Sean Chen. A Prototype Collaborative Directory of Metadata Standards. A Master's Paper for the M.S. in L.S. degree. April 2014. 44 pages. Advisor: Jane Greenberg

The implementation and development of a prototype metadata directory to support the discovery and identification of metadata standards, used in scientific research, is reported on. The project was undertaken as an initiative of the Research Data Alliance's Metadata Standards Directory Working Group in 2014 and makes more accessible information about metadata standards used by scientists in a range of disciplines in order to support the exchange, management, curation and preservation of scientific data. Other important benefits include: supporting the reuse of standards to eliminate duplication of effort, and to enable reproducible scientific research.

The report examines the design of the directory; including the use of a distributed version control system as a mechanism for content management. A discussion of the sustainability and risks, and a comparison to other collaboratively managed information systems is made. Future features and extensions applicable to the prototype are also addressed.

Headings:

Metadata

Electronic directories

Web development

Website authoring programs

Electronic data processing

A PROTOTYPE COLLABORATIVE DIRECTORY OF METADATA STANDARDS

by
Sean Chen

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina
April 2014

Approved by:

_____
Jane Greenberg

A Prototype Collaborative Directory of Metadata Standards

Introduction

Metadata is necessary for the discovery, access and management of data. It enables a whole of range of activities and features in many systems including search and retrieval, user experience and reusability. The report that follows details the implementation and development of a prototype directory that facilitates the exchange, management, curation and preservation of research data as proposed in the Research Data Alliance (2013). The directory continues and extends preliminary data gathering that was conducted by Perez (2013).

A metadata directory is an information search and retrieval system containing information about metadata standards. Similar to a directory of personnel in a corporate information system in providing relevant information about people in the corporation or business, such as personal identifiers, contact information, and physical locations, a good metadata directory should provide users with information about metadata standards to make decisions about their use. The depth of information can vary, but can commonly include: descriptions of the standard, relevant tools, and who or what is using the standard. Such information contained in a metadata directory can help system designers, scientists, and data curators in exchanging data, or even as a first step, discovering the relevant standards that might apply to a research project throughout its entire data management life cycle.

Efficiently sharing data with collaborators and other researchers is a primary concern of many scientists working in a number of domains (Tenopir et al., 2011). Sharing of data is a necessary step in the scientific process for a number of reasons including: testing validity, replicating experiments, and supporting reliable results. Validity means allowing other researchers, reviewers and funders to view the original data and make their own conclusions. Similarly, replication is a basic function of the scientific method and having data available affords this purpose. Finally, reliability is improved by offering a transparent view of research results and the data that supports them, thus it is evident that being able to share data would be a prerequisite to reaching steps of reliability and transparency.

Sharing of data depends upon agreements between the funding agencies, data producers, data users, data managers, and maintainers. These agreements are the affordances that allow communities of interest to exchange and share their research data between themselves and the rest of the world. The scope of agreements can include protocols for exchange of data, which can cover legal and regulatory protocols protecting privacy, to computer protocols that provision access to data within a networked environment. Other portions of these agreements could include regulatory requirements to funders and to requirements by institutions that mandate the sharing of and public accessibility of scientific. Agreements can be based on the structure of the data; but also include specifications about the packaging or structural level of detail that facilitate the sharing and exchange of the data. Lastly the semantic specification of the data needs to be agreed upon. This could include the description of the required data elements and

attributes; it could specify the relationship between data elements. This often falls under the scope of what is commonly described as metadata standard (Tenopir et al., 2011).

People who are interested in this work include the following: (a) scientists who are publishers of the data, (b) scientists who are consumers, (c) data curators and people responsible for the management of the research data, (e) funding agencies and personal, and (d) the general public. These many stakeholders can possibly use information from within a metadata directory in a number of identified ways. More importantly having a directory would enable uses that cannot easily be identified, in a sense providing essential infrastructure for future uses and reuses.

This sort of system exposes information about metadata standards that can enhance the discovery and use of the most appropriate standard for a project. By having a visible directory on the open web, metadata standards can be quickly identified and compared using such a system. Framed in another way, directories solve a problem of discovery. Identifying what appropriate metadata standard to use requires integrating a number of factors. Some of these include requirement issues include: what does the research project require in terms of its goals. Others include resource levels: how much time, or, what are the available money and staffing available to implement the metadata standard for the research project. Social and adoption factors can also play a role in selecting a standard. A metadata directory that provides information about who has currently adopted a standard or is developing a tool can give insight for possible adopters as to the currency, support and prevalence of a metadata standard.

Using a metadata standard that has been adopted by other researchers and institutions solves a range of problems with managing research data. These include

interoperability issues: such as making data generated from different research projects be analyzed be integrated for larger scale analysis. In addition, having access to the fruits of research in the form of data that subscribes to a particular standard allows the public to access and use the data for novel purposes. Another benefit is that research data that is presented with standardized metadata becomes more usable in the teaching and enabling students to become fully conversant with their chosen scientific disciplines (Kriesberg, Frank, Faniel, & Yakel, 2013). However, standards need a level of adoption in order to achieve the positive externalities that such widespread adoption can offer. If only one scientist is using a standard and no one else has adopted it, the benefits are minimal. On the other hand, if an entire community of scientists has adopted a particular standard for their projects then that whole community benefits. In other domains this is known as a network effect, for example, in computing Glider (1993) recounted Metcalf's Law that states that the value of a network is in proportion to the square of the number of nodes in a system. Similarly, directories help project designers and data managers identify what is most appropriate for their projects. Having a well-accepted standard helps smooth over many obvious problems such as: resource requirements, interoperability, consistency and consistency.

This work reports on the work and design of a functional prototype metadata directory. A prototype directory at the time of the writing of this report is located at http://rd-alliance.github.io/metadata-directory. The organization of the paper is as follows: (a) *literature review* of related topics, (b) *project overview* describing the context and background of the project, (c) *project design* specification, (d) *directory management*

issues (e) detailing of *project outcomes* (f) a *discussion* of the project with an eye toward sustainability and risks, and (f) a *conclusion*.

Literature Review

The development of a prototype metadata directory is informed by work done in a number of areas including: metadata, metadata registries, research data usage by scientists, collaborative work, and standards development and diffusion. Relevant literature on metadata in general and more specifically on metadata directories and registries is reviewed in order to inform the design and to discover prior art. The usage of research data by scientists and their communities is also examined in order to understand the user community for such a directory. Similarly, because the directory will be used in a collaborative manner, a review of computer supported collaborative work is also made. Lastly the development of standards and their distribution is considered in order to understand how the relevant communities of interest can further adopt the project.

Metadata Registries

Metadata is "data about data". Hillmann, Marker, & Brady (2008) categorize metadata into five functional areas: (a) administrative, (b) descriptive, (c) access (d) preservation, and (e) structural. A *metadata standard* is then the organized operationalization of a set of metadata aimed toward a particular application or domain of resources, documents, objects, or data (Chan & Zeng, 2006). Standards are also referred to as *schemes*, *schemas* or *element sets*. But for the purposes of this project are being called *standards*.

Metadata registries have been the primary focus of research and development in the metadata communities. A registry is an information system that allows both humans and machines to manage and maintain metadata standards; and more importantly, they provide an interface to technical information about the contained metadata standards.

This information about a particular standard include: (a) term vocabularies, (c) definitions of the term, (d) appropriate values, (e) namespace control, and (e) administrative information related to the standard (Day, 2003). Registries do provide a certain directory function by listing and providing information about a subject metadata standard. However, they are designed as technical systems, providing public interfaces that enable programmatic use of metadata.

Metadata registries have had ample standards development surrounding them. One example of that is the international effort to define a standard to describe metadata standards and metadata registries across a wide range of domains. The results of this effort is the ISO/IEC 11179 (2004) standard, also known as the ISO/IEC 11179 Metadata Registry (MDR) standard). The standard is a generalized framework for the description of data. The goals of the standard are to establish: (a) a standard description of data, (b) establish common understanding of those data across organization units and between organizations, (c) promote re-use and standardization of data across time, space and applications, (d) harmonization and standardization of data within organizations and between them, (e) facilitate the management of components of data, (f) and to facilitate the re-use of those components of data.

Metadata registries as a class of systems are oriented toward business use cases. With an origination in the Electronic Data Interchange (EDI) world, they are highly specified systems that enable interoperability between different information systems. They provide a way to mechanically access and analyze data in mediated fashions. A common use case would be providing a developer or designer with the technical information that would enable them to make data from one system work with data from

another (harmonization) or to be able to transfer data from one system to another (data exchange) (ISO & IEC, 2004).

For libraries and archives however the needs of managing and exposing metadata in a manner that works across applications is a primary concern. In addition, registries are used to help ensure that metadata is consistent and authoritative across information systems, a chief concern for work in digital preservation. The *Metadata Registry* (http://metadataregistry.org) is an effort by the library and archives metadata community to provide a platform to enable registering standards in a way that achieves similar goals to ISO/IEC 11179 (Hillmann, Sutton, Phipps, & Laundry, 2006). Similarly efforts in digital preservation and metadata interoperability have helped establish registries such as reported by Day (2003).

Scientific Research Data

Research data sharing varies widely by discipline, however rationales for sharing research data can include (a) enabling the reproduction or verification of research results, (b) providing public accessibility to research results, (c) enabling extensible research on existing data, and (d) to push the state of research and innovation forward (Borgman, 2012). Each of these facets is an important internal and external reason for developing a metadata directory.

Scientists want to share data but don't know how to and if they do, want to get proper credit for doing so (Tenopir et al., 2011). Getting credit and citing scientific data are important motivators (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013). Others have observed that there is a substantive duplication of effort being taking place among scientists and between communities of interest and practice

(Willis, Greenberg, & White, 2012). Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis (2011) observe that a substantial portion of research papers published in high-impact journals do not make available their data either by some sort of institutional policy nor to the very instructions of the publication. All of these factors indicate that one primary reason for difficulty in sharing research data is the lack of an appropriate metadata standard, thus justifying the need for a metadata directory as a clearinghouse and discovery tool for relevant metadata standards.

Collaborative Work

Wikis have been a long time tool for working as a collaborative tool for knowledge management. The success of Wikipedia is an example of that. Sharing information is difficult however and there aren't many examples of where the characteristics of a community will have an effect on the use of a wiki. Wang & Wei (2011) examine some of these characteristics that influence information sharing within wikis, finding that member interactions, participation and promotion are all important factors. Others have looked at some of the characteristics of the community members who contribute to wikis (Yates, Wagner, & Majchrzak, 2009) and found that the reconfiguration of existing content needs to be examined in a different way, and noted that contributions are heavily influenced by the technology platform and affordances that form norms from which content can be shaped.

Source Control Management (SCM) has been a principle issue for the development of software for many years (Conradi & Westfechtel, 1998). In recent years the dominant tools for source control management in the open source software communities have been moving toward a distributed model of version control

(O'Sullivan, 2009). A version control system allows members of a development team to perform two principal tasks: the first is viewing the history of changes within a number of files and documents. The second, being the ability to work independently of others working on the same documents while then later being able to merge the work from independent branches back into a canonical version of the project.

Standards Adoption

Spreading of standards can be looked at within a diffusion of innovation model (Rogers, 2003). Standards are an innovation in the sense that they represent something novel that needs to be adopted by interested parties or in the context of a system, users and designers. Rogers (2003) articulates five factors that influence the diffusion of innovations: (a) relative advantage, or, how much more improved the innovation is over previous innovations in the same domain, (b) compatibility, or, how much disruption the innovation will cause with existing patterns, (c) complexity, or, if the innovation is difficult to use, (d) trialability, or features that make the innovation easy to test or experimented with, and finally (e) observability, or how quickly the innovation can be spread through communications channels in a social group. The proposed metadata standards directory should address all of these factors in order to improve its adoption by users.

The directory by being an effective agent of standards information transfer should help improve the adoption of research data standards. Benefits of improved adoption can be viewed through the lens of network effects. Which are well understood benefits that come with standardization (Church & Gandal, 1992). Others have looked at modeling how standardization is aided by network effects (Weitzel, Beimborn, & König, 2006).

However, standardization can be a double edged sword where standards often lock industries or disciplines into a particular standard which is suboptimal for future developments (Farrell & Saloner, 1985).

The research reviewed above has implication for any metadata registry or directory initiative and provides insight into various aspects and issues. The literature reviewed has been particularly helpful in contextualizing the needs and goals for the RDA Metadata Standards Directory effort, and informed a series of steps to explore GitHub as an environment for prototyping a design. The remainder of this report specifically addresses a prototype for the RDA Metadata Standards Directory by providing a project overview, detailing the project design, addressing a proposed method for managing the directory, describing project outcomes, and discussing issues of sustainability and project risk.

Project Overview

This project reports on an international initiative to *develop a prototype directory* listing metadata standards applicable to scientific data. The directory covers metadata standards in wide use among scientists and the rest of the research data community. Requirements for this prototype directory were included in a project road map (Research Data Alliance, 2013). These requirements have been distilled as follows:

- *Collaboration*. The principal requirements of this prototype include an ability to openly and collaborative contribute and edit the directory, which will enable the documentation of metadata standards.
- *Categorization*. This metadata directory will list those metadata standards, categorize them and list important attributes of each of those standards.
- *Identification*. The metadata directory should support extensibility in its data model to include descriptions for and links to additional metadata standards and related vocabularies such as ontologies, authority lists and taxonomies that support the identification and discovery of metadata standards
- *Contextualization*. The directory should provide a platform for contextual information such as sponsorship, adoption and ease of use
- *Sustainability*. The directory should be able to be sustainably maintained by a collective group in a manner similar to the how the Research Data Alliance works

Secondary requirements include designing a framework that affords a wider range of participation as an interface or node for social interaction between contributors. Similarly, having data that can be exposed in ways that promote reusability and high

visibility on the World Wide Web is another important consideration. Previous efforts such as by Ball (2009) and Qin & Small (2008) provide evidence for a need for such a metadata directory, though neither effort has been subsequently sustained and updated in a collaborative and collective manner. However, this prototype metadata directory can be built in such a manner that addresses the need sof researchers and research data managers to find and locating appropriate metadata standards. Finally, a plan for operating the directory will be included which will discuss the necessary resources and steps for maintaining and sustaining the directory. A discussion of the risks to the success of a directory will also be included. Further information on the specification and requirements of the directory are contained in the Research Data Alliance's Metadata Standards Directory Working Group's project plan and preliminary report (2013) and detailed in the following project design section of this report.

Project Design

This project acts on the goals articulated by the agenda of the Research Data Alliance's Metadata Working group (Greenberg, Jeffery, & Koskela, 2013). This project concentrates on developing a prototype metadata directory, one of the Working Group's short-term goals:

> *The short term goals of the Research Data Alliance's Metadata Working groups include:*
> - *a prototype wiki-based directory listing metadata standards applicable to scientific data.*
> - *Use cases to analyze and document the functional sues of metadata (e.g. for discovery interoperation, reuse)*
> - *Operational plan for supporting collective, sustainable growth and maintenance of the RDA Metadata Directory and associated recommended practice guidelines.*

Working with existing communities and projects is an important consideration when developing this plan and in gathering resources to support the directory. Building off of, and, acknowledging existing efforts is an important step in order to minimize duplication of effort and to highlight the contributions made already to this effort. The prototype will provide a collaboratively sourced and maintained directory where contributions of data concerning metadata standards from across disciplines and domains can be accepted and integrated with other contributions. The prototype should cover a number of use cases including both human and machine ones. In addition a plan for sustaining and providing support for the directory should also be seriously considered. However, at this stage, there is not a sustainability plan, although Working Group members are aware of this need, and it is noted as a necessary steps in the group's charter.

Content

The UK-based Digital Curation Centre (DCC) has developed a listing of metadata standards that is being used as the basis for the prototype. Effort has been undertaken by the DCC to enhance that data but experience has shown that maintenance was not feasible for the limited amount of staff time available. In addition, there appear to be limits in extending the functionality of the existing DCC metadata directory in a direction that would meet all of the needs of the Metadata Standards Directory Working Group. But, existing descriptions of metadata standards could be enhanced and new content could be added with assistance, to the DCC website. Data on metadata standards and their associated implementations was added to the existing corpus through an information gathering project that included an email appeal and web form survey (Perez, 2013). As part of this project the data was incorporated into the existing DCC directory through a process of normalization, summarization and data entry. Following this, the data on metadata standards and their associated use cases tools and extensions would then be transformed and then imported into a prototype directory.

Structure

Directory content is structured into three main categories (a) standards, (b) extensions, and (c) tools. A simple ad-hoc functional metadata standard was used to describe the data about other metadata standards. This scheme for the description of metadata standards is functionally oriented toward the requirements of the metadata directory. This is justified because of the prototype nature of the metadata directory. Any expectations for export into a more interoperable form can be easily addressed once an appropriate mapping is generated, though the existing schema may not actually address

other requirements that the target may have. However, the schema for the metadata directory is easily extensible and the structure of the directory, with its simple architecture, allows modifications such as additional attributes, relationships to new entities, or taking advantage of community developments such as controlled vocabularies or other metadata standards. In fact the primary concern when extending the existing content would be the ability of the contributors to manage any extension of the data for previously entered data.

Standards. Standards are a set of metadata elements that are being used to describe data. Standards within the directory framework are not necessarily linked to an encoding or file format, but are just the sets of metadata elements as specified. Our identified attributes of a metadata standard are shown in Table 1.

Table 1

Metadata Standard Attributes

Attribute Name	Description
title	Full name of the metadata standard
name	Short filename for the metadata standard
subjects	Higher level subject category
disciplines	Associated academic disciplines
specification_url	URL for the standard's specification
website	URL for the standard's home website
related_vocabularies	Related controlled vocabularies
related_vocabulary:name	Name of the related vocabulary
related_vocabulary:url	URL of the related vocabulary
mappings	An encoding or file format associated with the standard
mapping:name	Name of the associated mapping
mapping:url	URL for the associated mapping
sponsors	Sponsoring organization
sponsor:name	Name of the sponsoring organization
sponsor:url	URL of the sponsoring organization
contact	Contact name
contact_email	Contact email address
standard_update_date	Latest date the standard was updated

version	Version numbering of latest standard version
description	Textual description of the standard

Extension. Extensions are metadata standards that are derivatives of an existing metadata standard. They are often designed or specified to fulfill a particular need by a community or software system.

Table 2

Extension Attributes

Attribute Name	Description
title	Full name of the extension
name	Short filename to identify the extension
website	URL to a website supporting the extension
description	Textual description of the extension
subjects	Higher level subject category
disciplines	Associated academic disciplines
standards	Related standards to the extension

Use Cases. Use cases are implementations of a metadata standard within the context of a service, project or organization. This data potentially helps users, stakeholders and others who might be interested in using a standard identify others who use the metadata standard or how the metadata standard is being used.

Table 3

Use Case Attributes

Attribute Name	Description
title	Full name of the use case
name	Short filename to identify the use case
website	URL to a website supporting the use case
description	Textual description of the use case
subjects	Higher level subject category
disciplines	Associated academic disciplines
standards	Related standards to the use case

Tools. Tools are piece of software or a software application that is known to use the metadata standard to do some amount of work. For example, tools can enable the maintenance, curation, publishing, analysis, or recording of scientific data.

Table 4

Tool Attributes

Attribute Name	Description
title	Full name of the Tool
name	Short filename to identify the Tool
website	URL to a website supporting the Tool
description	Textual description of the Tool
subjects	Higher level subject category
disciplines	Associated academic disciplines
standards	Related standards to the Tool

Content Management

Content that had been gathered during a preliminary information gathering phase (Perez, 2013) it was structured into the devised schema data. Then it was entered into a content management system that constituted the Duration Curation Centre's (DCC) metadata standards information. At the time of the writing of this report the DCC's metadata standards content is presented under a *disciplinary metadata* section of their website at: http://www.dcc.ac.uk/resources/metadata-standards. Using a fully featured content management system like the one being used by the Data Curation Center, poses development and sustainability issues such as: (a) not having a stakeholder who can maintain the website, (b) administrative issues, (c) and also extensibility issues.

Other options for content management system included a standard wiki interface such as supported already by the Dublin Core Metadata Initiative (DCMI) (http://wiki.dublincore.org/index.php/Main_Page) However issues with wikis were identified: (a) account management is difficult in a community which does not have an

identified entity that can support that infrastructure, (b) security issues include spam and spurious data, (c) and a lack of extensibility.

Also identified as an approach is the use of a simple web publishing framework that would generate a statically generated website. This approach uses a templating system to convert simply formatted content into a fully developed website using statically generated HTML web pages. This has advantages including: (a) easy of extensibility, (b) content that simply says what it is, (c) an elimination of the need to maintain some cumbersome authentication and access system, and (d) the ease with which other tools can be used to work on such a system. However, such a generated website system would need some sort of web hosting in order to deliver the content. Fortunately, there are a number of services and a possible way of actually managing contributions to such a system, which will be further investigated in this report.

One principal risk however is that members of the metadata and scientific community would be unable to use a version control workflow because of a perceived high barrier to entry such as a learning curve or needing to install local software. Secondary risks are similar to those with a more fully featured content management system such as not having (a) an identified maintainer of the project, (b) administrative and access issues, (c) and a need to have expertise in order to extend the existing code and template base.

Git: Distributed Version Control System

Distributed version control systems, such as Git, allow many editors to do changes on a shared set of documents. These systems work on a model of a shared remote repository being defined as canonical, from which all changes would be then

distributed. This central repository is often called the remote. Each user of the data makes a clone of the centralized repository onto their local system. These distributed nodes then can contribute changes back to the centralized repository.

Basic actions in the Git version control system include *committing* a set of revisions to a repository. This process adds changes to an index of changes along with information about the change recorded by the user committing the changes. Following the committing process those commits are then usually *pushed* to a remote repository. A *push* puts the changed code on the remote repository using a variety of protocols, such as HTTP. When changes are made to the centralized repository the changes are then distributed to other remotes through two linked processes, the first is the *fetch,* which retrieves the changes made to a remote or centralized version of the code. The second is *merging,* which is the Git feature that integrates those changes into the localized version. These two processes, the *fetch* and the *merge* are then combined together into what is typically called a *pull*; which is simply a fetch of a remote repository's differences from a local repository and then a subsequence merging of that data with the data or code on the local repository. The tricky thing that then happens with merges is that the code often can not be merged automatically because changes in the local and the remote versions conflict, this needs to be managed manually, using a tool that lets the user select a method to resolve each conflict in order.

One other major feature of the distributed version control system is the concept of a branch. A *branch* is a set of commits that differ from a master or canonical project. The way that Git works it that it is easy to create a branch, then work within a branch by making a number of changes, and then automatically *merge* those changes back into the

master or canonical version of the project. In fact that master version is often just called the *master branch*. There is a similar concept of a *tag*, which is a mark in the history of a code that represents a specific point in the history of commits that is important. This typically for organizational purposes, however, certain tags can be created that represent a whole set of code that can be identified with additional metadata.

The work of a project, edits to a codebase, merging of changes and the publishing of those changes, has developed under a number of models depending on the needs of the software project. The first can be described as a *centralized* workflow: where there is a shared repository and each individual user maintains a local copy of the centralized version and can push their changes back into the shared repository. The next one is an *integration-manager* workflow, where there is central repository that is considered to be the canonical one. Each person contributing has two copies, one is a public remote repository and the other is their private local one. Each contributor then makes changes and then publishes them to their personal public versions. A manager or editor then takes changes from each of those public remotes and then integrates them into the master centralized version. Once the changes are in the central repository they are then distributed back out to original contributors. The last workflow is known as a *dictator and lieutenants* workflow. This is where changes are gathered by the lieutenants and from many collaborators. Typically the lieutenants are divided along functional parts of a project. The larger set of changes, often a designated branch, and then are pulled in by the dictator who acts as the head of the project. All collaborators and lieutenants base their work on the remote through pulls and merges from the central repository.

GitHub is a web-based hosting service for projects using the Git source control management it provides services to host the central repository for a project. In addition to support for local tools for editing source code or content, the service has a web interface to handle integrating in changes from other user's local repositories. GitHub has an added wrinkle to a standard Git workflow by adding what GitHub calls *forking*. Forks are a bridge between a canonical repository and a local version being worked on by a user. The fork can be seen as a copy, but instead of being local onto a user's own system it acts as a remote repository being provided by GitHub of that same local version. Those changes are selected and then distributed back to the original repository by making a *pull request* to the maintainers of the original repository. The name comes because it is a request in the GitHub system, a notification or communication to the maintainers, who then evaluate the changes and then make a *pull*, which is a fetch of the forwarded changes and then a following *merge* of those changes into the project.

Metadata Directory Data Serialization

GitHub also provides a service that permits projects to produce a website based upon a branch of a project. This website can be a simple HTML one, or can utilize a variety of markup and templating languages to create a website. One of these markup languages is called YAML an acronym for YAML Ain't Markup Language. This website generator uses YAML as a language for defining variable data about a webpage. Combined with a language for templating and the stack of web technologies including HTML, CSS and JavaScript to create customized pages for each metadata standard and its associated implementations such as tools, extensions and use cases.

The core of the directory is the use of YAML to encode metadata on the metadata standards that are being presented in the directory. YAML is designed to be a human-friendly, encoding scheme for data. It reflects constructs such as scalars, arrays and associative arrays that are found in many programming languages. It is described as being "broadly useful for programming needs ranging from configuration files to Internet messaging to object persistence to data auditing." (Ben-Kiki, Evans, & Ingy döt Net, 2009).

YAML arranges data in scalars, lists or arrays, and associative arrays. Scalars are simple primitive values. Scalars are used with the other data structures and the structure of the document to record values and data. Lists or arrays would be sequences of data structures: these could be scalars or they include sequences of any of the other more complicated data structures. Associative arrays are sets of are key-value pairs; where the key is unique within the section or scope of the data that it is being applied to. This enables unambiguous description and access of the data being recorded and presented. Because YAML is designed to be processed easily by most programming languages, other serializations and mappings can be made with the data easily, an important consideration because of the requirement of portability of the directory data.

The advantage of using a lightweight markup format like YAML is that it provides a way of structuring data in a human readable way that should be instantly understandable by virtue of the structure of the document including: indentation, delimitation with punctuation, and the characterization and domains of the data recorded. In the prototype system, each attribute of a metadata standard is recorded as a key value pair. Attributes, which have a one-to-many relationship, are recorded as lists or

sequences, an example of this would be that a standard could have many subjects or disciplines it belongs too. Because of the nature of the YAML data serialization format complex data structures and relationships can be clearly expressed. For example, a metadata standard could have many related vocabularies which in turn those vocabularies can have a number of attributes such as a URL and a name. Similarly relationships between a metadata standard and its related implementations are maintained by recording that relationship as a key value pair in the related implementation as a foreign key to the related metadata standard.

Figure 1: Example YAML file

```
---
title: DDI - Data Documentation Initiative
name: ddi
layout: standard
type: standard
subjects:
  - general
specification_url:
http://www.ddialliance.org/Specification/
website: http://www.ddialliance.org/
related_vocabularies:
  - name: DDI Controlled Vocabularies
    url: http://www.ddialliance.org/
mappings:
  - name: DataCite Metadata Schema
    url: http://schema.datacite.org/
  - name: Dublin Core
    url:
http://www.ddialliance.org/resources/tools/dc
sponsors:
  - name: DDI Alliance
    url: http://www.ddialiance.org
update: 2009
version: DDI version 3.1
description: |
  A widely-used international standard for
  Describing data from the social, behavioral, and
  economic sciences. Expressed in XML, the DDI
  metadata specification supports the entire
  research data life cycle.
```

The final piece of the directory is a templating and site generation system that takes the structured data contained in individual YAML files and creates static HTML pages. Templates are usually HTML versions of the webpages which are systematically applied to the data for each metadata standard and in turn generate a complete HTML file with links, terms, graphics, and navigation all built into the website. This has the advantage of letting the design of the website be separated from the content. Because the templating system depends upon file structure which maps to appropriate sub directories in a website extending the structure of the website would then consist of adding additional files and folders as necessary. The GitHub service supports static website generation from a code repository, which the design of the directory takes advantage of.

`

Directory Management

Contributions to the prototype metadata directory are made using the GitHub service and a public Git repository hosted by the service. GitHub is a crucial part of the implementation supporting notifications and the creation of pull requests to the maintainers. There are a number of possible ways of making contributions: the first is an established workflow using the Git software on a local computer where each user has their own version of the project and they contribute using GitHub pull requests. The second would be to identify the contributors and have them share access to a repository. Lastly, as a lightweight option, possible contributors can use the online tools that GitHub provides to do changes completely within a web browser.

Fork and Pull

This contribution method would be the one that would involve installing Git on a local computer and using GitHub's features to make localized changes and then send them to the maintainers of the project. First, a contributor would need to make a fork within GitHub, which creates a copy of the project's canonical repository, but on the remote servers that GitHub provides. Then the contributor would need to make copy, *clone* in Git parlance, of their fork (their user-specific remote repository) onto their local machine: creating a local repository of the code or documents. That contributor makes contributions and edits: then commits those changes to the local repository. Those changes are then pushed (sent over the network) into the remote repository where those changes push the version that is current in the remote repository forward with those new changes. A *pull request* then is created and sent to the maintainers of the canonical

repository who then incorporate those changes from using the pull functionality of the project.

Shared Repository

A simpler alternative assuming that there is a smaller group of people interested in contributing to the repository would be to share access to the canonical repository. Each individual contributor could then contribute their changes directly to the canonical repository by cloning that repository, making changes and then pushing them into the remote repository hosted by the GitHub service.

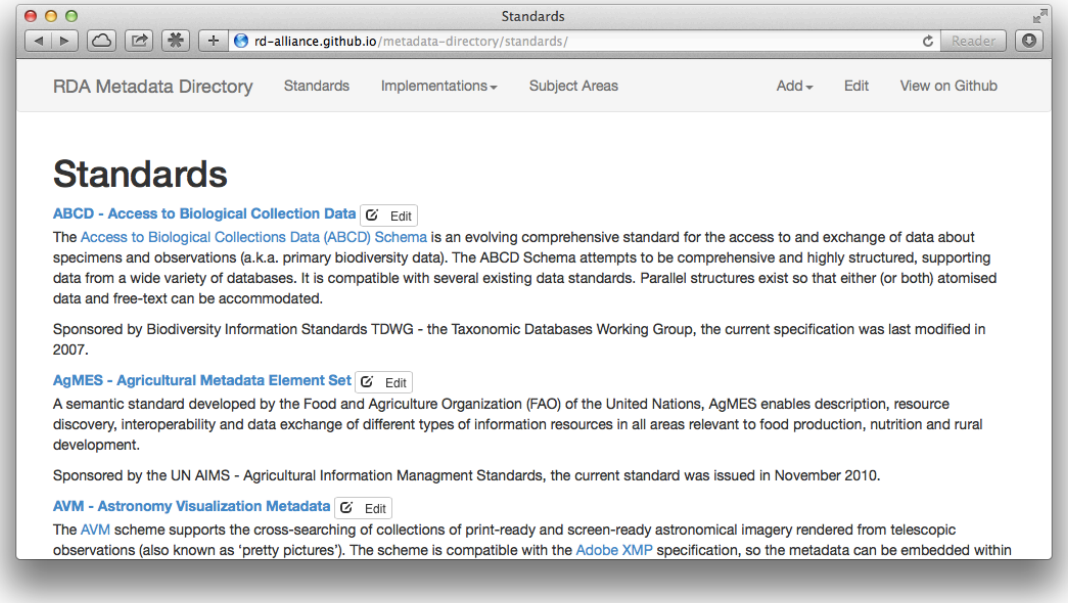
Online Editing

Finally, the service currently includes online editing of source files on a remote repository that would permit a contributor to make a contribution to a fork of the canonical repository, make online edits to files such as the description of a standard, commit those changes and then submit them as a pull request to the original maintainer. This would allow simple changes from a wide variety of possible contributors. However this method could also be combined with both the fork and pull and the shared repository models.

The design of the metadata directory includes links in a number of contexts to editing or adding of metadata standards and implementations. Contextual links would intuitively lead an interested contributor to edit a page while viewing the same page (see Figure 2). Similarly adding a page could be done from within a context or from a higher level if the contributor knows details of what they want to contribute. Editing interactively from within the browser would need a GitHub account; and would automatically fork the project to a personalized remote repository associated with the

contributor. They would make their contributions to their version of the metadata directory and then in turn would submit a pull request to the maintainers who in turn will incorporate the changes into the code that generates the directory.

Figure 2: Example Editing Context for Metadata Standards



Supporting interactive editing would be an important part of any collaborative project that manages a shared corpus of data. However, these considerations need to be balanced against access control and security considerations in an environment with a requirement of providing authoritative and high quality data. Similarly, many existing content management systems already have an editorial workflow that would be even more cumbersome if those sorts of roles and work were necessary. The advantage of using GitHub is that there is an existing workflow that can be repurposed to support an editorial workflow for the integration of content from a range of collaborators.

Extensibility

A primary requirement of the Metadata Directory is the capability of extending the functionality of the website in response to an unforseen need. Fortunately, the

architecture of the directory allows that to be done easily. Extending the directory would require a number of considerations to be made such as:

1. Identifying how the extension, new feature or data, relates to the unit of analysis, a metadata standard. For example, if there is a need to add a information about organizations that promulgate standards, establishing a new relationship or field to contain the establish the relationship as one where standards are governed or issued by these organizations is simply done by adding another field in the content YAML file or by creating a new entity.
2. Modeling whether or not that extension is a direct attribute of the metadata standard, or if it should rise to another entity within the existing scheme.
    a. If it is a direct attribute, including in our content structure for a metadata schema a new mapping or sequence that records the new attribute.
    b. If it is another entity, adding a new directory for the entity, recording the relevant attributes and then establishing relationships to metadata standards.
3. Developing templates for display that include the relevant attributes and relationships.
4. Evaluating the interest and resources of the community in providing the data and maintain the new data.

Adding additional functionality and providing an avenue for further features is an important requirement and represents a significant reason as to why the directory is being developed with an eye toward using the distributed version control system and its associated tools and applications.

Project Outcomes

The primary outcome of the project is that is a developed and deployed prototype located on the World Wide Web at http://rd-alliance.github.io/metadata-directory. The directory meets basic requirements to have an accessible list of metadata standards that can be edited by interested parties. It supports this functionality while providing powerful tools for version control, systems transparency, and documentation. It supports extensibility by being based on a simple scheme that is both human readable and can also support programmatic parsing for additional functionality. Finally, the system can be tested and deployed to a range of other platforms if necessary. The prototype can be evaluated and additional recommendations for its development and enhancement can be made, while being supported by the rich feature set and offered by the host platform.

Content from the Digital Curation Centre's disciplinary metadata standards directory was incorporated into the prototype. This included thirty-six (36) metadata standards, forty-two (42) extensions, ninety-four (94) use cases and fifty-three (53) tools. Content had been previously added as part of an information gathering project that resulted in the addition of eleven (11) standards, ten (10) extensions, twenty-three (23) use cases, and twelve (12) tools (Perez, 2013) . This content is summarized in the following table.

Table 5

Directory Content Additions

Type	From DCC	Perez 2013	Total
Standards	25	11	36
Extensions	32	10	42
Use cases	61	23	94
Tools	41	12	53

Content was mapped from the existing DCC website to a list of subject areas. Because the schema determined that metadata standards could be associated with more than one subject area our distribution of subject areas is greater than the total number of standards. The greatest number of standards is associated with the physical sciences and mathematics, while the life sciences comprise another large group. This is not surprising since many of the needs expressed in the Research Data Alliance's proposal are those of scientists in those disciplines. This content is divided into a number of subject areas as detailed in the following table.

Table 6

Subject Area Distribution in Directory

Subject Area	Count
General	8
Life Sciences	9
Physical Sciences & Mathematics	14
Arts and Humanities	3
Engineering	4
Social and Behavioral Sciences	5

Discussion

One of the principal requirements of the directory is to establish a platform for sustainable development and maintenance. A following section will discuss the issues with sustainably maintaining and developing the directory along with other risks with such an approach.

Sustainability

Ensuring that the directory is sustainably maintained and updated is a major requirement of the project. An important requirement of the project is the identification of persons or a group that can take the lead in maintaining the metadata directory and adding content to it. The inclusion of high quality information that is usable for others is a public good and one that would be hopefully rewarded by the rest of the research data community.

Ease of use is also an important consideration for the directory. Basing the usage patterns of the directory on existing user interface patterns and language affords users features which they may have encountered in other web applications and systems. Furthermore, taking advantage of affordances such as a simple *What You See is What You Get* editor will open up the possibilities of collaboration and contribution to a wider range of participants. Similarly, reusing a social pattern of contributions being reviewed with editorial guidance before being merged into the web site's content will bring a higher quality of information into the directory, thus improving its relevancy, accuracy, and understandability for the relevant communities.

Other parts of an effort to improve sustainability that are incorporated into the design of the website is the integration with an existing web service such as GitHub that

already provides many of the other features that support collaborative online communities. These features include a way of tracking issues or conversations about extensibility or problems of content in a manner that identifies participants along with contextual linking to the solutions or features that address those issues or requests. Other features include having a built in wiki that can be used for project documentation or to publish content that would fall outside the scope of the metadata directory directory, but are important guidance and information to support the community or people that are responsible for sustaining and maintain the directory. Lastly, by leveraging an existing service the prototype metadata directory can simplify many important services and features such as authentication, user account maintenance, rights management, permissions and hosting. In other words, the project can concentrate on improving and creating new content instead of the sundry of other things that are involved in deploying a website.

Risks

However there are other risks with the project plan, one of which is the reliance on someone or a group of interested people to contribute and collaborate on the project. Sustaining engagement and calling upon work from a range of people is always difficult for many projects. An association with a number of international efforts such the Research Data Alliance and DataOne, aiming to improve metadata and research data practice and services will help mitigate the risk of abandonment for the directory and will provide a ready pool of resources to help maintain and contribute to the directory.

Similarly there is a risk that the project would suffer from maliciously submitted content. Without expertise in every form of research data, it is conceivable that bad actors

could include information in the directory that is neither, accurate nor, neutral in presenting standards and their extensions with a fair editorial norm. However, this risk is shared my many open source projects that work on a massive collaborative scale. With a service like GitHub, there are many tools and policies built into the service, such as spam blocking, account limits, and reporting workflows, that scale these issues in a manner that is addressable even by small teams.

Engagement with the community is important in mitigating the problem of not having a number of people to contribute to the directory. Even though the number of contributors may be low to the directory, having the information be relevant to the wider community and toward will help make the directory successful. Online tools like issue tracking, discussion forums, and documentation will help clearly communicate the directory's architecture, policies, and content, to a wide-ranging audience, alleviating problems with having a limited number of contributing maintainers. Finally, project leadership is a quality that needs to be recognized in order for the directory to be a success.

Conclusion

A number of future directions for the prototype come to mind. These include (a) integration with existing Research Data Alliance (RDA) efforts, (b) additional testing and evaluation with a group of trusted people, (c) developing of new functionality, and (d) improvements in the usability of the application. All four of these efforts would address direct concerns mentioned in the Metadata Directory Working Group's project plan or represent best practice in the development of information search and retrieval applications.

The RDA is a wide-ranging effort that is attempting to address many points of difficulty in the research data management process. Bringing the directory prototype more fully into the work of the wider effort that represents the RDA would ensure the long-term sustainability of the project by increasing its value and visibility. The scope of this integration would need to be investigated more fully. Some possibilities could range from a loose association with linkages made to and from other RDA projects. Alternatively, a more tightly integration is possible where data from the directory is more directly embedded into applications and systems that are being developed in other parts of the RDA's efforts, engagement with other RDA members would represent a significant step toward addressing this further development.

Further possibilities for ongoing enhancements could be pursued by engaging with a trusted group of interested people who could provide additional information on the prototype. Additional training and documentation could be provided for this group and they could be charged with improving the content to the prototype. They could expand

the scope of the directory by including and gathering additional metadata standards and implementations that could improve the richness of the directory.

Additional functionality could be introduced into the directory. These could include many suggestions that were made at the Metadata Standards Directory Working Group's plan including:

- *Social media integration*. This can be across different dimensions, maybe allowing standards users to signal adoption or as contact and directory information.
- *Linked data support.* Export and implementation of a Linked Data service could permit reuse of the directory's data and support other applications using the data. Similarly, relationships to other resources could be discovered.
- *Communication tools.* Improvement of existing tools for managing issues, discussions and documentation could be made. Screen casting documentation could be made to help adopters.
- *Data model.* Improvements to the directory's data model could be made. This is supported by the extensible framework and could also include changes in the organization and structure of the directory to support novel uses.
- *Browse and search functionality.* Development of an improved browse and search interface that goes beyond other techniques for optimizing the search and retrieval of directory information in web search engines.

In closing, the projected has developed and deployed a working metadata directory that collaborators throughout the world can make meaningful contributions to. Earlier investigations have shown that there is a widespread community of research data

managers, curators, and users that are willing to submit data to such an application. Using a standard set of tools typically used in the development of software, the prototype leverages an existing pattern of content development and tracking that presents an advantage toward that of other systems such as a web content management system or a wiki. This prototype is built with extensibility in mind on a generalized scheme that is designed to present a low barrier to entry for users of all skill levels, while taking advantage powerful tools that help manage complex software development projects.

References

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLOS ONE*, *6*(9), e24357. doi:10.1371/journal.pone.0024357

Ball, A. (2009). *Scientific Data Application Profile Scoping Study*. Bath, UK. Retrieved from http://www.ukoln.ac.uk/projects/sdapss/

Ben-Kiki, O., Evans, C., & Ingy döt Net. (2009). YAML Ain't Markup Language (YAML™) Version 1.2. Retrieved February 27, 2014, from http://www.yaml.org/spec/1.2/spec.html

Borgman, C. L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. doi:10.1002/asi.22634

Chan, L. M., & Zeng, M. L. (2006). Metadata Interoperability and Standardization - A Study of Methodology Part I. *D-Lib Magazine*, *12*(6). doi:10.1045/june2006-chan

Church, J., & Gandal, N. (1992). Network Effects, Software Provision, and Standardization. *The Journal of Industrial Economics*, *40*(1), 85–103. Retrieved from http://www.jstor.org/stable/2950628

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, *12*, CIDCR1–CIDCR75. doi:10.2481/dsj.OSOM13-043

Conradi, R., & Westfechtel, B. (1998). Version Models for Software Configuration Management. *ACM Computing Surveys*, *30*(2), 232–282. doi:10.1145/280277.280280

Day, M. (2003). Integrating Metadata Schema Registries With Digital Preservation Systems to Support Interoperability: A Proposal. In *International Conference on Dublin Core and Metadata Applications (DC-2003)*. University of Bath. Retrieved from http://opus.bath.ac.uk/23599/1/101_paper38.pdf

Farrell, J., & Saloner, G. (1985). Standardization, Compatibility, and Innovation. *The RAND Journal of Economics*, *16*(1), 70–83. doi:10.2307/2555589

Glider, G. (1993). Metcalfe's Law and Legacy. *Forbes ASAP*, S158.

Greenberg, J., Jeffery, K., & Koskela, R. (2013). *Case Statement Proposal: Metadata Standards Directory (MASDIR) Working Group*.

Hillmann, D. I., Marker, R., & Brady, C. (2008). Metadata Standards and Applications. *The Serials Librarian*, *54*(1-2), 7–21. doi:10.1080/03615260801973364

Hillmann, D. I., Sutton, S. A., Phipps, J., & Laundry, R. (2006). A Metadata Registry From Vocabularies UP: The NSDL Registry Project. In *Proceedings of the International Conference on Dublin Core and Metadata Applications* (p. 9). Digital Libraries, Dublin, OH: Dublin Core Metadata Initiative. Retrieved from http://arxiv.org/abs/cs.DL/0605111

ISO, & IEC. (2004). *Information technology—Metadata Registries (MDR). Framework* (2nd ed., Vol. 2004). Geneva: ISO/IEC. Retrieved from http://standards.iso.org/ittf/PubliclyAvailableStandards/c035343_ISO_IEC_11179-1_2004(E).zip

Kriesberg, A., Frank, R., Faniel, I., & Yakel, E. (2013). The Role of Data Reuse in the Apprenticeship Process. In *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*. Retrieved from http://www.webjunction.org/content/dam/research/publications/library/2013/faniel-data-reuse-apprenticeship.pdf

O'Sullivan, B. (2009). Making Sense of Revision-Control Systems. *Communications of the ACM*, *52*(9), 56. doi:10.1145/1562164.1562183

Perez, C. (2013). *The RDA's Metadata Standards Directory: Information Gathering*. University of North Carolina at Chapel Hill. Retrieved from https://cdr.lib.unc.edu/record/uuid:60ae1a94-dd39-411e-8127-ca656ec7c29c

Qin, J., & Small, R. (2008). The Science Data Literacy Project. Retrieved February 28, 2014, from http://sdl.syr.edu/?page_id=32

Research Data Alliance. (2013). *Metadata Standards Directory (MASDIR) Working Group*. Retrieved from https://rd-alliance.org/filedepot?cid=95&fid=73

Rogers, E. (2003). *Diffusion of Innovations*. New York: Free Press.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, *6*(6), e21101. doi:10.1371/journal.pone.0021101

Wang, W.-T., & Wei, Z.-H. (2011). Knowledge Sharing in Wiki Communities: An Empirical Study. *Online Information Review*, *35*(5), 799–820. doi:10.1108/14684521111176516

Weitzel, T., Beimborn, D., & König, W. (2006). A Unified Economic Model of Standard Diffusion: The Impact of Standardization Cost, Network Effects, and Network Topology. *MIS Quarterly*, *30*, 489–514. Retrieved from http://www.jstor.org/stable/25148770

Willis, C., Greenberg, J., & White, H. (2012). Analysis and Synthesis of Metadata Goals for Scientific Sata. *Journal of the American Society for Information Science and Technology*, *63*(8), 1505–1520. doi:10.1002/asi.22683

Yates, D., Wagner, C., & Majchrzak, A. (2009). Factors Affecting Shapers of Organizational Wikis. *Journal of the American Society for Information Science and Technology*, *61*(3). doi:10.1002/asi.21266

Appendix

Figure A1: Example Metadata Standard Presentation

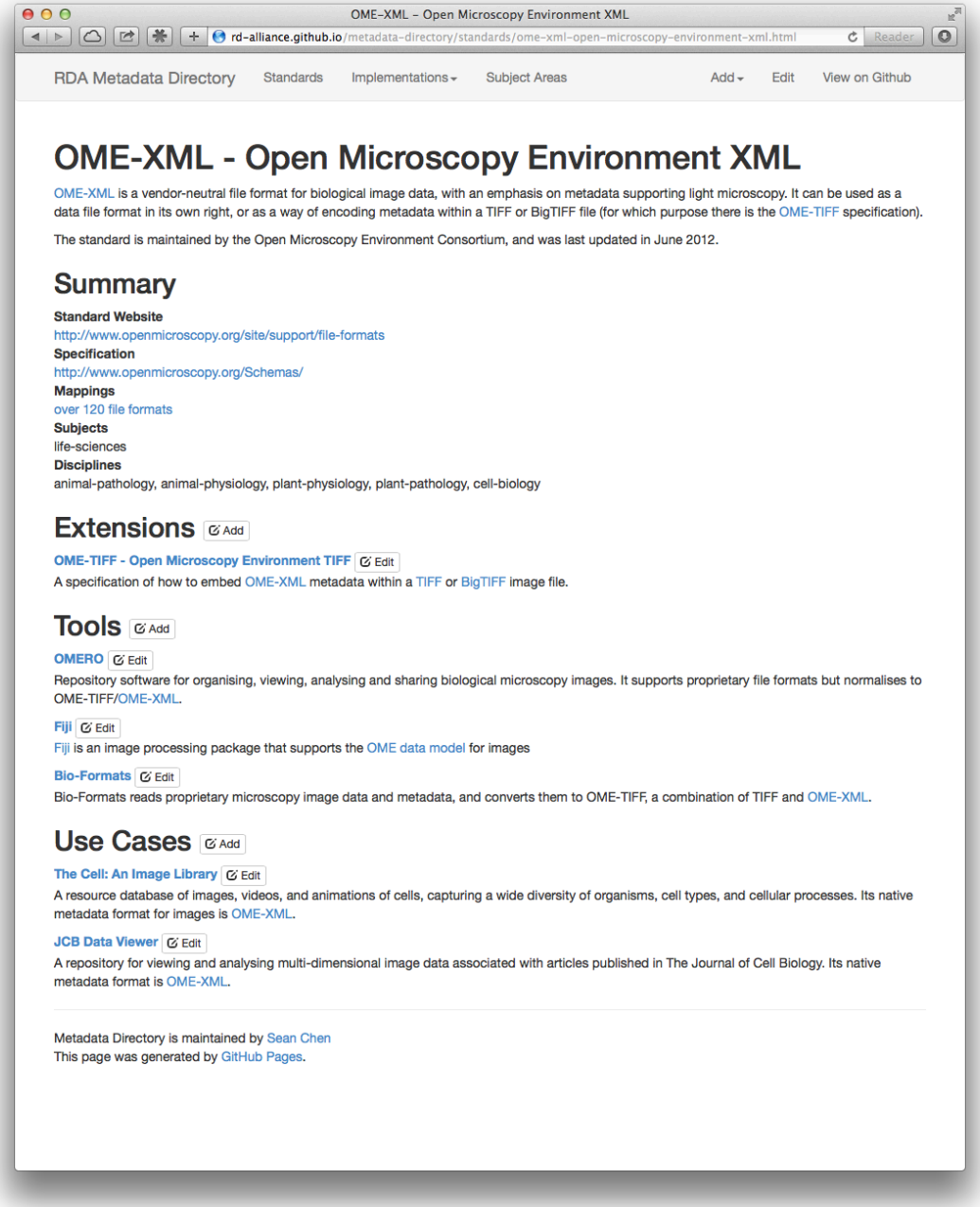


Figure A2: Subject Area Organization