-inventing the wheel.

Moreover, such issues are currently addressed by a number of initiatives, most of them running in parallel, sometimes ignoring relevant work carried out in other communities and often proposing not interoperable or conflicting solutions. The government and scientific community could create a critical mass that might promote a better coordination towards the definition and adoption of effective and consistent solutions for cross-domain interoperability.

References

OJ. (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). OJ L 108, 25.4.2007, p. 1–14.

OJ. (2008). Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata (Text with EEA relevance). OJ L 326, 4.12.2008, p. 12–30.

OJ. (2010). Commission Regulation (EU) No 1089/2010 of 23 November 2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services. OJ L 323, 8.12.2010, p.11.

Perego, Andrea, Cristiano Fugazza, Lorenzino Vaccari, Michael Lutz, Paul Smits, Ioannis Kanellopoulos, and Sven Schade. (2012). Harmonization and Interoperability of EU Environmental Information and Services. IEEE Intelligent Systems, 27(3):33-39, May-June 2012. doi: 10.1109/MIS.2012.22

Usage data for metadata properties to support open data registries and semantic wikis

Muriel Foulonneau
PRC Henri Tudor, Luxembourg
muriel.foulonneau@tudor.lu

Sébastien Martin
Université Paris 8, France
sebastien.martin84@gmail.com

Jacques Ducloy
France
jacques.ducloy@loria.fr

Thierry Daunois
Université de Lorraine, France
thierry.daunois@inpl-nancy.fr

Slim Turki
PRC Henri Tudor, Luxembourg
slim.turki@tudor.lu

Abstract

Metadata and ontology repositories are critical to ensure the discovery of existing vocabularies and the reuse of vocabularies and/or individual properties. However, these infrastructures should take into consideration the decision making process and criteria for the selection of a vocabulary of individual concept or property. Usage data in particular are important and can reassure on the maintenance of the vocabulary by a third party. This data is to some extent available through dedicated tools, such as semantic search engines. We illustrate the need for integrating usage data in the vocabulary infrastructures in order to support the reusability of vocabularies and therefore interoperability and data usability in science.

1. Introduction

Research is becoming more and more data centric (Hey, 2012). Researchers make use of many datasets which are available on the Web, including social data and open data sets, leading to the use of tools for manipulating large quantities of data (Big data). At the same time the development of open science and open access to scientific publications and datasets has led scientists to integrate more and more of their production the Web infrastructure. In this context, the description of datasets with common metadata models and domain ontologies to represent knowledge are critical to the reuse manipulation, representation, and exchange of data and knowledge. They constitute key challenges of the Web based research infrastructure.

Despite the widespread use of standard vocabularies, most implementers have a need for tailoring them or adding properties. Instead of creating a new homemade property, they can reuse an existing metadata property.

While metadata and ontology repositories provide information on the structural context of metadata properties, i.e., the vocabulary they come from, it is difficult to identify their maintenance conditions and their actual usage context.

In this paper, we propose a mechanism to provide this type of metadata in order to inform metadata implementers in their choice. We also illustrate some of the difficulties related to the reuse of vocabularies and individual properties and concepts that make this type of data very relevant through two projects we have carried out.

2. Reusing metadata properties

Despite the dissemination of standard vocabularies and the availability of metadata registries and ontology repositories, the reuse of vocabularies and individual metadata properties reuse is still not matching expectations. This decreases significantly data reusability. We show these challenges through two projects on Open data reuse and on the dissemination of scientific resources through scientific wikis.

2.1 Reusing vocabularies to describe Open data sets

In the scope of a study on Open Data, we have gathered statistics on metadata sets published on the Public Data EU catalogue[i] which federates multiple catalogues in Europe in order to identify the level of openness of datasets as well as the variety of the data providers. The metadata collected in May 2013 on 17.027 datasets used 236 properties. Indeed the JSON interface of the catalogue provides

data from original catalogues. A lot of information is therefore expressed through many distinct properties suggesting that the metadata properties were not well reused from existing catalogues.

TABLE 1. Distinct values by license field

Fields	Distinct values	Fields	Distinct values
licence	140	License_summary	55
License	15	License_title	26
license_url	5	License_uri	2
License_details	34	License_url	14
License_ID	26	mandate	29

Licensing information for instance can be found in 10 different metadata properties (TABLE 1). 79% of the metadata records have a meaningful value for at least 1 of these metadata properties. Nevertheless, the usefulness of a metadata property with a value in only 2 cases is very limited. This illustrates the lack of reuse of metadata properties.

2.2. Reusing vocabularies for a semantic Wiki

In the scope of the Wicri network[ii] of scientific semantic wikis (Ducloy et al., 2010), a set of properties had to be defined to create the semantic layer of wikis. In this network, a given topic (e.g., "metadata") can be studied on several wikis, displaying several points of view (e.g., "Computer science" or "Information science"), and/or in several contexts (e.g., France, Luxembourg, Europe, etc). Thus, on each wiki, a specific ontology is needed, dealing with the wiki's thematic, in addition to a common ontology in order to guarantee coherency and semantic interoperability among the whole network.

Many properties can be reused from existing ontologies. For instance, "Has PC member" comes from " semanticweb.org" to link a conference with a person, PC member for this event. In the same way, many terms can be reused from existing vocabularies, for instance EuroVoc as global vocabulary.

Nevertheless, identifying such properties is not always easy and, mainly, a consistent work of customization is requested. As with many examples of reuse of ontologies, special situations require adaptations (usually extensions). For instance, "Has PC member" must be extended if a large number of events get diverse committees. The Wicri framework, and especially the network, introduces a new set of problems. For instance, on a given wiki, it could be necessary to use simultaneously MeSH and EuroVoc, two vocabularies using common concepts, but with different terms (and different relationships). A new set of problems emerges when considering the interoperability between Wicri and the semantic Web : for instance, naming entities is not obvious, when the French Wikipedia refers to "Luxembourg (ville)" where most bibliographic bases refer to Luxemburg. Next step in this process, we now intend to use semantic features of Wicri in association with bibliographic metadata, as learning set for data mining among large corpus : this raises again new issues, displaying a completely new "landscape" of reusing metadata and properties.

3. Usage data to support the selection of vocabularies

Vocabularies usually include definitions of concepts and properties, their identifiers, and their relation with other concepts and properties. Knowledge Organization Systems may include scope notes as well as hierarchical or associative relations. The presentation of metadata models and ontologies including the objective and context for which they were conceived can usually be retrieved from the Web.

However, this does not inform on vocabulary maintenance. This is usually deduced from the source (e.g., the Library of Congress may be considered more reliable for maintaining vocabulary than a university laboratory which created a vocabulary for a particular project). The maintenance is particularly critical in order to enable data linkage for Linked Open Data for instance, since data publishers have to rely on the fact that the meaning of a concept they are using remains the same over time and that its description remains accessible. The lack of versioning mechanism represents a major weakness of the current Web infrastructure to support data linkage (Van de Sompel et al., 2010). The maintenance conditions are therefore critical to the decision making process, just like they are for software selection (Ruth, 2008).

The ability to relate to a community of metadata implementers who have implemented a property, or vocabulary is in this regard very important. The popularity of the resource is however not provided in metadata registries and ontology repositories such as the NSDL registry, the Dublin Core Metadata Registry, and the TONES ontology repository[iii].

In order to assess an open source software project, production data (e.g., new releases), usage data (e.g., number of downloads), and communication data (e.g., mailing list activity) can be analyzed (e.g., Wynn, 2003; Ahmed et al., 2010).

Regarding metadata models, properties, vocabularies, and ontologies, production data are represented by potential versions which can to a certain extent be found in metadata registries. There is usually no bug report system or module added to the metadata model. Communication data can be retrieved from mailing lists activity for instance which have to be analyzed. Usage data relate to the type and number of implementers and by the type and number of collections which have implemented the properties and vocabularies.

Usage data on vocabularies can be found through semantic search engines for properties and concepts which have been published using one of the Semantic Web standards, including RDFa, RDF, microformats, Schema.org, and OpenGraph for instance. The Sindice search engine[iv] for instance suggests known properties from a term (FIG 1**Error! Reference source not found.**). It provides the number of documents it has indexed which contain this property as well as the list of documents which can provide initial insight on the type of actors who use this property.

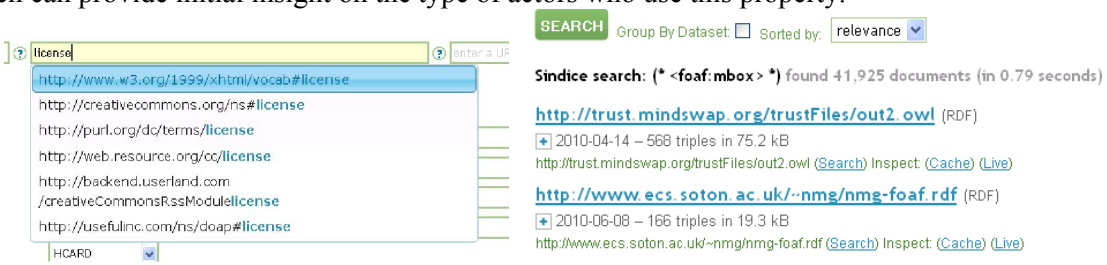


FIG 1. Occurences of the <foaf:mbox> property on Sindice

4. Conclusion: including usage data in the data infrastructure

Usage data can therefore be found through semantic search engines for instance, while structured information on the vocabularies can be retrieved from metadata registries and ontology repositories. Registries should include tools to automatically capture usage data and potentially general information on communication data, such as the liveliness of mailing lists that support the vocabularies. We suggest integrating these two infrastructures in order to support the decision making process of the metadata model, property, vocabulary, and ontology selection, thus facilitating reuse and improving the interoperability and accessibility of data.

References

Ahmed, F., Campbell, P., Jaffar, A., & Capretz, L. F. (2010). Myths and realities about online forums in open source software development: an empirical study. *Open Software Engineering Journal*, *4*, 52-63.

Ducloy, J., Daunois, T., Foulonneau, M., Hermann, A., Lamirel, J. C., Sire, S., & Vanoirbeek, C. (2010, September). Metadata for Wicri, a network of semantic Wikis for communities in research and innovation. In *International Conference on Dublin Core and Metadata Applications* (pp. 94-102).

Hey, T. (2012). The Fourth Paradigm–Data-Intensive Scientific Discovery. *E-Science and Information Management*.

Ruth, N. (2008). A Multi criteria decision making support to software selection. Master thesis, Makerere University

Van de Sompel, H., Sanderson, R., Nelson, M.L., Balakireva, L., Ainsworth, S., Shankar, H. (2010) An HTTP-Based Versioning Mechanism for Linked Data. Proceedings of the 3rd Workshop on Linked Data on the Web (LDOW2010)

Wynn, D. E. (2003). Organizational structure of open source projects: A life cycle approach. In *Abstract for 7th Annual Conference of the Southern Association for Information Systems, Georgia*.

[i] http://publicdata.eu
[ii] http://ticri.univ-lorraine.fr/wicri.fr/index.php?title=Accueil
[iii] http://owl.cs.manchester.ac.uk/repository/
[iv] http://sindice.com

Provenance Central: More Mileage from Provenance Metadata

Bertram Ludäscher
UC Davis, USA
ludaesch@ucdavis.edu

Paolo Missier
Newcastle University, UK
paolo.missier@ncl.ac.uk

Members
DataONE Provenance
Working Group

Abstract

Provenance has long been recognized as an important form of metadata that helps to "get more mileage" from data. Provenance describes the lineage and processing history of data and can be used to better understand and interpret data products, to assess and improve their quality and fitness for use, for debugging and reproducibility of results, etc. Important sources of provenance generation are scientific workflow systems and other controlled environments and cyberinfrastructure that can be instrumented to capture provenance.

We argue that to get the most value out of provenance it is critical to provide *provenance integration* and *analysis* capabilities. For the former, we are developing D-PROV, an extension of the W3C standard PROV that enriches the generic PROV model with important observables from scientific workflow systems and other provenance-enabled systems such as R. For the latter, we are developing PBase, a system prototype and associated language technologies to query and analyze provenance. PBase will be part of the as DataONE data preservation infrastructure for Earth Science Observation (www.dataone.org).

Our envisioned *Provenance Central* will be able to load and analyze provenance in order to connect data through its provenance with other datasets, workflows, and ontologies, but also with papers, scientific hypotheses, protocols, and users (i.e., authors and scientists). Discovering these connections requires *analytical techniques* that have not yet been applied to provenance. For example, since such provenance metadata will include, amongst other properties, data attribution information, we propose a novel type of analysis, which involves mining provenance through the entire repository, to elicit implicit *social connections* amongst the owners of the data. In summary, Provenance Central will be a new way of making data and social connections explicit, thus increasing data (re)usability in unprecedented ways.

Keywords: provenance; metadata; data integration; social network analysis.

Persistent Identifiers for Metadata Terms in a Crowd-Sourced Vocabulary

John Kunze
California Digital Library
USA
jak@ucop.edu

Greg Janee
UC Santa Barbara
USA
gjanee@ucop.edu

Christopher Patton
UC Davis
USA
cjpatton@ucdavis.edu

Abstract

Unique, persistent identifiers for vocabulary term concepts are critical for metadata (DC[1], SKOS[2], etc). This comes as no surprise to followers of Linked Data[3], for whom this first principle of the semantic web is a *sine qua non* for automatic reasoning with web content. It is even more important to metadata users who need a precise way to reference a particular concept when the term may have more than one definition.

Such is the case for the SeaIce Metadictionary[4], a crowd-sourced online dictionary of metadata terms in which multiple competing definitions are expected to be common and to co-exist indefinitely. Anyone can register and login in order to create new terms, edit their own terms, and comment and vote on others' terms. Typical use will be that someone, without logging in, searches for and inserts terms they find into metadata that they're creating to describe their own research. If unsatisfied with the terms that they found – or didn't find – they can login and take action, which means anything from up- and down-voting terms, commenting on others' terms, or adding and editing their own terms. Typical users will be research scientists trying to describe their datasets.

At the moment, SeaIce's internally generated database table row numbers are the only unique identifiers to disambiguate term concepts. As these numbers are not globally unique and not reproducible if the database is ever reloaded in a different order, we envisage creating another table column and populating it with stable identifiers from a recognized naming authority. To populate this column, we plan to extend SeaIce to mint identifiers from automatically from EZID[5] at the University of California, an established identifier service known for its open, identifier-scheme-agnostic architecture in support of research datasets and cultural heritage material.

In this presentation, we will discuss and seek feedback on a number of questions arising from the application of persistent identifiers (pids) to metadata elements. Questions include:

- Does it make sense to assign pids to element values as well as to element names?
- Should term relationships have pids? E.g., broader, narrower, synonym, antonym, etc.
- Research data relies on units, types, and other sub-vocabularies. Should they get pids?
- What should the experience of resolving a term pid be?

Participating in CAMP-4-DATA with attendees from diverse disciplines will provide valuable feedback for our development plan.

---

[1] Dublin Core Metadata Initiative. Expressing Dublin Core metadata using RDF. Jan 2008.
[2] Simple Knowledge Organization System
[3] Berners-Lee, T. et al. The Semantic Web. *Scientific American*. May 2001. p. 29-37.
[4] A product of the Preservation and Metadata Working Group within the NSF-funded DataONE project.
[5] UC Curation Center. Long-term identifiers made easy (n2t.net/ezid)

Separation of Concerns: PID Information Types and Domain Metadata

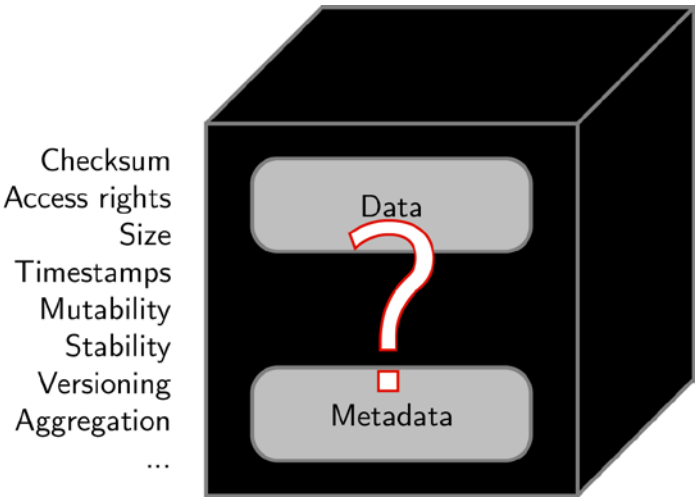
Tobias Weigel
DKRZ /
Universität Hamburg
Germany
weigel@dkrz.de

Timothy DiLauro
Data Conservancy /
John Hopkins University
USA
timmo@jhu.edu

We must define a pragmatic separation of concerns between metadata activities and the typed information associated with Persistent Identifiers. This distinction is important for ongoing debates within respective communities as well as in the RDA working groups.

From a data archive's viewpoint, a useful metaphor is that of the "black box" or "envelope": Data management is increasingly done by machinery rather than human users. So the machinery must know what to do with the boxes that come in through various channels, but it cannot open them for various reasons. We propose that metadata is a concern that is – from this particular view of automated data management – located inside the black box. A metadata description may actually be a black box object that must be managed just like all the others. Still, some information must be written on the outside of the box to be interpreted by the machinery. This information may be a subset of metadata, but it may also contain additional information not interesting as domain metadata.

These metaphors also work well with a technological layer stack view. Conceptual architectures such as the OSI model use distinct layers of abstraction. Understanding neighboring layers is not required; layers are defined through clear interfaces at the boundaries and can be changed independently. Consequently, PID information forms a lower layer, while metadata and all services working with them form a higher layer. Both layers are used by different actors, and there may be some transition of information between them. Services working independently in the framework of an individual layer will benefit from such an architectural division.



Ontology-Enabled Metadata Schema Generator: The Design Approach

Jian Qin
School of Information Studies
Syracuse University
jqin@syr.edu

Xiaozhong Liu
School of Informatics and Computing
Indiana University Bloomington
liu237@indiana.edu

Miao Chen
Data to Insight Center
Indiana University Pervasive Technology Institute
Mchen14@syr.edu

Metadata standards are important for normalizing descriptions of publications and research data and for information discovery and use. Large, complex metadata standards, however, can complicate the creation, sharing, and maintenance of metadata and incur high costs for metadata operations, especially in the domain of scientific data (Qin et al., 2010; Qin Ball, & Greenberg, 2012; Qin & Li, 2013). One strategy to solve the problems of large, complex metadata standards is to break them into independent modules to allow for reuse of elements and maximal possibility of automation. To implement this strategy, we need a metadata infrastructure that contains elements, vocabularies, and other metadata artifacts and that is easy to use. This short paper describes the design approach to an ontology-enabled metadata schema generator as part of the metadata infrastructure.

Elements in metadata standards in the scientific data domain tend to follow a pattern that a small number of (super-) general elements co-occur in a large number of standards and those co-occurred in 2-4 standards tend to be field-general. Even though semantically same elements co-occurred across different standards, they often varied in singular-plural forms, capitalization, or complete different words (Qin & Li, 2013). These inconsistencies and varying naming conventions can be mitigated by ontologies. These ontologies as the semantic underpinning for scientific metadata will have different types, e.g., entity ontologies for person, organization, project, study, dataset, and so on, or temporal ontologies for date and time. They will be built by following the portability principle (Qin, Ball, & Greenberg, 2012).

Many semantic resources have been made available in linked data format and can be utilized to avoid reinvent the wheels in the process of building a metadata infrastructure. Using the open identity metadata ORCID and ResearcherID as an example, FIG. 1 shows the design approach that uses the open identity metadata and portable metadata schemes in the form of ontologies as the input for the metadata scheme generator. The generator will then output the elements and relations selected by the user in the form of an RDF schema or other format.

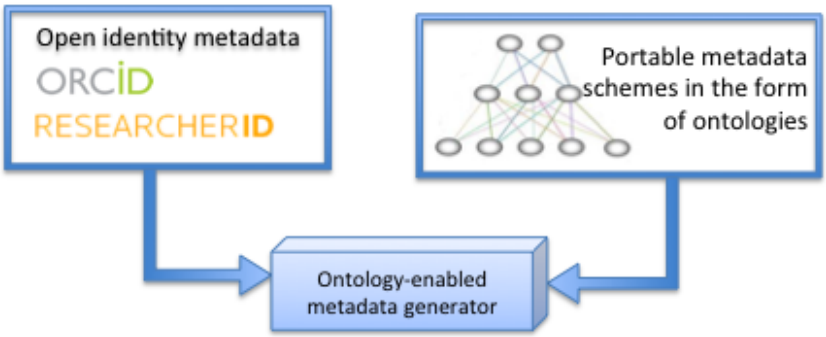


FIG. 1. A conceptual structure of an ontology-enabled metadata generator

Envisioning how such a metadata infrastructure might affect the metadata creation process, we developed a case scenario for the ontology-enabled metadata schema generator. The Ecological Metadata Language (EML) and Content Standard for Digital Geospatial Metadata (CSDGM) are two standards used to describe ecological data and geospatial data respectively. Both of them contain sections/modules of metadata elements and share some general descriptive elements. The EML eml-resource module contains elements such as title, creator, pubDate, abstract, keyword, and keywordThesaurus, which overlap with similar elements in CSDGM, namely title, originator, pubdate, abstract, keywords, and themekt. The two standards also share similar elements for describing time and space. For example, EML's eml-coverage module contains granular modules of "time", "beginDate", and "endDate", which correspond to similar temporal elements in CSDGM. The element "boundingCoordinates" is identical in both standards. These observations demonstrate that different metadata standards contain similar elements and the shared elements may be generalized as ontologies and reused in building new metadata schemas.

When a collection of ontologies is created, differences in naming semantically same elements can be smoothed out and structures established between elements based on their relations. A metadata generator can take the advantages of the ontology collection as well as external semantic resources for building customized metadata schemas. The ontology-enabled metadata generator will have the functions and capabilities to:

- Facilitate interactive metadata selection and structure definition in developing customized metadata schemas;
- Preload instances as the values for frequently used elements , e.g., project team members' names and affiliations, or project/study description;
- Automatically acquire identity information for entities and elements; and
- Build portable metadata to enable faster dissemination, access, and reuse.

To fulfill this design, there are more technical issues beyond metadata to be solved. For example, what kind of architecture for the back-end will be needed to enable a visualized, interactive, and dynamic front-end for the metadata schema generator? What are the boundaries between metadata modules and how will they affect the size and structure of ontologies? How will the ontologies be defined and maintained? For a novel thinking of metadata schema creation, we recognize that there may be more questions than answers at this early stage of development. Work is already underway for building a prototype as a proof of concept. We are hoping to have an example for the demo at the time of this workshop.

References

EML. http://knb.ecoinformatics.org/software/eml/

FGDC. http://www.fgdc.gov/metadata/geospatial-metadata-standards

Qin, J., A. Ball, & J. Greenberg. (2012). Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. *Dublin Core International Conference DC-2012, Kuching, Malaysia, September 3-7, 2012*.

Qin, J. & K. Li. (2013). How portable are the metadata standards for scientific data? A proposal for a metadata infrastructure. *Dublin Core International Conference DC-2013, Lisbon, Portugal, September 2-6, 2013*.

Qin, J., M. Chen, X. Liu, & A. Wiggins. (2010). Linking entities in scientific metadata. In: *Proceedings of the Dublin Core International Conference DC-2010*, Pittsburg, PA, October 20-22, 2010.

Metadictionary: Advocating for a Community-driven Metadata Vocabulary Application

Jane Greenberg, Angela Murillo,
Metadata Research Center, UNC,
USA
janeg@email.unc.edu,
amurillo@email.unc.edu,

Rob Guralnick
University of Colorado at Boulder
USA
robgur@gmail.com

Greg Janee
UC Santa Barbara
USA
gjanee@ucop.edu

John Kunze, California Digital Library
USA
John.Kunze@ucop.edu

Nassib Nassar
HW Odum Institute
UNC-Chapel Hill, USA
nassar@email.unc.edu

Christopher Patton
UC Davis
USA
cjpatton@ucdavis.edu

Sarah Callaghan, British Atmospheric Data Centre
UK
sarah.callaghan@stfc.ac.uk

Karthik Ram
UC Berkeley
USA
karthik.ram@berkeley.edu

Abstract

Metadata disorder and unnecessary costs are increasing due to the expanding population of scientific data schemes and standards. Metadata challenges are reviewed; and SeaIce, a community driven metadata vocabulary application, is introduced as a potential solution. SeaIce functions and development challenges are presented. CAMP-4-DATA participants are called upon to experiment with the SeaIce application and actively participate in a discussion targeting noted metadata challenges.

The Problem: Duplicative Metadata Efforts

Metadata is essential for managing research data. Scientists, data managers, and the full range of data information systems (e.g., repositories, grid computing, and cloud resources) rely on metadata to operate effectively. Today, driven by the digital data deluge, we find a plethora of discipline-oriented metadata standards supporting the same or similar functions (Willis, et al, 2012). For example, basically *all* descriptive metadata standards support discovery via topical subject terms/keywords; some include more granular properties for spatial and temporal data. Efforts establishing property semantics and defining content are duplicated time-and-time again, resulting in schemes that have marginal if any difference. The population of metadata standards that has emerged presents a disorder and cost concern, particularly given the overlap in supported functionalities.

Clearly overlap among metadata schemes aids interoperability, specifically data exchange and cross-system searching. Benefits aside, duplicative efforts incur unnecessary costs realized via the following:

- Metadata requires human and financial resources (Russom, 2010; Greenberg, et al, 2013).
- Intellectual demand and system development incur costs when aiming for metadata interoperability.
- Extending an existing scheme with new properties increases metadata costs.

Dublin Core Metadata Application Profiles (DCAPs)[1] and linked open data (LOD) can, on some level, help circumvent duplication and cost by leveraging existing metadata work. An approach built around virtual and social communities of practice may provide a complementary and alternative way to address these challenges.

The DataONE Preservation and Metadata Working Group (PAMWG)[2] advocates for a social approach to metadata vocabulary design. PAMWG has prototyped a metadictionary called **SeaIce**[3] that uses crowdsourcing for establishing metadata terms and engaging metadata stakeholders. The remainder of this paper introduces SeaIce, documents current features and goals, and discusses next steps. The last section of the paper calls upon CAMP-4-DATA participants to experiment with SeaIce and engage in a discussion to address metadata challenges.

Introducing SeasIce: Context for a Crowsourced Metadictionary

SeaIce Context

The SeaIce metadictionary is being developed to host community-driven metadata terms and definitions. Chief goals include reducing duplicative metadata activity and unifying metadata practices across disciplines. Functional requirements are presented in Table 1.

Table 1: Functional Requirements (Greenberg, et al, 2012)

Low barrier for contributions.
Transparency in the review process.
Collective team review, with rotating responsibilities among community members (scientists, developers, organizations, curators, etc.)
Consideration of elders (experts) to guide the review process and maintain thoughtful, balanced discussion.
Voting capacity of all users on the candidacy of terms submitted and their use.
Collective ownership of any user or organization.
Stakeholder engagement in the design and review process.

DataONE[4] serves as the target implementation community, although SeaIce has implications for any domain seeking to reduce duplicative efforts. DataONE is an ideal environment for launching SeaIce given the range of disciplines represented (e.g., ecology,

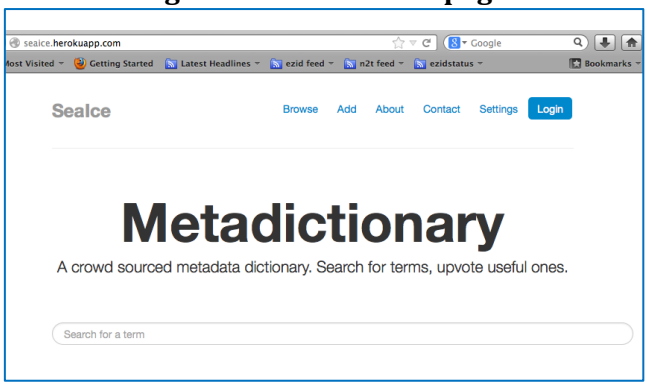
biology, geology, astronomy, etc., and the many sub-disciplines) and the diversity of metadata stakeholders (data creators, curators, system developers, and administrators).

DataONE is a community and a distributed framework providing steps toward a sustainable cyberinfrastructure. The SeaIce metadata dictionary supports this overriding goal by exploring an innovative means for a persistent and robust metadata infrastructure (Kunze, et al, 2013). By utilizing crowdsourcing techniques, the SeaIce metadictionary can help eliminate duplicative efforts, reduce associated costs, and provide an innovative framework for metadata interoperability across disciplines for stakeholder communities. The aim is a 'high-quality social ecosystem' in which the community of metadata stakeholders dialog, confirm terms and definitions, and unify metadata practices.

SeaIce – Prototype and Framework

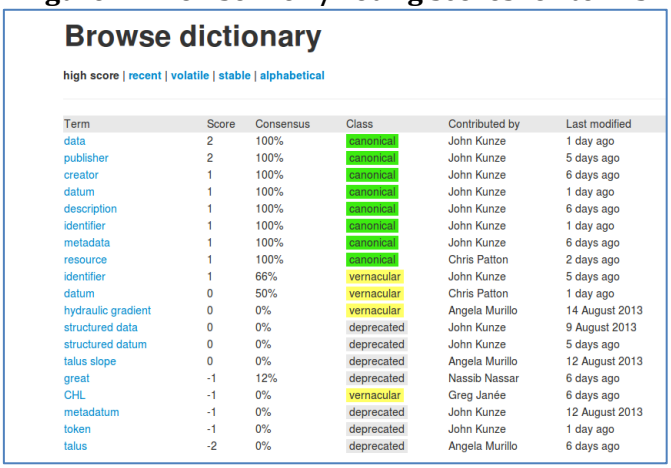
SeaIce is modeled on StackOverflow[5] and other social software services. Figure 1 presents the SeaIce homepage.

Figure 1: SeaIce Homepage



When logged in, users may vote terms 'up' or 'down' based on the definition and other aspects of importance; engage in online discussions about a term/definition/use, etc.; and propose new term(s) for discussion and voting. Figure 2, shows voting activity for a series of terms.

Figure 2: Browse View/Voting scores for terms



Browse dictionary

high score | recent | volatile | stable | alphabetical

Term	Score	Consensus	Class	Contributed by	Last modified
data	2	100%	canonical	John Kunze	1 day ago
publisher	2	100%	canonical	John Kunze	5 days ago
creator	1	100%	canonical	John Kunze	6 days ago
datum	1	100%	canonical	John Kunze	1 day ago
description	1	100%	canonical	John Kunze	6 days ago
identifier	1	100%	canonical	John Kunze	1 day ago
metadata	1	100%	canonical	John Kunze	6 days ago
resource	1	100%	canonical	Chris Patton	2 days ago
identifier	1	66%	vernacular	John Kunze	5 days ago
datum	0	50%	vernacular	Chris Patton	1 day ago
hydraulic gradient	0	0%	vernacular	Angela Murillo	14 August 2013
structured data	0	0%	deprecated	John Kunze	9 August 2013
structured datum	0	0%	deprecated	John Kunze	5 days ago
talus slope	0	0%	deprecated	Angela Murillo	12 August 2013
great	-1	12%	deprecated	Nassib Nassar	6 days ago
CHL	-1	0%	vernacular	Greg Janée	6 days ago
metadatum	-1	0%	deprecated	John Kunze	12 August 2013
token	-1	0%	deprecated	John Kunze	1 day ago
talus	-2	0%	deprecated	Angela Murillo	6 days ago

Modeled on StackOverflow, users may modify or delete their term and definition at any time. Once this occurs, those who have voted on the term will be notified. In addition, SeaIce provides listings of newly submitted terms, highly-rated terms, and highly-stable terms in order to guide users on which terms are ready for discussion and voting. Work is under way for SeaIce to provide a search mechanism that ranks highly-rated and highly-stable results.

SeaIce Features and Ongoing Development

SeaIce metadictionary presents a number of unique challenges not presented in other crowdsourcing environments. There are many social network systems rely on voting or ranking of answers. SeaIce is unique in accommodating a wide-array of stakeholders—data creators, curators, developers, administrators—anyone with a vested interest in metadata. The community of practice is quite diverse. Additionally, social technology is being used in SeaIce to identify a set of stable canonical terms; and these terms will form a common metadata practice specific to scientific data. This process must be fully automated and must reflect the consensus of the full stakeholder community. A central problem is that it is unlikely that every user will vote on every term. The PAMWG is exploring a heuristic for consensus based on user reputation. This heuristic involves *stability*, *class order* of term, and *voting impacts*. Ideas surrounding the heuristic functionality and SeaIce in general are captured in an open blog.* The percentages and time intervals presented directly below reflect truly preliminary considerations.

Stability

A term is considered stable if it meets two criteria: (1) the definition or term itself haven't been edited by the owner for some predefined period of time, and (2) the rate of change of the score drops below a certain threshold close to zero.

Classes

SeaIce has designated three term classes:

- Canonical - the set of stable terms with consensus over 75%.
- Deprecated - the set of stable terms with consensus under 25%. In the case that there is suitable replacement somewhere in the dictionary, we expect it will be standard practice to reference it in the deprecated term's definition.
- Vernacular - the set of unstable terms that cannot be classified as canonical or deprecated (unstable.)

---

* Christopher Patton's Blog is part of the Bi-level Metadata Registry Development project, DataONE 2013 Summer Internship program; see: https://notebooks.dataone.org/metadata-registry.

Voting and scoring

A SeaIce user may cast a single up or down vote on a particular term and they are permitted to change it at any time. Table 2 shows potential ways in which term classes may change. The weight of the vote is based on the ratio of his or her reputation to the sum of reputations of all users voting on the term. As the number of voters increases, the weights of the votes become more equitable. As a result, when a term has a small voting body, reputation is very important; this allows good terms to be promoted quickly and bad terms to be deprecated quickly. As the voting body increases a reputation loses significance. Reputation is used as a heuristic for consensus; and, therefore, the score becomes more equitable as the number of people with an opinion grows.

Table 2: Term Classes and Voting Impact

Vernacular → canonical -- term is stable after two days and consensus is above 75%.
Vernacular → deprecated -- term is stable after two days and consensus is below 25%.
Canonical → vernacular -- term has been updated, restabilized, and consensus has dropped below 75%.
Deprecated → vernacular -- term has been updated, restabilized, and consensus has risen above 25%.

Conclusion

Duplicative metadata efforts are not cost effective and require attention. SeaIce, a crowdsourced metadictionary, may help address this challenge and the disorder stemming from growing number of metadata schemes. SeaIce is in a development stage, and PAMWG members are experimenting with crowdsourcing metadata terms and definitions. Next steps include broadening participation and engaging others to experiment with SeaIce. The CAMP-4-DATA aims to "explore infrastructure design, applications, and policies that can advance the support of open, collective and sustainable access to metadata standards used for managing scientific data."[6] The SeaIce application fits this call, and DataONE PAMWG members welcome to opportunity to present SeaIce at the CAMP-4-DATA. We outline three key objectives for participants:

- Test the SeaIce application by entering a term(s)
- Test the voting mechanism for SeaIce by voting on a term(s)
- Engage in an open discussion with DataONE PAMWG members at the CAMP-4DATA.

In conclusion, demonstrating SeaIce and engaging in a discussion with international colleagues will allow the SeaIce effort to move forward the proof-of-concept. Further development SeaIce will allow PAMWG, DataONE, and other participants to contribute to the larger body of efforts addressing metadata challenges.

Acknowledgement

SeaIce and PAMWG are supported by the U.S. National Science Foundation (Grant #OCI-0830944).

References

Greenberg, J., Murillo, A., and Kunze, J.A. (2012). Ontological Empowerment: Sustainability via Ownership. Paper presented at the 23rd ASIS SIG/CR Classification Research Workshop, October 26, 2012, Baltimore, MD.

Greenberg, J., Swauger, S., and Feinstein, E. (2013). Metadata Capital in a Data Repository. Proc. Int'l Conf. on Dublin Core and Metadata Applications, 2-6, Sept., 2013, Lisbon, Portugal.

Kunze, J., Janee, G., and Patton, C. (2013, *in review*). Persistent Identifiers for Terms in a Crowd-Sourced Vocabulary. CAMP-4-DATA. Int'l Conf. on Dublin Core and Metadata Applications, 6, Sept., 2013, Lisbon, Portugal.

Russom, P. (2010). TDWI CHECKLIST REPORT: Cost Justification for Metadata Management. TDWI (The Data Warehousing Institute, Media, Inc.

---

[1] Dublin Core Application Profiles: http://dublincore.org/documents/profile-guidelines/.

[2] DataONE Preservation and Metadata Working Group: http://www.dataone.org/working_groups/data-preservation-metadata-and-interoperability-working-group.

[3] SeaIce Metadictionary: http://seaice.herokuapp.com/.

[4] DataONE: http://www.dataone.org/.

[5] StackOverflow: http://stackoverflow.com/.

[6] CAMP-4-DATA CfP: http://dcevents.dublincore.org/IntConf/index/pages/view/camp-4-data-cfp.

RUresearch - Open Source Metadata Application Profile and Research Object Handling for Research Data.

The Rutgers University Libraries have developed an open source workflow management system that includes a cataloging utility and a compound object handling system that enables the creation of metadata and intelligent object handling to fully support documenting and sharing research data. The cataloging system, which can be used independently and can work with any repository architecture, supports both MODS and Dublin Core metadata schemas. The MODS application profile includes an event-based subschema as a MODS extension schema, that can capture any useful event in the lifecycle of the data, from data capture, to data analysis, to data editing to data reuse. The application profile also includes elements for type of research, research methodology, type of data and type of subject, mapped to MODS and Dublin Core genre and subject elements. The data compound object supports documentation (lab notebooks, images, etc.) and instrumentation (data capture, data analysis, etc.). In addition to relating resources to each other using RDF, the resource handling also includes support for hierarchical file uploads, exactly as they are stored on the researcher's computer or server. The the metadata and object handling will be presented through examples from the RUcore (Rutgers Community repository) research data portal, RURsearch, http://rucore.libraries.rutgers.edu/research/

Authors: Cezary Mazurek, Marcin Mielnicki, Aleksandra Nowak, Krzysztof Sielski, Maciej Stroinski, Marcin Werla and Jan Wglarz

LEPSYDRA Data Aggregation and Enrichment Framework

Clepsydra is a flexible and scalable system for aggregation, processing and provisioning of data from heterogeneous sources. It was designed and developed to be a basis for services focused on aggregation and enrichment of (meta)data describing on-line collections of cultural heritage digital objects from Polish memory institutions. The first production deployment of this system is the PIONIER Network Digital Libraries Federation. While designing and developing the Clepsydra system we wanted to have the possibility to:

- store and access large amounts of heterogeneous data records;

- aggregate data records from many different kinds of sources;

- process data records from one format to another.

Clepsydra consists of three core components corresponding to the three groups of functional features:

1. Clepsydra Storage - stores and gives access to data records, utilizing NoSQL database.

2. Clepsydra Aggregation - gets data from various sources to Clepsydra Storage and is responsible for keeping Clepsydra Storage synchronized with the data sources. Aggregation activities are done by indepedent agents.

3. Clepsydra Processing - is responsible for processing data records stored in Clepsydra Storage according to dynamically configured rules. The processed data is stored back to Clepsydra Storage.

CLEPSYDRA is an open-source software available at http://fbc.pionier.net.pl/pro/clepsydra/.