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

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

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

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

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

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

Greg Janee UC Santa Barbara USA
gjanee@ucop.edu

Christopher Patton UC Davis USA
cjpatton@ucdavis.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[1], 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.

[1] The SeaIce project is also known as YAMZ (Yet Another Metadata Zoo). This change was instituted following the presentation of this work at the CAMP-4-DATA.
[2] Dublin Core Application Profiles: http://dublincore.org/documents/profile-guidelines/.

- Extending an existing scheme with new properties increases metadata costs.

Dublin Core Metadata Application Profiles (DCAPs)[2] 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)[3] advocates for a social approach to metadata vocabulary design. PAMWG has prototyped a metadictionary called SeaIce[4] 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. The DataONE Preservation and Metadata Working Group (PAMWG) advocates for a social approach to metadata vocabulary design. PAMWG has prototyped a metadictionary called SeaIce3 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[5] 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).

[2] Dublin Core Application Profiles: http://dublincore.org/documents/profile-guidelines/.
[3] DataONE Preservation and Metadata Working Group: http://www.dataone.org/working_groups/data-preservation-metadata-and-interoperability- working-group.
[4] SeaIce Metadictionary: http://seaice.herokuapp.com/.
[5] DataONE: http://www.dataone.org/.

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[6] and other social software services. Figure 1 presents the SeaIce homepage.

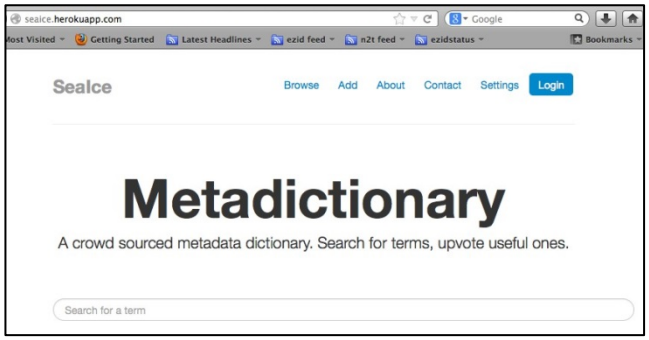
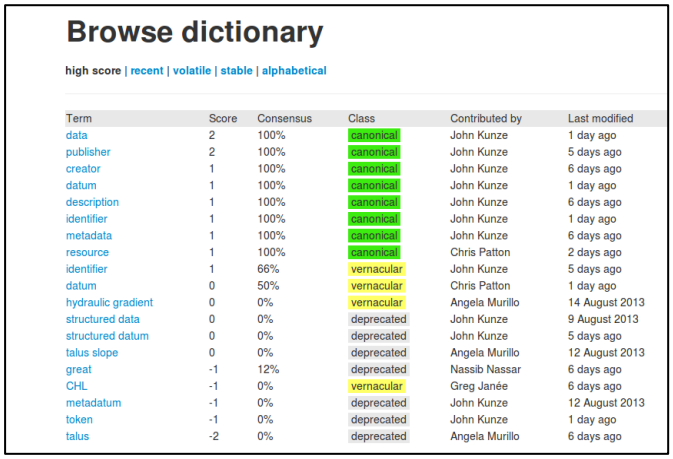


FIG. 1. Sealce Homepage

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.



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

FIG. 2. Browse View/Voting scores for terms

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

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

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.[7] 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.)

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.

[7] 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.

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."[8] 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.

Acknowledgements

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..

[8] http://dcevents.dublincore.org/IntConf/index/pages/view/camp-4-data-cfp.