Facilitating Data Discovery by Connecting Related Resources

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Antonia Rosati[a,*] and Matthew Mayernik[a]
[a] *Library and Archives, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80305, U.S.A.*
arosati@ucar.edu, mayernik@ucar.edu, (303) 497-1183

Abstract. In this study, we investigate two approaches to increase the discoverability and connectivity of resources on the web. The first approach is the use of semantic web data structures in RDF/XML, in particular the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) vocabulary for creating compound digital objects. The second approach is the use of Schema.org vocabularies for marking up html web pages to increase their visibility to web search engines. Through applying these two mark-up approaches to three case studies within the geosciences, we identify factors that help to evaluate their applicability to research data archives. Our analysis points toward the most efficient and effective markup for aggregating resources within research data archiving settings. We focus on factors that can lead to increasing public discoverability of datasets. Our evaluations are based on the following characteristics of each mark-up approach: ease of use, the available standards and vocabularies, the ease of interoperability, and the relation to data citation tools and methods.

Keywords: Linked Data, RDF, schema.org, datasets, environmental sciences

[*] Corresponding author. E-mail: arosati@ucar.edu.

1. Introduction

Federal agencies, professional societies, and research organizations across discipines are calling for researchers to make data that led to a given research result more discoverable and usable for secondary purposes [6, 12, 17]. These efforts promote transparency, traceability, and reusability of taxpayer-funded research. Much focus has been, and needs to be, given to the data sets and other resources that underlie peer-reviewed publications. Data sets are critical components of the research process, but are often very complex in structure and in their relations to other kinds of research products. The relations between varying resource types need to be formally documented and made visible on the web in both human and machine readable form in order to enable the widest possible reach for data sets to be discovered.

In order to use something, it must first be found; in order to find something, it must be named. More and more, research products, including data sets, publications, software, visualizations, etc., are posted on publicly available web sites. Many of those resources, however, are difficult to discover online because they are poorly described, stored in databases, or managed behind password protection, making them effectively invisible to web search engines and other web crawling systems.

This project focuses on the second step of the "name-find-use" sequence: enhancing discoverability of research data and related resources by making them more visible to web-based discovery tools.We explore how web-based mark-up infrastructures assist resource creators and users in making related scholarly resources more connected and discoverable. We investigate how to leverage the Semantic Web and other semantic mark-up systems to ensure the widest possible range of resource discovery methods. This is a more fundamental process than search engine optimization. Linking datasets, creators, standards, and users is an important step as it helps to increase the number of users and measure the impact of funded research.

1.1. Data Discovery

The amount of digital data online is increasing tremendously. Approximately one million government datasets are available online within the United States alone via Data.gov, and these represent a small proportion of the data available online. Scientific researchers are seeing a similar increase in data volumes and complexities [3]

Manually searching for data using search engines is an inefficient process. Web search engines, being designed to search for web pages and documents, perform poorly during searches for data sets [1]. Depending on the search, they can return too many results, too few results, and possibly irrelevant results due to terminology differences between the searcher and the data set creators. In addition, search results always require additional investigation because they return web pages, not data sets themselves.

An additional motivation for increasing the discoverability of research data is that data are often valuable to more than just those inside the science community. For example, the data within the Protein Data Bank (http://www.rcsb.org/) is widely used within the biosciences, but, unexpectedly, is also widely used by schoolchildren [4]. Effective data discovery systems can increase the impact of funded research by making these unexpected data uses more likely.

1.2. Data Connectivity

Research data sets do not exist in isolation from the rest of the world. Data sets are collected by people, using particular instruments and methodologies, and may be processed, filtered, or compiled into numerous derivitave data products. As such, many relations exist between data sets, as well as between data sets and data creators, data managers, research organizations, scientific publications, etc. In lieu of sophisticated data discovery tools, data users commonly use these connections to discover data, namely through reading about data sets in the literature, through receiving suggestions from personal contacts, or through particular organizations, such as NASA, NOAA, or NCAR, known for collecting trustworthy data [28].

At base, relationships are associations between two or more things. Relationships are central to all information systems [11]. Within information systems, relationships are used to associate meanings to entities, to combine entities that have common properties, and to enable the discovery of new resources and insights [22]. Relationships can be implicit or explicit, depending on the existence of specific named connections between two entities [23]. For web-based resources, formal syntactic structures allow relationships to be represented in standard machine-readable formats.

1.3. Study Outline

In this study, we investigate two approaches to increase the discoverability and connectivity of resources on the web. The first approach is the use of semantic web data structures in RDF/XML, in particular the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) vocabulary for creating compound digital objects. The second approach is the use of Schema.org vocabularies for marking up html web pages to increase their visibility to web search engines. Through applying these two mark-up approaches to three case studies within the geosciences, we identify factors that evaluate their applicability to research data archives.

First, we discuss the ways that research data sets are being made more identifiable and linkable through the use of unique identifiers.

2. Data Citations via Unique Identifiers

Creating a unique identity for a research data set is a complex task [19]. Data sets are often combined into composite data sets, or pulled apart into sub-sets. In addition, many data sets, such as climate observations, stock prices, and social media feeds, are highly dynamic, changing on a daily or weekly basis.

There is a growing movement across disciplines to assign actionable identifiers to data sets to facilitate the traceability and unique identification of data sets that led to published results. The "data citation" movement is making this trend more widespread. Data citations give datasets a persistent presence online by assigning a formal "name" and location, typically via Digital Object Identifiers (DOIs) [15]. DOIs provide a stable, permanent URL at which the resource can be found. DOIs are "actionable identifiers" in that they identify the resource, but also can be used in a URL form, which can take the user directly to the item being identified [13].

DOIs originated in the journal publishing world, but are receiving push to be used in a mainstream fashion for datasets [7].

The DataCite organization is building technology and services that enable organizations to assign DOIs to data sets and other research products (http://datacite.org). DataCite provides DOI registration and resolution services. DataCite requires metadata to be submitted along with a DOI registration request, and has created an XML metadata format to collect information appropriate to the discovery and management of data [24].

The significance of this activity is that this work is explicitly oriented towards making research data sets uniquely identifiable and named on the internet. While many data archives already serve data online, the act of creating persistent named identities for data sets is a recent product of these data citation initiatives.

3. Semantic Web

Web 1.0 is essentially a read-only form of the World Wide Web. It is a presentation of text, images, or even product catalogs to be browsed. Web 2.0 allows for interaction with a website, such as live chat customer service. Many websites, such as youtube.com and facebook.com rely on user-supplied content for their existence. Web 3.0 is essentially a way to bridge the communication gap between human web users and computerized applications. "One of the largest organizational challenges of presenting information on the web is that web applications aren't able to provide context to data, and, therefore, can't really understand what is relevant and what is not" [9]. Through the use of semantic markup, data are available in a form that is human and machine readable – allowing for a seamless interaction between humans and the vast amount of data on the internet.

In the following sections, we outline the data structures of interest to this paper. The Resource Description Framework (RDF), the Open Archives Initiative-Object Reuse and Exchange specification, which leverages RDF, and the Schema.org Microdata structure.

3.1. Resource Description Framework

The Resource Description Framework (RDF) is a World Wide Web Consortium (W3C) standard that specifies a model for data interchange on the internet [27]. RDF provides a machine-readable syntax in which named graphs can be created and used to managed networks of relations between entities on the web [8]. RDF is built from XML triples that use URIs to name entities on the web. RDF underlies the movement to create "linked data", namely data on the internet that can be published by anyone, links to standard vocabularies and ontologies, and can be linked to by others [4]. An important feature of the RDF model is that anyone can create a new vocabulary for any purpose. The linked data approach is that vocabularies should be made known to the larger linked data community, so that they can be leveraged by other users and linked back into the larger network of linked data.

RDF can be serialized into a number of formats in order to be made accessible for use. The most common serialization is RDF/XML, in which RDF statements are declared in XML structures. In RDF/XML, individual RDF vocabularies are declared via namespace designations and individual statements are made via standard XML syntax. RDF/XML is designed to be implemented as a stand-alone representation of data and metadata, with visual representations and computational processes built on top.

The exception to the standalone nature of RDF is RDFa, which is a serialization of RDF for embedding RDF statements into the HTML of web pages. RDFa statements are included in-line in the HTML code itself. RDFa is also an open and extensible specification, meaning that any user can create custom vocabularies. We did not directly compare RDFa to Schema.org, as prior comparisons have shown the significant difficulties that RDFa users have experienced, and the high error rates seen in RDFa implementations [20, 29].

3.2. Open Archives Initiative-Object Reuse and Exchange (OAI-ORE)

The Open Archives Initiative-Object Reuse and Exchange (hereafter called ORE) is a specification for how to describe and exchange sets of related objects on the internet (http://www.openarchives.org/ore/). The ORE project was created to address a gap in the ways that web resources were described, namely, the need for machine-actionable ways to create and describe groupings of resources in a web environment. Machine agents and Web services typically fail to interpret materials as related resources. Prior to ORE, groups of related resources could not be made visible on the web via URLs. ORE defines a standard for the identification and description of clusters of Web resources (known as "aggregations"). ORE provides a data model to assemble the URIs of related resources into an aggregation [14]. In the ORE model, each aggregation is gven a URI, as well as a description of its constituents and, optionally, the relationships among them.

When using the ORE data model, resources are brought together into "aggregations", which are in turn described by "Resource Maps". Resource Maps are machine-readable documents that describe ORE aggregations and can be expressed in a variety of formats, including RDF/XML.

A URI is used as the identity for referencing the aggregation. When the aggregation's URI is dereferenced by a human (e.g., via a web browser), a descriptive HTML page is returned that details the aggregation in a manner suitable for human consumption, typically as a "splash page". When the aggregation's URI is dereferenced by a machine agent (e.g., a web crawler), the aggregation's Resource Map is returned in whichever serializtion is specified.

Note that ORE allows a hierarchy of aggregations, such that one aggregation can refer to another; however, each aggregation must have its own Resource Map. The ORE ontology can be combined with other ontologies to achieve more expressive descriptions of the aggregated resources and their relationships. Through using additional ontologies, the Resource Map can leverage RDF/XML to express relationships and properties for the aggregation and its constituents in the linked data fashion.

ORE has been adopted for use in a variety of settings, including to aggregate digitized resources within the Library of Congress' National Digital Newspaper Program [26], climate data sets within the Advanced Climate Research Infrastructure for Data (ACRID) [21], and as a support mechanism for digital document repositories [25].

3.3. Schema.org and other microformats

Most Web applications sit on top of databases that contain vast amounts of structured data. However, in the translation between database and webpage, the structured data is converted to HTML for display, thereby losing its useful structure. What was once machine readable is now only human readable.

HTML5, the most recent update to the HTML standard, has limited semantic capability built in. Rather than complicate HTML5, Google, Yahoo!, and Bing announced Schema.org in June 2011. Schema.org uses HTML5 elements to embed semantic code into webpages. Microformats, such as Schema.org, are a collection of vocabularies for extending HTML with additional *machine-*

readable semantics [20]. Using microformats within HTML specifically allows search engines to understand the meaning of particular keywords. The structure allows a machine to make meaningful inferences and suggestions.

Schema.org is search engine focused. The simplicity of Schema.org is what allows search engines to make use of it. For example, Schema.org vocabularies can be extended by users appending self-made vocabulary terms onto any standard Schema.org vocabulary. However, search engines readily ignore this extra information until it becomes included in Schema.org as part of standardized vocabularies. Schema.org vocabularies are often commercial related – such as "product" or "recipe"; however, "scholarly article" and "dataset" are official pieces of the Schema.org vocabulary released in April 2013. Schema.org is in version 1.0 alpha at the time of this writing.

4. Case Studies

The case studies included in this study are all drawn from scientific research and information system design projects underway at the National Center for Atmospheric Research (NCAR). NCAR-related work were chosen based on heterogenaity of resources. NCAR is also in the process of assigning citable identifiers to data sets and other research products [16]. As noted below, the case studies here have also been involved in initial data citation tests.

We investigated how each markup type would apply to each case study. The recommendations derived from these tests point toward the most efficient and effective markup schema for aggregating resources within three case studies. Case (1), the North American Regional Climate Change Assessment Program (NARCCAP), is the most homogeneous of the three cases. NARCCAP provides access to data files in NetCDF format that were created through a closed group of collaborators on a finite project. http://narccap.ucar.edu. Case (2), is a conglomeration of animation and image files created by NCAR's Visualization Lab (VisLab). The visualizations from the NCAR VisLab are used to display and explain data from several NCAR groups, http://www.vets.ucar.edu. Case (3), the Advanced Cooperative Arctic Data and Information Service (ACADIS), houses multiple data types and formats across multidisciplinary and interdisciplinary boundaries from a variety of program frameworks and institutions. The ACADIS collection is a composite of data about Polar Regions. www.aoncadis.org.

4.1. NARCCAP

The North American Regional Climate Change Assessment Program (NARCCAP) is an international program to produce high resolution climate change simulations in order to investigate uncertainties in regional scale projections of future climate and generate climate change scenarios for use in impacts research.

Even though a variety of global and regional models are used to produce data, all data are in NetCDF file format and are made available through NCAR's Earth System Grid (ESG).

NARCCAP is the first NCAR dataset to acquire a DOI. It was used as the first test case for DOI assignment because the boundaries around the data set can be drawn relatively cleanly in comparison to some other data stores. The program has over 700 data users.

4.2. VisLab

The VisLab provides visualization products and services to many researchers and programs within and outside of NCAR. The VisLab manages their products on a lab website, and some items of high community interest are posted on NCAR-managed youtube.com channels. Often there will be many versions of the same visualization, such as successive iterations of particular products, or the same product posted in multiple locations online. In addition, sometimes there are shortened edits or interviews interjected into the video presentations for particular visualizations. VisLab resources also include articles or still images, which are often related to the animations or videos discussed above. Thus, multiple sources and versions of items need to be linked in this case study.

The focus for this paper was the Palmer Drought Resource Index (PDRI) videos on the VisLab website and on YouTube.

4.3. ACADIS

The Advanced Cooperative Arctic Data and Information Service (ACADIS) has created its own repository with a unique organization system. Data sets are contributed to ACADIS by hundreds of data creators in many file formats, including excel and word documents, NetCDF, csv, TIF, JPG, and shapefiles.

ACADIS is not a finite project like NARCCAP, it is open to user submissions on an ongoing basis. Version control on submitted products, therefore, is not simple. Also contributing to the complexity of the ACADIS case are the fact that submissions come from multiple authors, include data created through multiple methodologies and from multiple disciplines, and contain diverse readme and other filesfor each contributed data set.

Also, ACADIS is not wholly run by NCAR. The project is a partnership between multiple NCAR/UCAR entities and is led by the National Snow and Ice Data Center (NSIDC). With all of the above characteristics, it is the most heterogenous of the three case studies.

5. Results

5.1. NARCCAP and RDF

The application of RDF to NARCCAP resources was straight forward because XML metadata were already available and vetted through the DOI creation process. The NARCCAP data collection previously had been assigned one DOI. This DOI could be used as a central linking device for related journal publications and outside resources. The NARCCAP project maintains lists of publications and presentations which have been created using the NARCCAP data. The ORE protocol is well suited to aggregating these resources by declaring RDF vocabularies for describing publications, leveraging the metadata already created for NARCCAP data set via in the DOI XML document, and using the ORE vocabulary to describe the entire aggregation. Using the ORE vocabulary in this manner could make the NARCCAP resources visible for discovery as an aggregation, instead of as a distributed set of individual resources.

5.2. NARCCAP and Schema.org

The implementation of Schema.org with NARCCAP could not use the existing resource of the XML file as was the case with the RDF implementation. The standard DataCite metadata fields had to be remapped to existing Schema.org vocabularies. Implementation was limited to the NARCCAP information website since the repository (Earth System Grid) is controlled by an entity outside of NARCCAP's control.

At the time the research was conducted, the metadata field for "Data Set" was most closely related to a "Product" or "Creative Work". The NARCCAP DOI was most closely related to the available identifier fields of "productID" or "isbn". As Schema.org has under ongoing revision during the research, this limitation was temporary. Schema.org and Data.gov were working on a data set vocabulary. This vocabulary specific for data sets was released in the 1.0a release of Schema.org in April 2013.

An important aspect of the NARCCAP data in relation to the use of Microdata to enhance discovery, is that the data set is hosted on NCAR's Earth System Grid (ESG) and cannot be downloaded without registration, account approval, and logging into the system. Even though ESG has a login, the DOI enables linking using the DOI's URI instead of the password-protected individual data files.

5.3. VisLab and RDF

VisLab resources available on the lab website include videos, software packages, and other related resources. Each visualization is poster with certain metadata, such as Prinicple Investigator(s), visualization creators, and information about the data that underlie the visualizations, such as geospatial resolution. This information provides base level description, and, could be used within an ORE representation, once it wasmanually mapped into an RDF structure. Any additional metadata would also have to be manually added using appropriate RDF vocabularies for describing visualizations.

The YouTube versions of the visuzaliations could be included in the aggregation through the use of "sameAs" or "isVersionof" relationship declarations, which are standard relationship types within many vocabularies, including the Web Ontology Language (OWL, http://www.w3.org/TR/owl-ref/), and the Dublin Core Terms vocabulary (http://dublincore.org/documents/dcmi-terms/). Easy to use scripts or tools that allow users to input metadata in order to autogenerates RDF files using standardized vocabulaires would siginificantly make the process of using semantic mark-up simpler.

5.4. VisLab and Schema.org

Building on the NARCCAP case study, the VisLab web site has similar issues in potentially using Schema.org vocabularies. Firstly, the types of resources on the VisLab web site, scientific visualizations and software packages, are not standard types within the Schema.org vocabulary set. As noted in the RDF example, metadata are already available on VisLab pages for each visualization, and just needed to be marked up according to Schema.org vocabularies. The challenge is mapping the available metadata to Schema.org categories that were not developed for scientific materials.

Second, the resources posted on YouTube, although already benefiting from YouTube's high profile online, cannot be marked up in any semantc fashion, because there is no way for YouTube users to adjust the HTML code on YouTube pages. Relation links and Microdata markup can be done from VisLab webpages about resources posted on YouTube, but Microdata can be used to indicate relationships in the other direction.

5.5. ACADIS and RDF

In order to deposit datasets into the ACADIS repository, a creator must create metadata as well.ACADIS therefore collects a lot of metadata in their database, which is then displayed on the data set web pages.The metadata fields collected by ACADIS are organized in a custom schema. This custom scheme has been developed to address the diverse data description needs of of the submitted data sets. As of this writing, the fields are in the process of being mapped to the DataCite DOI XML schema, and DOIs have been assigned to a few data sets as pilot tests of the process. The ACADIS web site enables users to create nested collections of files, which is in essence a way of creating aggregations of related materials. The ORE vocabulary is thus less benfitial for internal creation of aggregations than for aggregating internally held data resources with matierals that are maintained elsewhere online, such as publications hosted on publisher websites. This would require mapping the internal data structure into the ORE vocabulary, along with other relevant metadata standards, such as the DataCite schema, once it is in place.

5.6. ACADIS and Schema.org

Applying Schema.org to the ACADIS Gateway is a fairly straight forward process of marking up the HTML on the data set web pages. However, most of the keywords used in ACADIS derive from the Global Change Master Directory (GCMD) keyword list, and are not in Schema.org vocabulary. While Schema.org has grown to include "dataset" as an entity, it does not seem likely that GCMD keywords will be included in the format because of their subject specific focus. The "dataset" vocabulary will be able to provide information to search engines that is not as specific as ACADIS maintains internally, but it will at least identify resources as data sets, thus enabling them to be more discoverable.

Schema.org can, however, assist in linking people such as the PIs, CoPIs, data managers, data collectors, etc., who participate in the scientific projects that ACADIS supports, by providing rich vocabularies to annotate information about people.

6. Discussion

In the following discussion, we outline the three evaluation criteria that we used to compare the possible uses of the mark-up approaches for the case studies. The evaluation criteria used were: ease of use, the available standards and vocabularies, ease of interoperability, and the relation to data citation tools and methodologies.

6.1. Ease of use

RDF requires a triple store, which may be overwhelming to novice and intermediate users. It is based on XML, which is by design extensible for customization via self-constructed vocabularies. These kinds of customization require a high degree of knowledge in order to implement robust XML schema definitions and uses. In the case of Microdata, however, any webmaster can use Schema.org markup. It can be implemented piecemeal, in the sense that it can be implemented on any webpage with any level of detail that the user requires. HTML, the foundation for microdata formats, is highly standardized and underlies the vast majority of the material found on the internet.

Microdata makes HTML code less human readable because the information is inserted into HTML code between sentences, instead of being placed in the header as easily visible <meta> tags. Overall, the distinction between RDF and Microdata is the contrast between simple and easy to use vs. powerful, extensible, and more complex.

6.2. Available standards and vocabularies

A large number of semantic web-enabled vocabularies have been created on innumerable topics, including standards for locations and datasets. Many vocabularies have been created in relation to geoscience information, including vocabularies specific to geonames, geospatial locations, and geoscience more generally [2, 18].

Up until April 2013, Schema.org has been lacking in specific vocabularies for academia and datasets. To use Schema.org for our case studies, we used the generic "Thing" and "Creative Work" tags within Schema.org due to the lack of any more appropriate options for describing academic work, and data sets in particular. This gap was noted in the microdata community, and more academic focused vocabularies have been developed, such as a vocabulary for datasets (http://schema.org/Dataset). The vocabulary is still in its infancy.

Oddly enough, adding complexity may take away from Microdata's appeal. The vocabularies in Schema.org are simple, small, and constrained in topic. RDF, in contrast, is a general purpose tool. The first task of figuring out what to do with it can be an enormous impediment. Without direction, the number of possibilities in vocabulary and ontology selection, triple-store technologies, and data display tools require significant investigation and expertise.

6.3. Ease of Interoperability

Both RDF triple stores and Microdata are growing in use. In addition, vocabularies are multiplying fast for both markup types. Because of this ever increasing number of available vocabularies and ontologies, data repositories are still in the wild west when it comes to data interoperability and standardization.

Microdata is an important part or function in HTML5. In part, HTML5 was created to harness Microdata's capabilities [20]. Web development tools are including Microdata functionalities. For example, the Drupal content management system has a module that will automatically markup pages with Microdata (http://drupal.org/project/microdata).

One issue potentially impeding the utility of any web-based markup for data discovery is the reality that many reearch data systems require authenticated logins. Logins are useful for data archiving teams for generating usage metrics and enabling user contact, but they also serve as a de facto firewall in preventing search engine indexing and discovery. Some techniques are available to instrument password protected web sites to allow crawling by search engines, such as the "First Click Free" approach to making restricted information available for indexing by search engines. These techniques, however, must be customized for each search engine, and in the process provide another barrier that web development must hurdle.

Thus, there may be little incentive to mark up data repository web pages that require logins with Microdata if the web search engines will not be able to see beyond the splash pages of the portals.

6.4. Relation to data citation tools and methodologies

Assigning DOIs to resources via the DataCite service involves a) registering a web-accessible name that is referenceable and persistent for a resource, and b) creating an XML metadata description of your resource that is registered along with the DOI. The DataCite data citation metadata XML is more easily leveraged in RDF applications than in Schema.org applications, though knowledge of XML is not the same thing as knowledge of RDF. The W3C have a working draft of a conversion of microdata to RDF [10], but conversions in the other direction – from RDF to Microdata – necessarily have to be custom written for each RDF vocabulary used.

Thus, assigning DOIs to digital resources such as data sets is a first step to making the data set more discoverable and connected through semantic mark-up. Resources are given a uniquely referenceable name on the web, and as such can be leveraged as an entity in RDF or Microdata statements. Additional benefits require custom metadata mapping into the RDF or Schema.org schemas being used for a give application.

7. Conclusions

RDF is an accepted standard backed by the W3C for the exchange of web data. Schema.org is a more recent development with strong backing from Google, Yahoo!, and Bing. Many RDF vocabularies are well established and connected, whereas the vocabularies within Schema.org are still in early development.

Schema.org has the benefit of having a particular use case: increasing the visibility of web pages to search engines. Schema.org opens up data discovery for a whole different audience and it can be implemented with minimal effort, but it lacks the vocabularies in comparison to RDF. RDF can be leveraged for resource discovery outside of the search engine environment, and can be leveraged to produce new methods of enabling resource discovery through machine-readable relationships, as for example through the ORE vocabulary.

For a large research organization like NCAR, Schema.org offers the simplicity of mark-up in HTML code, which can be implemented on any locally managed web site. With RDF linked data, the biggest benefits will be seen for an organization if there is wide adoption and significant resources put in to enable data linkages through appropriate vocabularies and distributed data stores.

The official naming of a dataset with a DOI allows each markup schema to link resources better, because the data become uniquely referenceable entities online.

Even though Schema.org is based on triple statements, like RDF, Schema.org vocabularies are very simple to use and understand. Schema.org has grown tremendously in the last few years, and is expanding in scope with regulariety. The aforementioned Schema.org vocabulary for data and data sets, released as Alpha 1.0 in April 2013, could be a significant boon for labeling data resources in a way that will enable them to be more visible to public search engines.

For most other applications of semantic technologies, however, RDF provides flexibility and the possibility for integration with external vocabularies and data sources.

8. References

[1] G. Antoniou and F. von Harmelen, A semantic Web primer, Cambridge, Mass.: MIT Press, 2004.

[2] C. Baru and K. Lin, Mediating among GeoSciML resources, International Journal of Digital Earth, 2(S1), (2009), pp. 18-28, http://dx.doi.org/10.1080/17538940902912437

[3] G. Bell, T. Hey, and A. Szalay, Beyond the Data Deluge, Science, 323 (5919), (2009), pp. 1297-1298, http://dx.doi.org/10.1126/science.1170411

[4] C. Bizer, T. Heath, and T. Berners-Lee, Linked Data - The Story So Far, International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), (2009), http://dx.doi.org/10.4018/jswis.2009081901

[5] P. Bourne, Towards Data Attribution and Citation in the Life Sciences, In P.E. Uhlir (Rapporteur) For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. Board on Research Data and Information, Policy and Global Affairs, National Research Council of the National Academies. Washington, D.C.: The National Academies Press, 2012, pp. 43-48, http://www.nap.edu/catalog.php?record_id=13564

[6] M.J. Costello, Motivating Online Publication of Data, BioScience, 59(5), (2009), pp. 418-427. http://www.jstor.org/stable/10.1525/bio.2009.59.5.9

[7] R. Duerr, et al, On the utility of identification schemes for digital earth science data: an assessment and recommendations, Earth Science Informatics, 4(3), (2011), pp. 1-22, http://dx.doi.org/10.1007/s12145-011-0083-6

[8] H. Ebner and M. Palmér, An information model for managing resources and their metadata, Semantic Web Journal, 2011, http://www.semantic-web-journal.net/content/information-model-managing-resources-and-their-metadata

[9] B. Getting, Basic definitions: Web 1.0, Web. 2.0, Web 3.0. Practical eCommerce, April 18, 2007, http://www.practicalecommerce.com/articles/464-Basic-Definitions-Web-1-0-Web-2-0-Web-3-0

[10] I. Hickson, G. Kellogg, and J. Tennison, Microdata to RDF: Transformation from HTML+Microdata to RDF, W3C Working Draft, 12 January 2012, http://www.w3.org/TR/2012/WD-microdata-rdf-20120112/

[11] W. Kent, Data and reality: basic assumptions in data processing reconsidered, New York: North-Holland Pub. Co, 1978.

[12] T. Killeen, Dear Colleague Letter -Data Citation, NSF 12-058, March 29, 2012, http://www.nsf.gov/pubs/2012/nsf12058/nsf12058.jsp?WT.mc_id=USNSF_25andWT.mc_ev=click

[13] J. Kunze, Towards Electronic Persistence Using ARK Identifiers, 2003, http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf

[14] C. Lagoze, H. Van de Sompel, M.L. Nelson, S. Warner, R. Sanderson, and P. Johnston, A Web-Based Resource Model for Scholarship 2.0: Object Reuse and Exchange, Concurrency and Computation: Practice and Experience, 2010.

[15] M.S. Mayernik, Bridging data lifecycles: Tracking data use via data citations workshop report, NCAR Technical Note NCAR/TN-494+PROC, Boulder, CO: National Center for Atmospheric Research (NCAR), 2013, http://dx.doi.org/10.5065/D6PZ56TX

[16] M.S. Mayernik, et al., Data citations within NCAR/UCP, NCAR Technical Note, NCAR/TN-492+STR. Boulder, CO: National Center for Atmospheric Research (NCAR), 2012, http://dx.doi.org/10.5065/D6ZC80VN

[17] National Science Foundation (NSF), Dissemination and sharing of research results, 2013. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[18] R. Raskin, Development of ontologies for earth system science, In Sinha, A.K., (ed.), Geoinformatics: Data to Knoweldge: Geological Society of America Special Paper 397, 2006, pp. 195-199, http://dx.doi.org/10.1130/2006.2397(14)

[19] A. Renear, S. Sacchi, and K. Wickett, Definitions of dataset in the scientific and technical literature, Proceedings of the American Society for Information Science and Technology, 47(1) (2010), pp. 1-4. http://dx.doi.org/10.1002/meet.14504701240

[20] J. Ronallo, HTML5 Microdata and Schema.org, The Code4lib Journal, 16 (2012).. http://journal.code4lib.org/articles/6400

[21] A. Shaon, et al., A Linked Data Approach to Publishing Complex Scientific Workflows, E-Science (e-Science), 2011 IEEE 7th International Conference on, 2011, pp.303-310, http://dx.doi.org/10.1109/eScience.2011.49

[22] A. Sheth, I. B. Arpinar, and V. Kashyap, Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships, in Enhancing the Power of the Internet, vol. 139, M. Nikravesh, B. Azvine, R. Yager, and L. Zadeh, Eds. Springer Berlin Heidelberg, 2004, pp. 63–94.

[23] A. Sheth and C. Ramakrishnan, Relationship Web: Blazing Semantic Trails between Web Resources, IEEE Internet Computing, July-August 2007, pp. 84-88.

[24] J. Starr and A. Gastl, isCitedBy: A Metadata Scheme for DataCite. D-Lib Magazine, 17(1/2) (2011). http://www.dlib.org/dlib/january11/starr/01starr.html

[25] D. Tarrant, et al., Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications, The Code4Lib Journal, 6 (2009), http://journal.code4lib.org/articles/1062

[26] M. Witt, Object Reuse and Exchange (OAI-ORE), Library Technology Reports, 46(4) (2010).

[27] World Wide Web Consortium (W3C), Resource Description Framework (RDF), 2004, http://www.w3.org/RDF/

[28] A.S. Zimmerman, Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse, International Journal of Digital Libraries, 7(1/2) (2007), pp. 5-16.

[29] Adams, S. (2012). HTML5 Case Study 1: Semantics and Metadata: Machine Understandable Documents. *UKOLN, May 21.*