Metadata and Semantics Workshop Summary

February 23-25, 2015

Indianapolis, Indiana

Acknowledgements

This report is a summary of the Metadata and Semantics Workshop that was held February 23-25, 2015 in Indianapolis, Indiana and funded by US-Research Data Alliance (NSF # 1349002) and DataONE (NSF #0830944). All the materials from this workshop are available on the RDA Metadata Interest Group web page.

We would like to thank Indiana University and the staff for their support of this workshop. In particular we thank Jenny Olmes-Stevens and Jodi Stern for their support with logistics both prior to and during the workshop. Thanks to Bolanie Solomon and Scott McCaulay for taking notes and Inna Kouper for help with facilitation. We would also like to thank Jamie Petta for her help with logistics and reimbursement and RDA-US for making this workshop possible.

The editors for this paper are Gary Berg-Cross (Spatial Ontology Community of Practice (SOCoP) and Rebecca Koskela (DataONE, University of New Mexico).

We would like to acknowledge the contributions of the workshop participants:

Laura Bartolo, Kent State University

Dave Dubin, University of Illinois Urbana-Champaign

Michel Dumontier, Stanford University

Peter Fox, Rensselaer Polytechnic Institute

Ted Haberman, HDF Group

Robert Hanisch, National Institute of Standards and Technology (NIST)

Pascal Hitzer, Wright State University

Krzysztof Janowicz, University of California Santa Barbara

Inna Kouper, Indiana University

Adila Krisnadhi, Wright State University

Marshall Ma, Rensselaer Polytechnic Institute

Emilio Mayorga, University of Washington

Scott Peckman, University of Colorado

Beth Plale, Indiana University

Ray Plante, National Data Service

Charles Vardeman, Center for Research Computing Notre Dame

Mary Vardigan, Interuniversity Consortium for Political and Social Research (ICPSR)

John Westbrook, Rutgers University

Ilya Zaslavsky, University of California San Diego

Contents

Executive Summary

The RDA Metadata and Semantics Workshop was held February 23-25, 2015 in Indianapolis, Indiana. The workshop was intended to provide an opportunity for outreach to communities not already acquainted with the Research Data Analysis (RDA). The workshop was also an opportunity for attendees to review and share current progress and experience with semantic metadata approaches, including linked data, the use of lightweight, opportunistic methods, bottom-up and top-down approaches, useful tools and ontology design patterns. The workshop explored the readiness of semantic approaches for addressing the variety of Big Data and metadata challenges and the handling of diverse data and infrastructure issues.

As an outreach effort there were a number of successes. These included involvement of the Biomedical community, leveraging some of their experience with semantics. There was also successful outreach to the Materials Genome Initiative where new work was proposed and indeed is underway. This included the launching of efforts to show the value of the Linked Data Concept by creating RDF connections to other linkable datasets such as in the Materials Measurement repository. In addition outreach with NDS will continue.

A number of the groups are considering using Semantic Web Technologies, Linked Data, and Ontologies for a variety of things including federated queries over multiple data sources based on the workshop. Two proposals were developed for submission to the RDA Data Share Fellow program, both affiliated with the Metadata Interest Group. Additional standards were added to the Metadata Standards Directory and new use cases were submitted for the RDA Metadata effort. In addition a process for utilizing existing ontologies such as Semanticscience Integrated Ontology was developed. A new ontology design pattern was developed for the Community Surface Dynamics Modeling System vocabularies.

Introduction to Workshop

Rationale and Purpose of the Workshop

Metadata is a core building block area of the Research Data Alliance (RDA) community as reflected in the numerous groups involved in this area. A summary of metadata interests and principles was articulated at the RDA Collaborative Working Group session held at National Institute of Standards and Technology (NIST) Nov. 13-14, 2014. As noted there documenting data, services, and workflows plays a key role for resource discovery, access, organization, interpretation, sharing and usage across many research areas. A strong interest was expressed for the improvement of metadata by formalizing its semantics including formalizing relations among metadata elements. A major issue concerns the identification of best practices for integrating semantic technology and its methods with traditional metadata approaches and technology. A small, hands-on workshop was proposed to address this by bringing together a core group of active, representative RDA members working on metadata along with people outside of RDA who are interested in and are engaged in efforts to improve semantics for metadata. Discussion and joint work included metadata use cases as well as some semantic technology practitioners to help with:

- Formalization of metadata by identifying possible practical paths for the needed formalizations and technological solutions for documenting data resources
- Aligning and integrating representations from different sources

- Identifying technological and/or social obstacles that may prevent formalization and integration
- Development of formal solution examples from submitted use cases and extant metadata

The workshop was intended to provide an opportunity for attendees to review and share current progress and experience with semantic metadata approaches, including linked data, the use of lightweight, opportunistic methods, bottom-up and top-down approaches, useful tools and ontology design patterns. It also provided an opportunity for outreach to communities not already acquainted with RDA. The workshop explored the readiness of semantic approaches for addressing the variety of Big Data and metadata challenges and the handling of diverse data and infrastructure issues.

Metadata

RDA Perspective and Metadata Context

The four 'core' metadata groups of RDA, Metadata Standards Directory Working Group (MSDWG), Metadata Interest Group (MIG), Data In Context Interest Group (DICIG), and Research Data Provenance Interest Group (RDPIG), have worked together to facilitate the use and reuse of data. As the results from a Science survey[1] showed, two major problems with the current state of scientific data management in a research lifecycle: a lack of funding and staff support for managing active data and the lack of metadata standards and tools for managing research data. Building on previous work with metadata standards, the core metadata groups have brought forward a set of principles for metadata that the groups believe RDA should adopt and promote. Those principles are:

- The only difference between metadata and data is mode of use
- Metadata is not just for data, it is also for users, software services, computing resources
- Metadata is not just for description and discovery; it is also for contextualisation (relevance, quality, restrictions (rights, costs)) and for coupling users, software and computing resources to data (to provide a Virtual Research Environment)
- Metadata must be machine-understandable as well as human understandable for autonomicity (formalism)
- Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…)

Some examples and implications of these principles are:

1. Consider a library catalogue stored electronically. To a researcher it is metadata – using the catalog finds the book or article. To the librarian it is data: she can count how many books or articles exist on biochemistry compared with clinical medicine.

2. In a VRE (Virtual Research Environment) the amount of work a researcher has to do manually does not scale. Autonomic services are required. In order to achieve this data, services, users and computing resources need to be described to middleware, which manages the scheduling, allocations, connection of the components, etc. These descriptions are metadata.

[1] Science Staff, "Challenges and opportunities," Introduction to special section Dealing with Data. *Science*, 11 February 2011: Vol. 331, pp. 692-693.

3. Metadata for discovery followed by manual selection and connection is already achievable. However the selection of appropriate datasets or software is greatly enhanced by using contextual metadata; that is metadata characterizing the object of interest. Contextual metadata concerns persons, organizations, projects, funding, outputs (publications, products, patents), facilities and equipment – in short attributes that allow the end user or software representing the end-user to assess the relevance and quality of an object (dataset, software) for their current purpose.
4. The mantra is formal syntax and declared semantics. This allows machine processing rather than manual processing.
5. Management metadata links with (3); the contextual metadata can also be used for evaluation of research, policy-making and other management functions at institutional or funding organization level.

Metadata Purposes

Metadata serves many uses: (1) data discovery, (2) contextualization, and (3) information for detailed processing[2]. The metadata for discovery must be sufficient for the human or computer system to find the dataset/data objects of interest. The higher the quality of the discovery metadata, the greater the accuracy and completeness of the discovery. Typical 'standards' in the discovery area are DC (Dublin Core)[3] and CKAN (Comprehensive Knowledge Archive Network)[4]. For contextual metadata one of the widely used 'standards' in Europe is CERIF (Common European Research Information Format)[5], which covers persona, organizations, projects, products (including datasets), publications, patents, facilities, equipment, funding and – most importantly – the relationships between them expressed in a form of first order logic with both role and temporal attributes. Software has been written to generate DC and CKAN from CERIF (and several other 'standards'). The detailed processing metadata is typically specific to a research domain or even an individual experiment or observation method. It includes schema information to connect software to data and also parameters necessary for correct data processing such as precision, accuracy or calibration information.

[2] K.G. Jeffery, A. Asserson, N. Houssos, B. Jörg: "A 3-layer model for Metadata", in CAMP-4-DATA Workshop, Proc. International Conference on Dublin Core and Metadata Applications, Lisbon September 2013.
[3] http://dublincore.org/
[4] http://ckan.org/
[5] http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

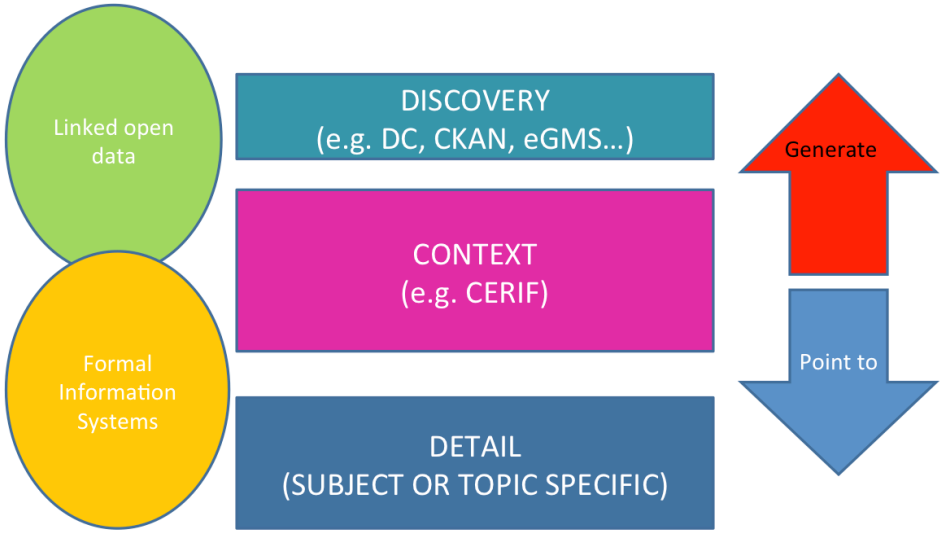

Figure 1 Three-layer model

CERIF, referenced in the Figure 1, provides a "much richer metadata than the standards used commonly with Linked Open Data (LOD) and so improves greatly the experience of the end user (or the advantages of providing metadata.)

Use Cases and Profiles

Metadata use cases were collected using a standard template from the MSDWG members and members of other RDA Interest Groups and Working Groups. The set of use cases is compiled by the MIG. At the RDA Fifth Plenary meeting, the metadata groups agreed on a joint plan. It consists of the following steps:

1. Collect additional use cases: a form has been prepared and is available on the website together with a use case example both written and on the form;
2. Collect metadata 'standards' into the MSDWG directory;
3. Analyze content of (1) and (2) to produce a superset list of all elements required and a subset list of common elements by purpose – so called 'packages' of metadata elements:
4. Test those 'packages' with research domain groups in RDA and adjust based on feedback;
5. Present the 'packages' to the TAB (Technical Advisory Board) of RDA for authorizing as recommendations from RDA to the community.

Metadata Requirements and Obstacles to Data Sharing

Overall metadata requirements were well stated by Peter Fox that:

> "Scientists should be able to access a global, distributed knowledge base of scientific data that appears to be integrated and appears to be locally available."

However data are obtained by multiple means (instruments, models, analysis) using various protocols), in differing vocabularies, using (sometimes unstated) assumptions, with inconsistent (or non-existent) metadata. It may be inconsistent, incomplete, evolving, and distributed. And, it is almost always created in a manner to facilitate its generation *not its* use." (Thus there is no such thing as "raw data." –

From Peter's Fox's presentation). Data are always created for a certain purpose, following workflows and observation procedures, dependent on the sensors and technologies used, come with an intrinsic uncertainty, reflect the theories and viewpoints of the people that recorded the data, and so forth. To give a concrete example, the body position at which a blood pressure measurement is taken matters for the interpretation of the results. Therefore, it is important to agree on a terminology for body positions. SNOMED CT, for instance, provides a body position for blood pressure.

A DataONE baseline assessment of international scientists' data practices and perceptions was carried out in 2010/2011. The survey queried researchers in various domains on their data practices – data accessibility, discovery, re-use, preservation and, particularly, data sharing. In addition, the survey addressed questions related to perceptions of barriers that may hinder data sharing and reuse. Two thirds (67%) of the respondents agreed that lack of access to data generated by other researchers or institutions is a major impediment to progress in science. The addition of metadata to datasets helps makes the dataset more accessible by others and into the future. The results showed that approximately 80% of scientists use no metadata standard and the second largest category for metadata standard used was "MyLab". A follow up assessment carried out in 2013/2014 shows that the current metadata standard use by scientists listed as none has dropped to a little above 50%. Although smaller than the baseline assessment, this still reflects a large number of researchers not using established metadata standards.

Despite our best efforts to document research data with metadata, it is still hard to find answers to questions for several reasons. One of these reasons is that there are many ways to represent the same data and each dataset representation may be different for various reasons including:

- Multiple models for the same kind of data do emerge, each with their own merit
- Massive proliferations of alternate metadata/vocabularies to describe things

This problem was illustrated by Michel Dumontier's presentation of alternative models for protein volume. Direct models of protein volume are simple, physical, and show that a decimal number expresses protein volumes. Models with more context specific ideas of volume express it as one quality of proteins with its own volume units, while still more contextually oriented models express the fact that volumes have data values, which have units of measure for volume (Figure 2) Different database tables may have columns to capture such information.

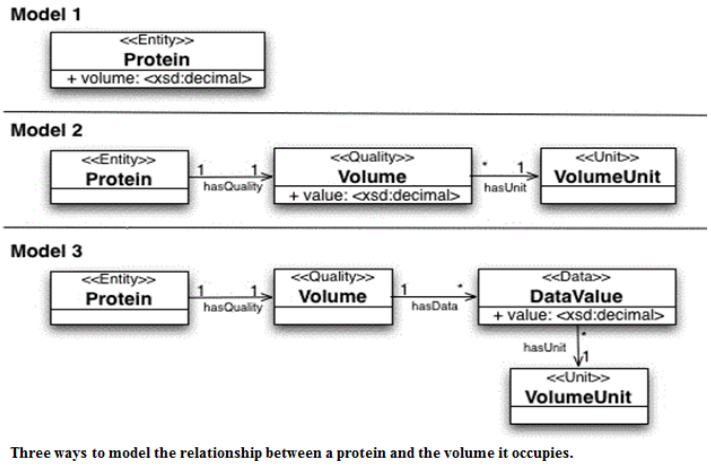

Figure 2 Three Alternative Ways of Modeling Protein Volume

A broad view of why there are major metadata obstacles and challenges include those cited by Ilya Zaslavsky as demonstrated in the EarthCube project's CINERGI[6] pipeline:

- There are different classifications of resource types: Common resource types are: Organization, Webpage, Collection, Dataset (note: EPOS -Users, SW services, computing services)
- A title may be non-descriptive such as
  - Insufficiently unique ("Roads" –what does this mean)
  - Meaningful, but opaque naming patterns (e.g. "AXXX34nn1")
- Keywords
  - May be missing or may be too specific to domain
  - May lack references to a thesaurus or controlled vocabulary, i.e., they are just freeform text acting as a label

Thus obstacles can be as simple as a problem of understanding how to format people's names, a example explored by Pascal Hitzler (see his presentation) showing that more than a text string is needed (**Figure 3**).

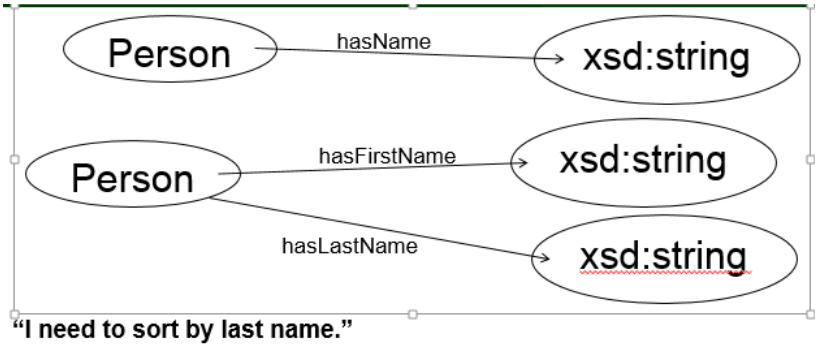

Figure 3 Examples of Different Schemas for Person Name

However, we can have various problems with different type of names, when for example:

- "My name is "Artur d'Avila Garcez". I'm Brazilian of Spanish descendancy.
- "My first name is Anna-Maria, but I live in the U.S. and the ID systems didn't accept a hyphen in my name.
- "My name is Pan Ji. What do you mean by 'last name'?"
- "My name actually changed recently ..."

Similar things can happened with location so there are real "**modeling choices you may regret later**", such as over committing to some particular model of name or location may cause problems.

Other challenges with metadata include:

- Information missing such as an Abstract, or a Contact field saying "call", location, time without reference, wrong URL. Pascal's examples include:
  - Actually need to know the location's degree of certainty
  - Do not need an uncertainty value, a probability distribution is needed
  - What measurement device was used at that location and when?

[6] Community Inventory of EarthCube Resources for Geosciences Interoperability (http://workspace.earthcube.org/cinergi)

    - Working with historical data, how is the prime meridian specified? Before 1884 different people defined it differently. At times temporal description may be missing, or inconsistent, without a reference to temporal scale

- Duplicate records of metadata, which makes automated processing difficult or confusing. Are they the same or 2 instances at different times or locations? This causes problems in grouping: a range of metadata records from a single source that appear to be very similar (only differ in one parameter e.g. location) – they may be discovered as a group of records
    - Records for services and records for datasets/databases behind from data streaming services may be the same thing. Again this is inadequate documentation of relations and is also a provenance problem.
    - Several metadata records from different catalogs may point to the same physical dataset (or have overlapping subsets of distributions). This may also be a provenance Issue.

Efforts to Mature and Better Share Metadata

The RDA Metadata Standards Directory Working Group (MSDWG) has built a directory of descriptive, discipline-specific metadata standards in collaboration with the UK Digital Curation Centre that is easy to update. The directory promotes the discovery, access and use of such standards, thereby improving the state of research data interoperability and reducing duplicative standards development work.

Metadata standards, both general and domain specific are challenging. Ilya Zaslavsky provided examples illustrating this when we have to convert from one form to another. Conversion is difficult because there are different metadata models and profiles involved as previously illustrated for Protein Volume. In research we may run into different details of required/ mandatory and optional fields that vary between Dublin Core and ISO standards. Metadata standards, such as Dublin Core and ISO provide a series of fields to be populated, but can be different meaning of such fields based in part on the initial purpose/emphasis of data collection. As Ted Haberman pointed out in his presentation, we have different dialects or local interpretations of how these fields should be filled out. A simple example is that the roles of “authors” and “contacts” are often mixed up.

Standards vary by and within Domain

Materials science has standards for different levels from quantum to macro scale and from thermodynamic to diffusion. Diffusion Mobility Description for example includes the follow standard elements:

- Elements: e.g. A, B
- Phases: e.g. FCC_A1 and BCC_A2 (crystal structure)
- Software/version used
- Reference info (as previously noted there may be local interpretations of this)
- Bibliographic info
- DOI link
- Contributor info (very under-specified, which provide flexibility but little semantics)
- Name (of what or who one might ask)
- E-mail
- User Comments (also very under-specified which provide flexibility but little semantics)
- Links to other files

Recommended Standards and Differing Metadata Standards as Dialects

Ted Haberman provided some ideas based on his experience of sharing Metadata Recommendations as part of his presentation. He noted that many standards (e.g. Directory Interchange Format or DIF) distinguish among required, recommended, and suggested metadata elements.

Each concept in a recommendation could be represented in any number of dialects. That is, there is heterogeneity in metadata just as there is in data. Eight such dialect variations in recommendations were presented and compared with differing concepts used in each as examples:

NASA Dialects

- Directory Interchange Format (DIF)
- EOSDIS Clearinghouse (ECHO)
- EOS Core System (ECS)
- Service Entry Resource Format (SERF)

Other Dialects

- Data Catalog Vocabulary (DCAT)
- FGDC Content Standard for Geospatial Metadata (CSDGM)
- ISO 19115 and ISO 19115-2 / ISO 19139 and ISO 19139-2 (ISO)
- ISO 19115-1 / ISO 19115-3 (ISO-1)

Pairwise mapping between dialects is a problem. If we reconcile differences between the resources in a pairwise manner, the amount of work, etc. grows quickly: **Cost (N) = N (N-1) / 2 ~ N^2**

Controlled Vocabulary

As previously noted some standards like Dublin Core add a degree of standard vocabulary for the types of information to be filled out as metadata. Some domains such as medicine and earth sciences have developed extensive vocabularies to facilitate communication among researchers and to facilitate data sharing and integration. A rational for a controlled vocabulary (CV) approach, such as with **Community Surface Dynamics Modeling System** (CSDMS), is to introduce a new, generic or standard representation that acts a "hub" between differing vocabularies, which effectively describe the same entity or resource of interest[7]. With a controlled hub vocabulary we may then map resources to and from it. The amount of work, maintenance, etc. drops to **Cost (N) = N** (thus avoiding the combinatorial explosion previously cited).

Within a domain there may be several data models/metadata standards and competing vocabularies. For example continuing with the CSDMS, which deals with the Earth's surface observable and projected changes are that are constantly taking place. For one thing, there are many domains involved including the lithosphere, hydrosphere, cryosphere and atmosphere each with one or more vocabularies. CSDMS allows open source code sharing and re-use through a model repository broken into categories of about 250 different models. Examples of variable names generated as part of CSDMS include some as*sociated with Processes:*

> *snow_melt_volume_flux, atmosphere_water_rainfall_volume_flux or flow rates and fluxes (incoming or outgoing) such as:*

[7] What is defined as a resource may differ substantially across projects.

lake_water~incoming__volume_flow_rate

Such vocabularies seem like a useful first step to standard terms, but see the report section of this paper for workshop efforts to show on how these complex vocabularies can be formally modeled as more useful ontology design patterns.

Adding Structures and Schemas to Metadata

Database structures have long been used to organize metadata. This is natural since metadata is data, but the role of data as metadata does make this a more difficult task since database semantics are limited to relations between tables, as shown in Figure 4, and important relations among data attributes are not explicit as most models.

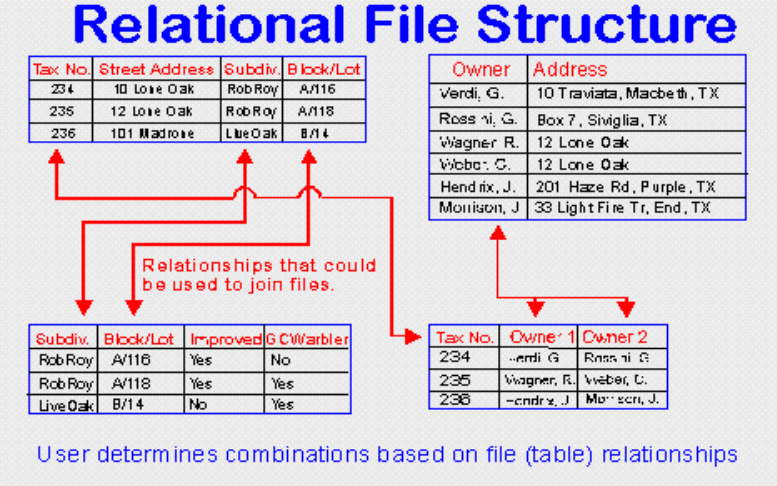

Figure 4 From Database Concepts developed by Kenneth E. Foote and Donald J. Huebner, Department of Geography, U of Texas at Austin, 1996: http://www.colorado.edu/geography/gcraft/notes/datacon/datacon.html

Structures like this can be improved by the use of Dublin Core-type annotations, but there remain issues with the meaning of the vocabulary. Database structures are part of the history of metadata evolving from documenting files in an attempt to make data more human understandable. Even the major data hubs such as Data.gov still rely on keyword-based search rather than concept understanding and as a result can have unreliable, incomplete, and missing metadata. For this type of retrieval problems, even 'a little semantics goes a long way' (http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html). Metadata is now a complex affair with many interacting pieces including Metadata Application Profiles and Workflow like entities. One obstacle to more effective metadata is complexity and as Thomais Stasinopoulou notes this still does not adequately address semantics:

> **"**Metadata schemas are created for resources' identification and description and - most of the times - they do not express rich semantics. Even though the meaning of the metadata information can be processed by humans and its relationship to the described resource can be understood, for machine processing the actual relationships are frequently not obvious. In contrast to metadata schemas, ontologies provide rich constructs to express the meaning of data"

Lessons from Metadata Use

In addition to the use cases provided by RDA metadata groups, several examples of different ways in which metadata is used were presented. Ilya Zaslavsky provided seven types of use:

1. Dataset Identification, Description, Licensing and Provenance
2. Dataset Discovery (via Catalog or repository)
3. Exchange of Dataset Descriptions
4. Dataset Linking
5. Content Summary
6. Monitoring of Dataset Changes
7. Contextualization (*documented in detail sufficient to allow reproducibility)*

Distinct uses of data (discovery, access, understanding) have different metadata needs. The requirements or suggestions for these metadata fields, which may be associated with a specific metadata dialect, are called recommendations in this terminology.

An example presented as part of the Materials Genome briefing concerned search within a CALPHAD type repository (see Figure 5):

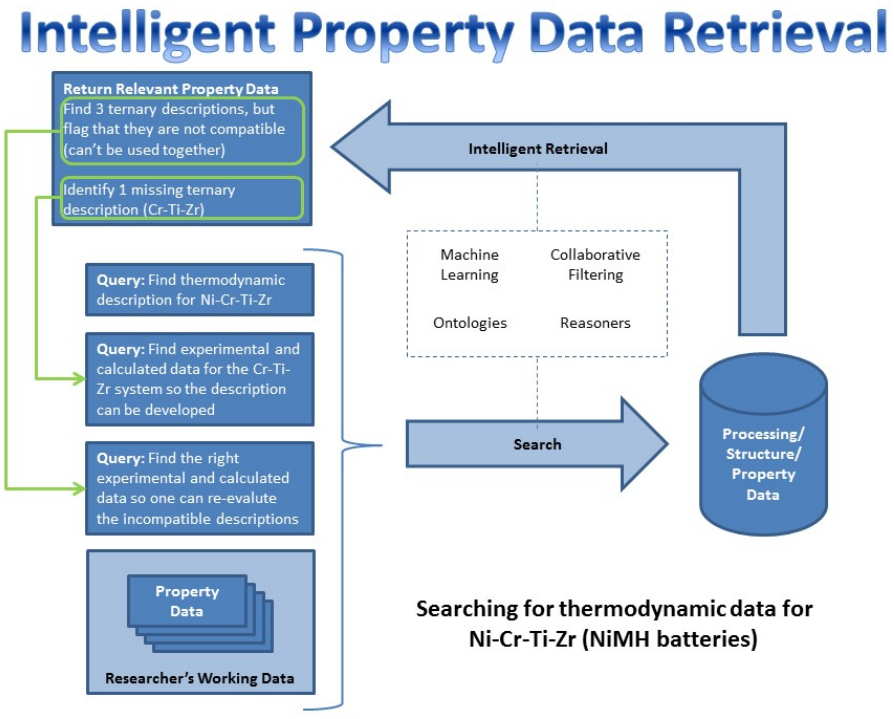

Figure 5: Thermodynamic properties of Ni-Cr-Ti-Zr systems are stored and can be searched to see if they are appropriate for battery needs. Searching for battery data based on properties (CALPHAD-type Repository)

Metadata-provided contextualization is part of the RDA effort, but it is a complex topic. For example, to Peter Fox, context includes provenance information such as:

- Origin or source from which something comes
- Intention for use, who/what generated for,
- Manner of manufacture

- History of subsequent owners,
- Sense of place and time of manufacture, production or discovery

Note, provenance about time and derivation can be asserted in Research Description Framework (RDF)/ontological forms such as done in nanopublications (discussed later in this paper). A small illustration of provenance assertions that can be made in RDF form includes:

```
:provenance {
  :assertion prov:generatedAtTime "2012-02-03T14:38:00Z"^^xsd:dateTime  .
  :assertion prov:wasDerivedFrom :experiment .
  :assertion prov:wasAttributedTo :experimentScientist
```

Metadata Semantics

Over the last 20 years metadata and semantics to support scientific data have developed through a series of evolving paradigms. Metadata schemas are varied, and documented at their best may be sufficient to be processed by humans. They are not, however, typically semantically rich enough on their own for machine processing. As part of this evolution, formal semantics have been added to support data and system interoperability via machine processing[8]. In practice semantics and efforts like standard vocabularies function in multiple tiers between the levels of data and human understanding.

Illustrating Three Semantic Approaches

Different types or degrees of semantics may be appropriate for different tasks. For example, annotation for discovery of information can be done with a smaller degree of semantics than for other uses such as semantics to provide usable context. Adequate contextualization recognizes data heterogeneity as a challenge and as noted by Peter Fox at the workshop:

> "There exist(ed) significant levels of semantic heterogeneity, large-scale data, complex data types, legacy systems, inflexible and unsustainable implementation technology"

How do we handle this varying heterogeneity? Different degrees of semantics are called for by the varying heterogeneity as was illustrated at the workshop starting with light semantics used for annotations.

Semantic Annotation, Tagging and Linked Data

Search across heterogeneous data seems to be one of the simpler tasks to address and can be improved by using better annotation of keywords. One source of improvement is via the semantic web standard RDF, which uses subject-verb-object type triples to annotate data. Such annotated data is made openly available by publishing them as linked data on the web.

Linked Open Data provides an incredibly dynamic, rapidly growing set of interlinked resources that can be leveraged. An example of this work summarized by Michel Dumontier is Bio2RDF (see Figure 6), which is a large, open source project (scale of over 11 billion RDF triples from 35 datasets) that provides Linked Data for the BioLife Sciences based on large datasets such as PubMed and ChEMBL. There are

[8] Semantic Interoperability is considered to be one of the problems of this decade. Lack of interoperability leverages a cost on productivity, lives and annual dollars. [Ref. OMG, & SIMF citations]

several advantages of converting informal biomodels into formal representations of bio-knowledge. Each dataset, which has an informal model, is described using its own schema but triples add formal semantics.

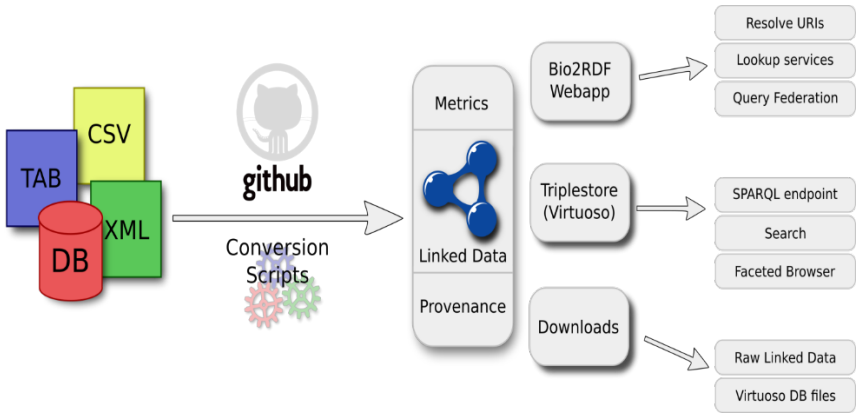

Figure 6 Bio2RDF, an open source project to unify the representation and interlinking of biological data

Semantic annotation was also illustrated for the BioModels Database, an open-source repository for storing, exchanging and retrieving quantitative models of biological interest. Curated models can be submitted in a variety of forms including **Systems Biology Markup Language (**SBML) and these can be further annotated using RDF, as shown in Figure 7, with the *intent* to express the fact that the species represents a substance composed of glucose molecules. In effect we describe complex associations between research statements/facts to build a larger model useful for scientific discourse and publication.

We also know from the SBML model that this substance is located in the cytosol and with a (initial) concentration of 0.09765M

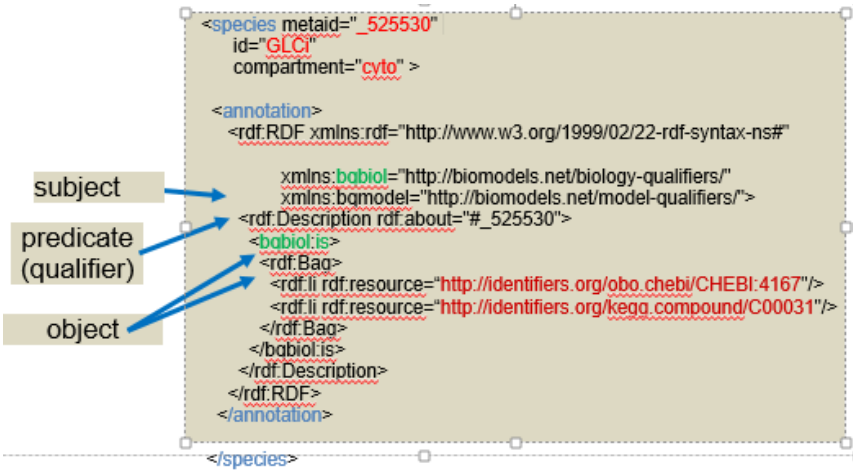

Figure 7 Annotation of SBML using RDF

Converting informal biomodels into formal representations of bio-knowledge provides these advantages:

1. Captures the semantics of models and the biological systems they represent

2. Leverages knowledge explicit in linked terminologies
3. Validates the accuracy of the annotations/models
4. Discovers biological implications inherent in the models
5. Queries of the results of simulations in the context of the biological knowledge

Tactically you can turn linked data into whatever you need. You can be APPLICATION specific such as drug repurposing (see Repurposing Drugs with Semantics - ReDrugS[9]), verify annotations in biomodels, or discover aberrant pathways.

RDF can be used to express standards like W3C, community or application standards or even visualization standards. Linked data in RDF form can be used to query the data, ontologies, and services simultaneously using federated queries over independent (unchanged) SPARQL Databases.

For example, you can use a SPARQL query to find all protein catabolic processes (and more specific ones) in biomodels with a query <example, http://bioportal.bio2rdf.org/sparql>. A query is usually carried out on one platform such as a biomodel.

Supporting/ bridging ontologies are used to unify the representation and interlinking of biological data. This illustrates how heterogeneity is allowed in data but with proper annotation and bridging ontologies, one may also support data integration of these. The general idea of bridging ontologies is discussed and illustrated in the sub-section on ontology design patterns, but use of a specific ontology (SIO) for BioMedicine was also discussed at the workshop and some work with this was demonstrated for the Earth Science domain.

The Semantic Science Integrated Ontology (SIO), with its pattern shown in Figure 8, is used to ground Bio2RDF, SADI[10] and semantic web services. SIO has 1300+ classes, 201 object properties (including. inverses) and a single data type property. It is built to support transferring data and models using the ontological schema

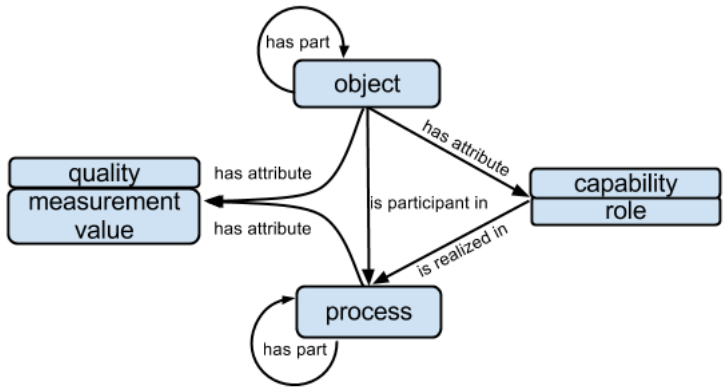

Figure 8: High-Level Concepts of the Semantic Science Integrated Ontology (SIO)

[9] ReDrugs uses nanopublications, which are structured digital objects that associate a statement composed of one or more triples with its evidence/provenance, and digital object metadata. They are dynamically generated from a SPARQL query and thus are convenient value-added products.

[10] SADI stands for Semantic Automated Discovery and Integration and is a type of semantic web service API that provides search, graph traverse, and composite probabilities for the resulting graph of biological entities using a SADI web service framework.

CSDMS Standard Names

Another illustration of annotation, with less explicit semantics, is via standard and controlled names. This was discussed at the workshop as part of the CSMDS, effort to deal with vocabulary issues with minimum effort. CSMDS maintains an extensive list of standard names of objects, quantities and operators, using an intelligent multi-level naming scheme. CSMDS has various standard names for quantities, objects and operations. There are also five delimiters used in CSDMS standard names.

It was noted that on the surface CSDMS Standard Names have a similar pattern to RDF triples for creating unambiguous and easily understood standard *variable names or "preferred labels"* that are concatenated according to a set of rules. These are then used to retrieve values and metadata. The triple pattern is:

> Object name + [Operation name] + Quantity name

Examples of this pattern include terms for variables such as:

> atmosphere_carbon-dioxide__partial_pressure
> atmosphere_water__precipitation_leq-volume_flow_rate
> earth_ellipsoid__equatorial_radius
> soil__saturated_hydraulic_conductivity

The relations in these triples (+ plus or -) work at the word level with the semantics informally understood in the words. That is, "operation name" is a label for some type of operation; carbon dioxide after atmosphere implies a part-whole relation and no formal relations can be processed by computers from specified CSDMS terms. For this reason the work group decided to try to employ the SIO model (as seen below) with RDF triples to further formalize CSDMS terms. See the summary report from the Earth Science group.

The vocabularies of variable names are used for connecting models in a plug and play manner. If one can find things that should have the same name then they can be connected. Mapping to these names is part of the Basic Model Interface that enables the plug-and-play functionality of the models. CSMDS has 3 major semantic uses in search and discovery:

1. Semantic similarity of terms for search and discovery; this casts a big net and we can use thesauri and synonym ideas
2. Semantic equivalence for which we need unambiguous, human and machine readable labels as part of a more focused view, for proper mediation[11] and
3. Matching and semantic meaning showing how term concepts are related to other terms and their concepts as needed for machine processing. Classification is used here, but ontological analysis and design are more generally applicable.

Ontology Design Patterns

Ontologies are theoretical or computational artifacts/tools constructed to express some intended meaning of a vocabulary that is understandable to humans. They are like metadata but use terms for primitive categories and relations more formally to describe the nature and structure of a domain of discourse. Ontologies are particularly useful under several circumstances, such as situations where subtle distinctions are important or where precision and accuracy of the meaning role are played by

[11] *We need to know what terms mean the same thing for* connecting resources (e.g. model, data); automation and to pass data from a provider to a user. Different labels for the same things are synonyms.

metadata for data and system interoperability. For example, forms of precipitation such as rain or snow may be important or the form of a compound may influence the nature of reactions. Differences of data meaning/ heterogeneity are also due to cultural variations, progress in science, viewpoints, and granularity. As noted in Krzysztof Janowicz's talk, "Sensemaking is difficult and meaning fits into our contextual purposes.

> Because different data fit different purposes these will be heterogeneous, and it is important to recognize these and not paper over them by slamming them into some arbitrary commitment to one word or phrase. Disagreement is to be expected during ontology creation as commitments are made. Good ontological models reflect explanation and justification for formal commitment. "

An ontology design pattern (ODP) is in effect a small, reusable ontology that can be a successful solution to a recurrent modeling problem. ODPs provide a modular approach to ontologies that helps make them reusable and flexible since a good pattern includes replaceable pieces in a plug-and-play style. The intended vision of using ODPs to improve metadata is to create a metadata ecosystem in which patterns can be assembled and shared as needed.

There are several types of ontology patterns but the one most applicable here is the so-called *content pattern* that usually encodes specific abstract notions, such as process, event, agent, etc. SIO, mentioned previously, provides one example of a simple pattern with 4 top-level related entities.

One pattern illustrated at the workshop was for **Oceanographic Cruise.** This is shown in Figure 9. More examples from the GeoLink Project are available at www.geolink.org.

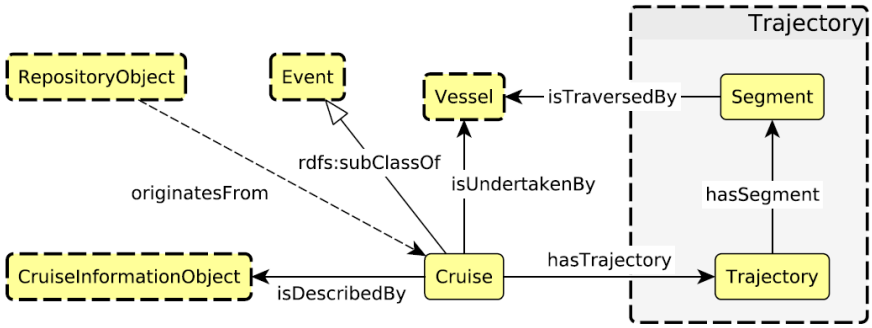

Figure 9 Oceanographic Cruise ODP based on earlier Trajectory Pattern. Includes plug-ins for Vessel, Event etc. as needed

Different patterns can be specialized for new use and also combined with other ontology design modules such as illustrated in Figure 10 for cruises.

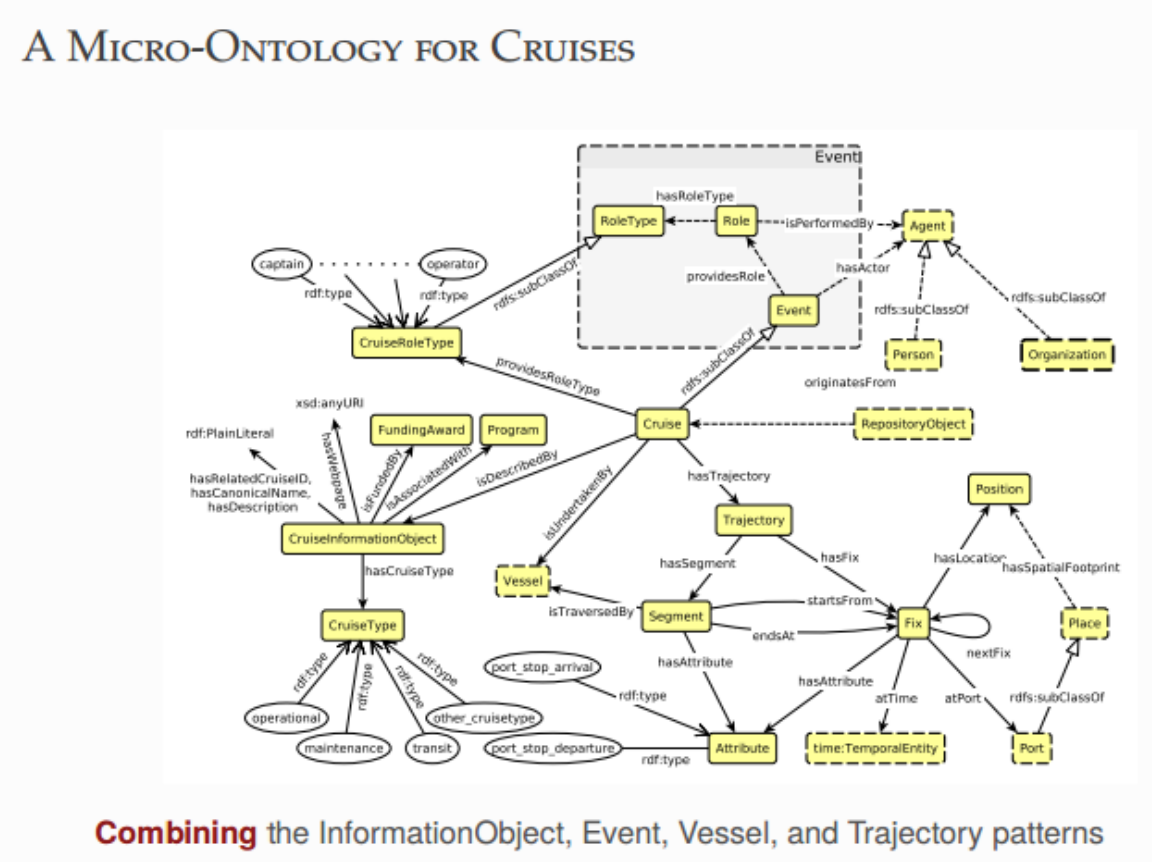

Figure 10: A family of Cruses forming a Micro-Ontology combining InformationObject, Event, Vessel, and Trajectory patterns

RPI Tetherless World Constellation Methods

Metadata and semantics to support scientific data have matured through a series of evolving paradigms over the last ~20 years. The Tetherless World Constellation methods for adding semantics, for example, have evolved over this time. The initial attempt (circa 2005) leverages the semantic web and its standards such as RDF triples. Ontologies were used for developing useful metadata with semantics.

After 2009 some enhancements included substantial knowledge provenance ontology work (PROV-O) and a Semantic Data Framework (SeSF), which can be used as a configurable and extensible frame for semantic eScience work.

> "SESF builds upon previous work in the Virtual Solar-Terrestrial Observatory. The VSTO utilizes leading edge knowledge representation, query and reasoning techniques to support knowledge-enhanced search, data access, integration, and manipulation. It encodes term meanings and their inter-relationships in ontologies and uses these ontologies and associated inference engines to semantically enable the data services. The Semantically-Enabled Science Data Integration (SESDI) project implemented data integration capabilities among three sub-disciplines; solar radiation, volcanic outgassing and atmospheric structure using extensions to existing modular ontologies and used the VSTO data framework, while adding smart faceted search and semantic data registration tools. The Semantic Provenance Capture in Data Ingest Systems (SPCDIS) has added explanation provenance capabilities to an observational data

ingest pipeline for images of the Sun providing a set of tools to answer diverse end user questions such as *Why does this image look bad?*

..... The proposed eScience framework is based on semantics, and software built on and around the semantics. A sustainability path for communities is essential so that use may continue into the future. Sustainability extends beyond software to ontology development and vetting, and communities of practice (scientist, data providers, and technical teams)."[12]

Tools for Better Data Curation and Semantic Annotations

The workshop did not include a specific session on metadata or semantic tools, but tools were part of several presentations and part of the discussions. For example, obstacles to wider data sharing included inadequate metadata tools along with data discoverability challenges, lack of best practices and educational materials, insufficient funds and low sustainability of key infrastructure.

Materials Genome Initiative tools, for example, will focus on a phase-based data search for specific data. Their tool suite will include Guided Data Capture (http://trc.nist.gov/) and the use of a curation tool, as shown in the Figure Figure 11, which leverages the a DSPACE repository and creates a Materials Data Curation System[13] using XML-based schemas with a focus on phase-based data search for specific data; and not just a paper.

There is also a focus on developing curation tool interfaces to integrate directly with instruments and computational tools. See https://github.com/usnistgov/MDCS

While the benefits of semantics and interoperability are understood, practices and tools for practical and widespread implementation are not obvious. Some progress is being made at the lower end of semantics for improved annotation.

Tools will be needed to facilitate template construction and semi-automated metadata annotation. Examples of text and ontology annotation tools for example, include: KIM, GATE, MnM, OntoMat and Melita (Benchmarking of annotation tools).

[12] http://tw.rpi.edu/web/project/SESF

[13] The curation system is written in Python, backed by MongoDB with a SPARQL Query interface, an XML-based Schema and supports table input. Its new features include an ability to store templates. Some schema management tools, a REST API interface, a Schema Composer and a Link large data w/ DSpace.

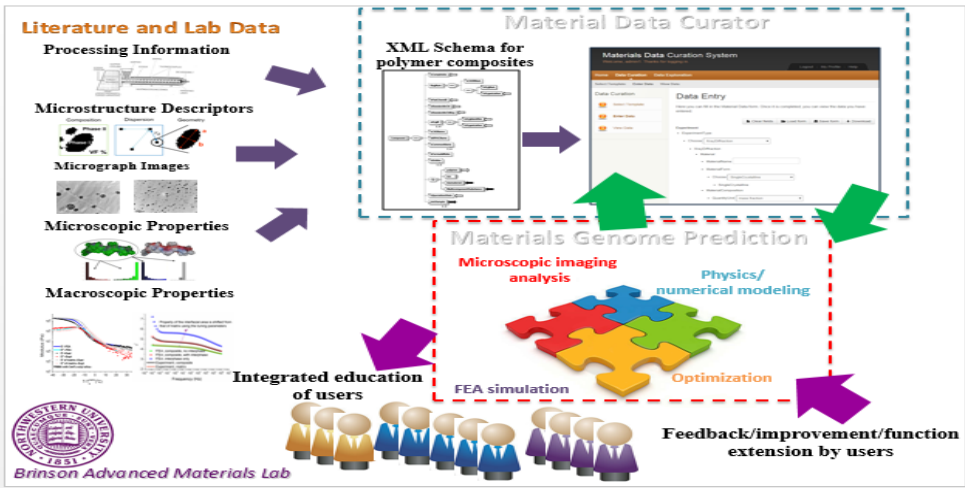

Figure 11 Curating Materials via a Data Curation System

In addition better tools for visualization of ontologies and ODPs are needed. Various efforts that are underway (at RPI and Stanford for example) are part of the “Center for Expanded Data Annotation" that will develop tools to facilitate template construction and semi-automated metadata annotation.

Common Data/Metadata Architectural Themes

Like tools, architecture did not have a specific session devoted to it, but architecture was part of several presentations and implied in the discussions. One example depicted in Figure 12 from Peter Fox’s presentation illustrates semantic interoperability within an architectural view. Of note is the integration afforded by a semantic mediation process connecting vocabularies and models by the Software Layer.

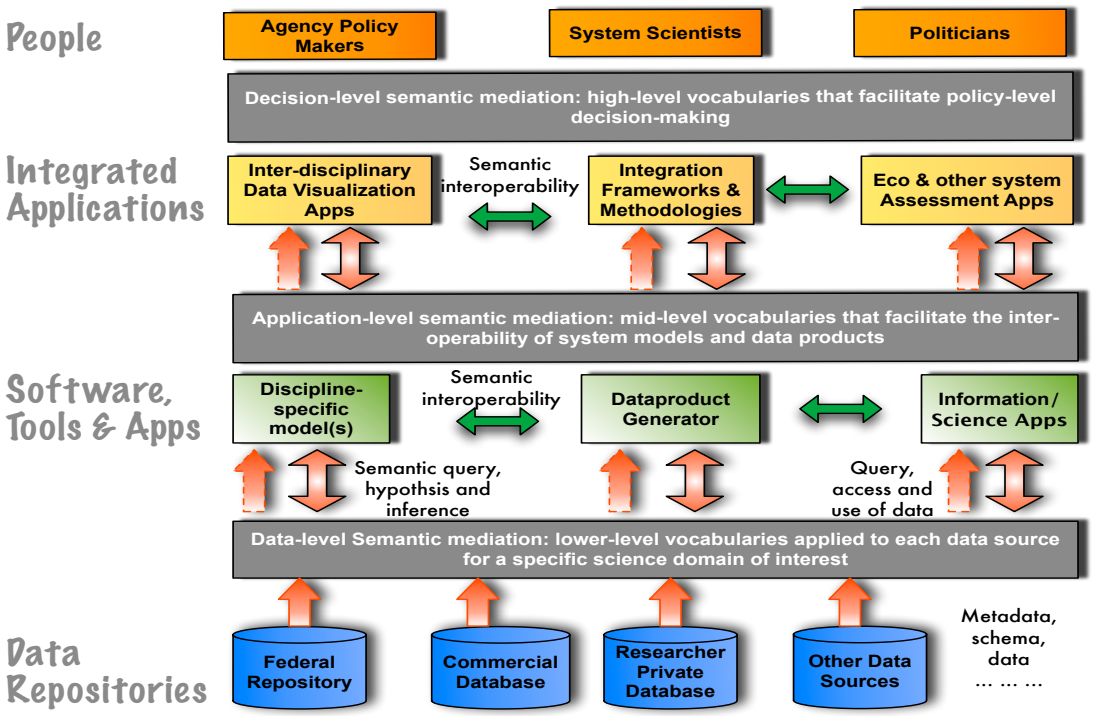

Figure 12: Semantic interoperability visualized within an architectural view

EarthCube "Architecture" has requirements that other data efforts tend to share such as information integration, federated data discovery, and the flexibility of modular aspects. In an architectural context it was noted that software and data systems should or must be:

- modular
- extensible
- sustainable
- sliceable (i.e. you can adopt part of it without adopting all)
- simple enough for easy adoption
- complex enough to solve real problems
- scalable in terms of breadth of topic coverage
- elastic, in that it allows partners to decide how much they want to share
- respectful of individual partner/project/domain modeling choices

These are reasonable system requirements. In the context of metadata and ontologies to support architectures, several of these remain challenging. For example making the adoption of semantics easy, as well as complex enough to solve real problems and doing this in a scalable way is difficult.

Panel Discussion

A panel session was held during the first afternoon of the workshop. Panelists included: Bob Hanisch, Ray Plante, Emilio Mayorga and John Westbrook. Among the topics discussed were:

- Minimal vs. Maximal Metadata, Tools and Motivations
- Standard Reference Data and How Linked Data and Ontologies Can Help
- National Data Service

- Connecting Linked Data Work and Traditional Metadata
- No Grand, Common, Metadata Schema

Minimal vs. Maximal Metadata, Tools and Motivations

Bob argued that minimum metadata is a good thing, but maximum metadata is a problem. Trying to get everything expressed for all circumstances is just too complicated and the result too rich. Bob provided an example from a NIST rich metadata effort that wasn't taken up. Users leave blank fields if there are too many things to fill out thus requiring (later) curation, which is expensive.

Ted wasn't sure that more is less and the problem may be lack of suitable tools to: create more metadata, find what you need and to view it in a user-friendly way. Ted also noted that adding metadata is similar to spiral development idea[14]. You break the effort into parts (modules) that are integrated over time. Ongoing evaluation helps and some European projects are working on user feedback about datasets that amounts to a spiraling, curation opportunity.

Michel suggested that there are examples of tools that make it easier for metadata authors to fill in the blanks. Tools that help with context make the process easier. An example is someone using a microarray that specifies mice as the target. In this context the tool can present the author with a list of suitable microchips to document. Scott Peckham is prototyping a tool, TurboSoft, analogous to TurboTax that helps users at pain points such as the metadata needed to bridge between different models. User-friendly tools allow data producers to be better drivers of improved metadata.

The question was asked, "What about the producers of spreadsheets?" We are not giving them the tools to provide more information about their data. An issue mentioned here, "Is there a role of linked data?" Cooperation is afforded in linked data - I do my part and you do your part and then we connect. So better tools here would be good to have and might provide a good-sized payoff.

Emilio cautioned that we need motivated users to be accurate with a TurboTax-like tool. There is an issue of validation. Just because a form is filled out doesn't mean it's accurate. We need carrots as well as sticks. Bob asked how to motivate the researchers and Michel suggested the role of a funder's mandate; but not all funders do this, so we need to think of other incentives. Data citation is one way, but there is a lot of work to make this happen.

Ted suggested that one motivation is that others will discover your data. Many scientists are happy with existing ways of sharing (meetings, etc.) so we need a richer set of benefits other than that others will be able to use your data. NOAA provides an example. If you create some metadata, you then get some service (this can be done by data centers or NDS). Ray noted that this was the way the Virtual Observatory (VO) worked. The 100 papers that cite the VO show the real research impact from VO. However, most papers do not usually acknowledge the infrastructure that supported it. The research community needs to make good use of infrastructure in order to do good science. Ray suggested that this might mean that we need to build our own metrics for how data is reused and cited.

Charles asked people to think of how blogs work on the web. There is a mechanism for reuse and re-annotation of data. It is not just the end users but also the middleman that are involved in aggregating information.

[14] Metadata development can follow a spiral idea similar to the spiral model of software development. A spiral includes a collection of concepts required to support a particular documentation need or use case as discussed above for differing dialects

Ray cited Github as the type of environment that allows developers to know when others are having problems and allows feedback for this.

For a discussion of Minimum Information Biological and Biomedical Investigations (MIBBI) see Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project, which discusses how "minimum information guidelines for reporting proteomics experiments and describing systems biology models are gaining broader support in their respective database communities."

Standard Reference Data and How Linked Data and Ontologies Can Help

A related issue is the use of Standard Reference Data. Some are too specialized to be used by others and those that are not general may not support interoperability. The same may be true of some ontology work. A perceived problem is that there is no immediate payback for the work of developing ontologies or standardizing reference data. Unless there is a compelling science use case, pursuing general interoperability may seem to many to not be worth the effort.

Dave Dubin provided some of his experience justifying and explaining ontologies and linked data. Students want to know how this will help. To them it looks for very boring. One needs some examples (even mundane ones) to show how linked data structure (versus structured data) will help. But the real examples presented today by Michel and others show how amazing it could be if it were scaled up. How great it would be if we all got on board. We could show that the Java programs are simpler and don't include all the hard to create checks. We get to build in sanity checks. Think of it as a content-based websites where we don't have to explicitly store things because we can deduce things just in time. This allows social engineering and cooperation in small groups.

National Data Service

Ray Plante, representing the National Data Service (NDS), provided a summary on their ideas and efforts. NDS is a consortium of many institutions dedicated to "an emerging vision of how scientists and researchers across all disciplines can find, reuse, and publish data". The core ideas are to build and operate services for doing common things that make sharing data, reusing data easier. Metadata is an important component of this.

The NDS approach is not to be the silo to end all silos. Rather it is to connect what is going on or successful so that there are common ways to accomplish various things (transporting data, etc.).

Specific NDS challenges include:

- Data publishing workflow - how to instill the use of workflow by researchers? The goal is to try to get researchers to input metadata as early in the process as possible (example, vendors of instruments providing an API).
- There are various tools to consider but also the repository where data will reside. Data and metadata, it is argued, need to travel together.
- Formats like JASON LD format - with the ability to state how one metadata maps with other metadata. Thus you can preserve metadata schemas within a community but also map to other schema. Charles Vardeman noted that he is doing this at Notre Dame and that you have something to hook into from what the application developer had in mind.

Connecting Linked Data Work and Traditional Metadata

Gary suggested that at times we seem to be talking about 2 different realms: linked data and traditional metadata. Emilio, for example, talks about the realities of producers and consumer within a large and diverse community that includes individual PIs in NOAA and NSF-based projects using OGC and the Unidata standards. In contrast to this world based on traditional standards, we have the realm of linked data with its standards that Michel, Pascal, and Krzysztof are talking about.

So a question is, "Can we bridge this repository world of data and linked data that is open and in the cloud?"

The issue for Ray, Scott and others is that they want to know how to leverage each realm's strengths and investments. Can we convert data/metadata to linked open data and test that this is easy and useful? The idea is to pull data content in and reformat it to RDF. The data is then massaged to the requirements of particular uses. Charles noted some relevant experience with the Semantic Sensor Network, which has an ontology, but was supported by the Open Geospatial Consortium. We need some matchmakers to bring the realms together.

Michel noted that one problem with this conversion is that data (hence also the metadata about it) is dirty. He has found that there are more exceptions than not, so while you can have scripts to transform raw data into RDF form, the raw data are imperfect and changing and thus you have to keep up.

No Grand, Common, Metadata Schema

People agreed with Ray as part of a discussion of drilling down from general to specific ideas for data discovery that you "can't expect to have a grand, common, metadata schema." The solution needs to be a hybrid of what is available, such as from Google, Kayak etc. We have some ingredients such as SchemaGraph and Wikipedia/DBPedia but there at least 2 challenges:

- How to pool together the different ways of finding data to present a common view that is easy to navigate? Ray noted that this is needed in the astronomy community.
- Too many results problem - we know that author and title are not sufficient and we have to leverage the meaning at the community level. This brings up metadata mapping, registries, and linked data issues.

The PDB Story

John Westbrook provided a history of PDB over the past 40 years as a small community that used a repository. The data architecture that has evolved grew out of a crystallography international body that provides for standards. The community embraced metadata at an early point (1970).

The original goal was electronic publication but in the early 1980s this was extended to macromolecular (from micromolecular). Generating the metadata for macromolecular took 10 years with a range of expertise. Representation of data and representation of science were very similar to what was discussed today. The project started before XML and well before the semantic web technology and leveraged-keyword value (simple and reproducible for data instances, domain metadata).

The database was extended beyond crystallography to experimental model data; most of the experiments in the database come from x-ray crystallography. Integration has been modeled in a data dictionary with mapping of the various types of experiments, and semantics as part of well-formed definitions.

Gary offered examples of the two types. You may have keyword pairs that are documented in Wikipedia; or you have the definition in PDB done much differently. They were developed by physical scientists working out the definitions. This takes a long time but you may get community based-agreement from which you can expand to different communities. In the process you may encounter the real heterogeneity of a different world. But this is reality.

Charles said that PDB was the starting point for some analyses but data taken out and other applications were used (e.g. AMBER).

John added that in addition to being able to define metadata, it is also important to define reliability at various granularities. In the last 5 years they have worked with the community to describe best practices for applications to work with the PDB archive. This is important for discussion of interoperability. It is important to add quality information.

Workshop Session Reports

Following the presentations, a major activity was small group work on topics of interest. The results of these group efforts are summarized below.

General Metadata Report

The group developed a project description that was submitted to the RDA Data Share Fellow program. This project will be affiliated with RDA Metadata Interest Group (MIG). This proof-of-concept project will evaluate metadata integration/ Interoperability for data discovery via Linked Open Data (LOD) publication using LOD protocols.

Still another effort discussed was linking portions of this project with the Linking across the Chasm project.

Materials Genome Initiative Breakout Report

The Materials Science Registry will be proposed to the RDA as a Materials Science Working Group. The WG proposal presented and discussed at RDA Plenary 5 to the Materials Science Interest Group. The work will define minimum or modest metadata extensions to Dublin Core to enable resource discovery across materials sciences. The group will seek organizations to do pilot implementations. The deliverable of this effort will be a document with metadata definitions. OAI-PMH will be used for harvesting and synchronizing metadata records.

The Materials Genome Initiative group also explored how the HCLS Dataset Description (DD) Profile[15] might be applied to datasets from the materials and biochemistry communities using the dataset records from the NIST Materials Measurement repository[16] and the Protein Data Bank (PDB)[17]. In both cases, conversion of the dataset descriptions to the DD profile using the JSON-LD format was quite straightforward. The HCLS Dataset Description (DD) Profile is interesting for its potential to enable cross-disciplinary dataset discovery because it is explicitly discipline-generic. Also interesting is that, as a "profile", it does not define any new metadata concepts; instead, it leverages metadata definitions from several existing metadata standards.

[15] http://tinyurl.com/mzstlxs
[16] http://materialsdata.nist.gov/
[17] http://www.rcsb.org/

At NIST, the Materials Measurement repository archivists must deliver their dataset descriptions to Data.gov[18] using their JSON-LD-based Project Open Data (POD) Metadata format[19]. We studied two dataset descriptions from the repository, one in POD format and one in the repository's native format. In both case's we found that most of the data mapped readily into DD profile. There were some pieces of data that could not be mapped; however, the POD format provided the necessary term URIs that allowed us to include that repository-specific information in the DD record. We concluded, then, that it would be quite trivial to create a script to convert repository dataset metadata records on demand or en masse.

With the PDB case, the example record describes the entire PDB data collection as a whole, which contains many separately accessible components. These components can also be accessed in multiple encodings that DD can capture. In effect, then, the PDB DD record represents a data "site-map" of the PDB data holdings.

Once it was determined how straightforward creating DD records was, the group discussed three potential next steps, each with increasing complexity, which could demonstrate the value of exporting records in this format. First might be a simple demonstration of dataset discovery. This would involve converting a large number of records into DD format, loading them into an RDF triple-store database, and creating SPARQL queries for selecting datasets. A simple demonstration like this is significant in how it points to a high-level mechanism for dataset discovery that can be applied not only to any discipline but also across disciplines.

A next-level exercise would be to try to show the value of the Linked Data Concept by creating RDF connections to other linkable data. Perhaps easiest would be to link datasets in the Materials Measurement repository related to particular chemical species to descriptions of those species in Wikipedia (which are available via DBpedia and are actively curated by the chemistry research community). Connections could also be made where possible to data in the Protein Data Bank, which already provides RDF descriptions. The linked data approach could also be used to demonstrate how to create links from NIST repositories to other community repositories such as the Materials Data Facility, the Materials Commons, and the Materials Project. This would also be an opportunity for exercising the recent W3C standard, Linked Data Platform, which defines RESTful service interfaces for accessing RDF descriptions.

Finally, it was determined DD records could be used to explore how connections might be integrated not only to discipline-specific metadata but also to discipline-specific tools. The linked-data approach may be a viable mechanism for smoothly migrating a user from discipline-generic tools capable of finding datasets across multiple disciplines to discipline-specific tools that can leverage the knowledge of a discipline and its specialized metadata to help the user refine their selections and analyze specific datasets.

Earth Science Work Group Report

The goal of this group was to develop an example of an Ontology Design Pattern for one portion of the CSDMS vocabulary dealing with radiation effects in the atmosphere. A sample of terms used to inform the effort is shown below:

1. atmosphere_clouds_radiation~incoming~shortwave__absorptance
2. atmosphere_clouds_radiation~incoming~shortwave__reflectance
3. atmosphere_clouds_radiation~incoming~shortwave__reflected_energy_flux

[18] http://www.data.gov/
[19] https://project-open-data.cio.gov/

4. atmosphere_clouds_radiation~incoming~shortwave__transmittance
5. atmosphere_clouds_radiation~incoming~shortwave__transmitted_energy_flux
6. atmosphere_clouds_radiation~outgoing~longwave__emittance
7. atmosphere_clouds_radiation~outgoing~longwave~downward__energy_flux
8. atmosphere_clouds_radiation~outgoing~longwave~upward__

The naming scheme suggests some underlying formal semantics that might be exposed as a modular ontology. As part of ontological-conceptual analysis we might ask if absorptance and reflectance are related and if water and aerosols in the atmosphere have similar phenomena.

The group used visual models as a guide to address some of these questions, such as in Figure 13 where relations are shown to a total radiation input. An implicit, underlying model is visualized in the figure including various activities such as absorption and reflection or upward vs. downward paths of radiation. This"knowledge" is not actually expressed in the terms since incoming may be downward or upward depending on whether it is reflected or not.

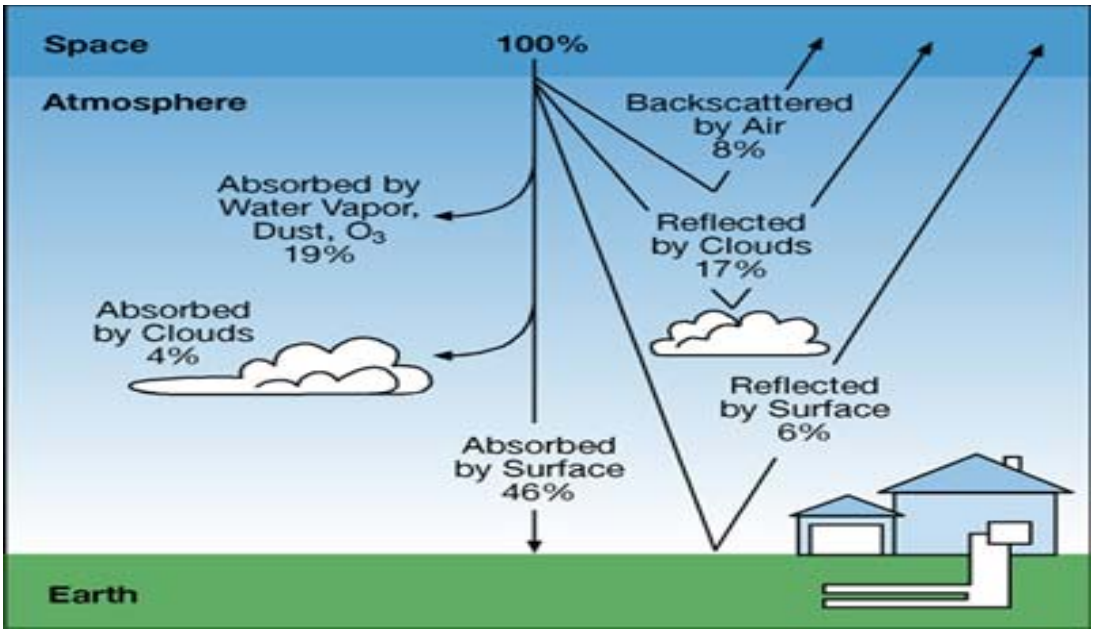

Figure 13 Big Picture of Atmosphere Radiation Interaction

Based on this analysis the group looked at implied ontological commitments needed to be formalized and employed the SIO (previously discussed) as a starting point for the work. SIO has 4 main, related entities Object and its sub-objects (in Blue in Figure 14 below) some Process (in Purple) and a role (in Yellow) for the Objects and Processes along with the Measurements (Green) of Attributes as diagramed in Figure 14.

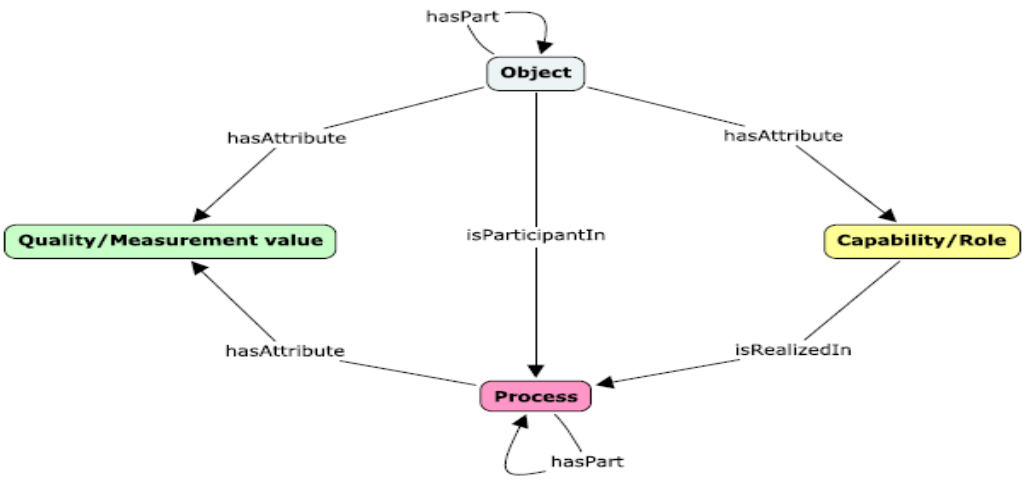

Figure 14 SIO Model used to Model a Portion of the CSDMS vocabulary

Based on group analysis SIO was deemed to be useful to show how various objects can play a role in a process. In the case of the previous list of terms this is the process or radiation interacting with different objects such as clouds or aerosols. The objects can play difference roles such as absorbing or reflecting. Based on this understanding a model was developed, shown in Figure 15 below, that aligns the Process of Atmosphere-Radiation Interaction and related objects with the SIO ontology model. As noted in the figure the model is general - many other processes might be represented with different objects and attributes.

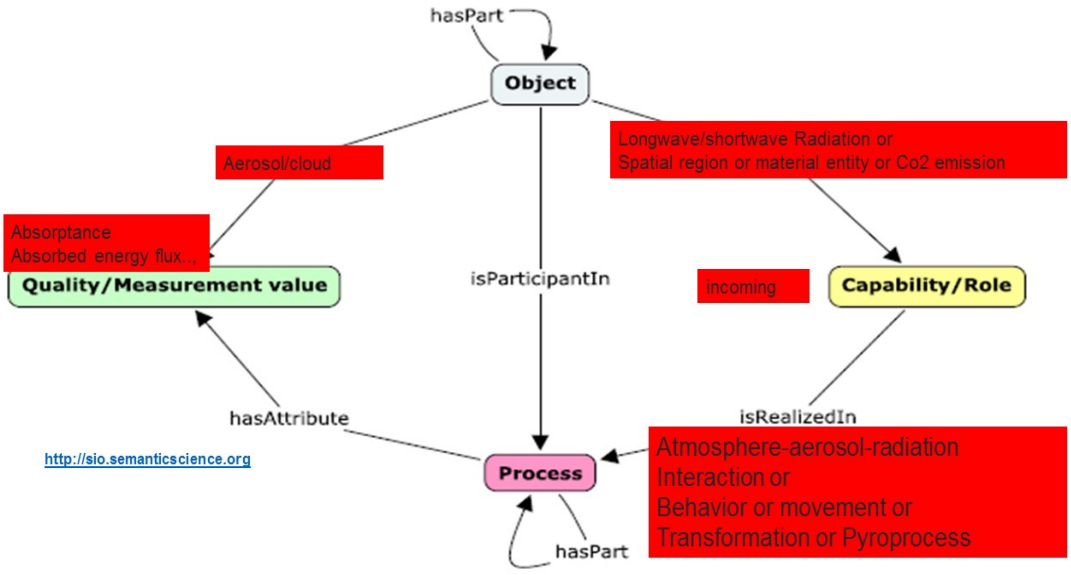

Figure 15 Alignment of some terms with the SIO Model

Using the alignment, a full model for a portion of the Radiation-Related Terms in the CSDMS vocabulary was developed into a preliminary model as shown in Figure 16. Roles such as incoming radiation

participate in a radiation interaction process with the aerosol part of the atmosphere. An object of that participation role is reflected in shortwave radiation that has a series of attributes including energy flux.

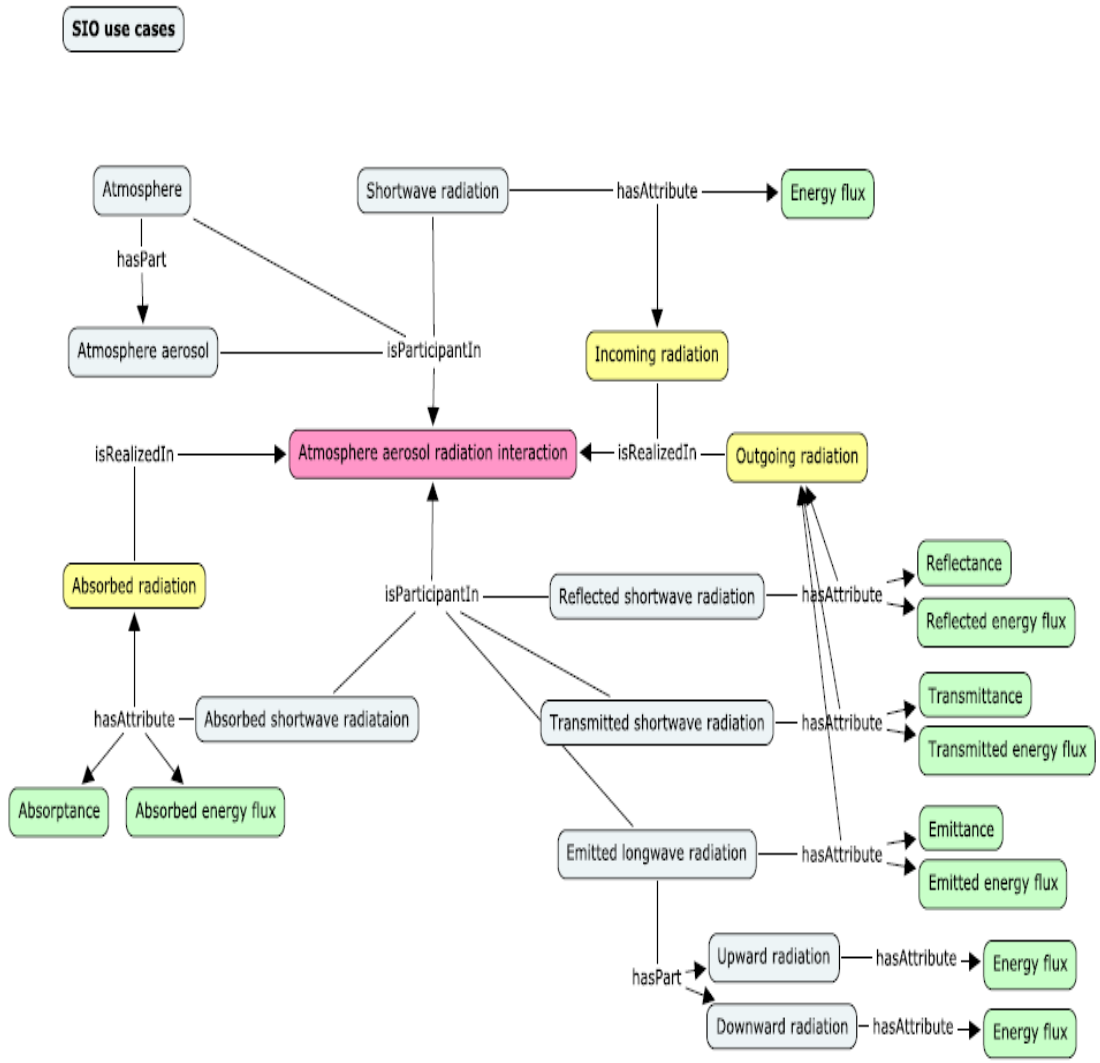

Figure 16: A Portion of the CSDMS vocabulary of terms mapped to the SIO Model

The integrated view of terms in the models allows more relations questions to be posed directly to the data such as attribute values of all objects that participate in reflected shortwave radiations or the emitted long wave radiation that has an upward radiation part. Potentially all of the CSDMS vocabulary may be analyzed and modeled.

Conclusions and Summary

All participants were excited by an enhanced and integrated metadata approach including the possibility of enhanced semantics.

Outreach

As an outreach effort there were a number of successes. These included involvement of the Biomedical community, leveraging some of their experience with semantics. There was also successful outreach to the Materials Genome Initiative where new work was proposed and indeed is underway. This included the launching of efforts to show the value of the Linked Data Concept by creating RDF connections to other linkable datasets such as in the Materials Measurement repository. In addition outreach with NDS will continue.

A number of the groups are considering using Semantic Web Technologies, Linked Data, and Ontologies for a variety of things including federated queries over multiple data sources based on the workshop. Potential benefits include:

- Unique global identifiers easy conflation and de-duplication
- Transparent data model;
- Reduced need for guessing
- No data silos,
- No API restrictions
- Many pre-defined lightweight vocabularies (ontologies)
- Smart data reduces the need for smart applications
- Access to machine reasoning support
- No need for agreement, since ontologies make hidden assumptions explicit
- Does away with the data – metadata distinction!

Two proposals were developed for submission to the RDA Data Share Fellow program, both affiliated with the Metadata Interest Group. The General Breakout group developed a project description for a proof-of-concept project that will evaluate metadata integration/ Interoperability for data discovery via Linked Open Data (LOD) publication using LOD protocols. As mentioned in the Materials Genome Initiative breakout report, a materials project proposal was developed for the RDA Data Share Fellow program.

As part of the outreach effort additional metadata standards were added to the Metadata Standards Directory. In addition, new use cases were also submitted for the RDA Metadata effort.

A process for utilizing existing ontologies such as SIO was developed. A new ontology design pattern was developed for the CSDMS vocabularies. In addition a presentation on how to use SIO for materials vocabulary was made at the RDA Plenary 5 Joint session: IG RDA/CODATA Materials Data, Infrastructure & Interoperability, WG Data Type Registries, WG PID Information Types, IG Data Fabric & IG Domain Repositories by Gary Berg-Cross and Charles Vardeman.

Ontology Issues

Throughout the workshop a number of semantic challenges and issues were explored, particularly as they related to improve data understanding and metadata practice. In an era of big data and the desire for data sharing without boundaries a top-tier problem is how to make sense out of the data. This is especially a problem if we don't understand what the data are. As several participants noted we are at a point where formal semantics can now make metadata and data more understandable and hence more useful. But sense making requires more powerful semantic technologies and ontologies.[20] How do we introduce this into architectures, standards and current practice?

[20] See Why the Data Train Needs Semantic Rails

One strategy suggested was to focus on how to make the data smart. This would ease the burden of making applications "smart" because smart data will make the applications and services flexible and easy to use. Instead of developing increasingly complex software, the so-called business logic should be moved to the (meta) data. The rationale is that smart data will make all future applications more usable, flexible, and robust, while smarter applications fail to improve data along the same dimensions.

In analogy to hardware virtualization it was proposed that given a set of ontology design patterns and their combination into micro-ontologies, we could abstract from the underlying axiomatization by:

- Dynamically reconfiguring patterns in a plug and play style
- Bridging between different patterns an micro-theories
- Providing ontological views and semantic shortcuts that suit particular provider, user, and use case needs by highlighting or hiding certain aspects of the underlying ontological model
- Map between major modeling styles, e.g., the use of instances versus classes.

But, as noted by Pascal Hitzler, we are still searching for the sweet spot with a commonality of commitments to meaning. These commitments reflect specific modeling decisions, which are made in annotation, controlled vocabularies, lightweight taxonomy, models or full-blown ontology, and ontologies. You can either make detailed specifications (ontological commitments), which will often hinder reuse for new purposes, or you can avoid the commitments, resulting in ambiguity that cannot really be resolved later, thus also hindering reuse**.** Data integration and reuse can be hindered at either extreme, by insufficient commitment to semantic constraints or by too rigid ontological commitment. The implication you can visualize is that there exists some point or region between those extremes where data reuse is optimized. Given suitable commitments in a sweet spot the cost of data integration and reuse can be addressed by using a flexible, plug-and play semantically enriched metadata architecture. Several things need to be considered to make such a strategy work. One of which is to make ontologies: small and modular so that useful pieces can be understood and applied for particular purposes. A modular approach may be useful as explored and demonstrated in the Earth Science session at this workshop. ODPs as previously discussed may represent a modular approach to help with this effort. In analogy to hardware virtualization, a related set of ontology design patterns could be reconfigured into a plug and play style that suit particular providers and users. Some general issue to further consider include:

- Are we ready for metadata semantics to be widely used?
- If so, what tools would help and where are the opportunities? Do we just try it?
- Can we agree on common or domain principles (like modularity or building blocks) or some formal semantic requirements?
- What is an Integrated Metadata and Semantic layer Architectural Vision?
- What is included in the Core and Framework Semantics
- Is the Semantic Data Framework with multi-tiered interoperability developed at RPI a place to start?