Project SIMILE: Semantic Interoperability of Metadata and Information in unLike Environments

Massachusetts Institute of Technology
July, 2005

Summary

The MIT Libraries are requesting funding from the Andrew W. Mellon Foundation for the sum of $1.5 million for a two-year project for a project on advanced development of prototype tools for and demonstrators of the Semantic Web for a range of large-scale educational data interoperability problems, and to integrate that work into the DSpace platform and other educational environments where interoperability of heterogeneous data and metadata is needed. This work, which will be conducted in collaboration with MIT's Computer Science and Artificial Intelligence Lab (CSAIL) and the World Wide Web Consortium (W3C), will be of immediate benefit to prior Mellon-funded MIT initiatives such as DSpace, OCW and OKI, as well as several other Mellon-funded initiatives, including ARTstor, JStor, Sakai, Fedora, and VUE. It will additionally improve understanding of the benefits and possibilities of Semantic Web technology in the general higher education community.

Introduction

Libraries and other managers of scholarly information face significant challenges in coping with the increasing amounts of digital material that they are acquiring from digital publishers and from members of their organization (e.g. the faculty and researchers of academic institutions). These digital materials require new methods for efficient and affordable long-term stewardship in order to cope with the scale, complexity, and heterogeneity of this digital data.

Libraries and other digital content aggregators must consider how these materials will be acquired, either from publishers or faculty, and deposited into managed repositories, as well as how to handle their long-term storage, management, and preservation. They must arrange for an array of managed digital services including indexing, search, and access as well as storage and life cycle management. Institutions need digital repository solutions that span these needs, allowing them to offer services for digital resources with no less functionality then they have offered in the past for physical resources.

DSpace is an institutional digital repository platform developed by the Massachusetts Institute of Technology Libraries and Hewlett-Packard Laboratories. Faculty and researchers or their proxies (e.g. library staff) can submit their digital research output: research papers, image collections, datasets, audio, video, etc., directly into their institution's DSpace for widespread networked access and long-term stewardship by the library.

In 2002 MIT and HP made the DSpace software available via an open-source, BSD-style license and it is now in use at over 100 research-generating universities and cultural

heritage organizations world wide (see http://wiki.dspace.org/DspaceInstances/ for the current list of registered sites). This community of organizations is collaborating to achieve a common vision of federation among DSpace repositories, and with similar organizations that run repositories built on other platforms (e.g. Fedora or EPrints repositories) and running services of interest to the research community (e.g. Google Scholar, A9).

Delivering on the vision of federated digital repositories that support seamless, location-independent discovery and retrieval of relevant research material requires advances in interoperability among the digital resources under institutional management – however and wherever they are physically hosted - and between those resources and the services that consume and/or produce descriptions, annotations, and additional resources. Such interoperability has been very difficult to achieve so far in ways that are both highly scalable and technically supportable by the library and other digital content management communities.

To advance solutions to this problem, the MIT Libraries are working together with the World-Wide Web Consortium (W3C) and the MIT Computer Science and Artificial Intelligence Lab (CSAIL) on the SIMILE project. SIMILE (Semantic Interoperability of Metadata in unLike Environments) is working to enable highly scalable interoperability among digital content, the metadata that describes that content, and the services that allow people to work with that content. SIMILE

SIMILE began two years ago with research funding from Hewlett Packard, and has already completed work to perform studies, capture representative data collections, and create tools that demonstrate the value of the Semantic Web to the problems of digital data interoperability for a variety of services. We wish to further develop these tools and technologies into production quality and to add new tools in this area for which the community has expressed a need.

Furthermore, SIMILE will further leverage and extend DSpace, enhancing its support for arbitrary schemas and metadata, primarily though the application of RDF and Semantic Web technology. To guide the SIMILE effort we will focus on well-defined, real-world use cases in the library and educational technology domains. Because of the widespread adoption of the DSpace platform by research libraries, we believe that there will be a powerful deployment channel through which the benefits offered by Semantic Web technology to improve data interoperability at Web scale can be compellingly demonstrated in a visible and global community.

Project Description

We propose a project that will create compelling use case demonstrations and prototypes at the intersection of community or individual information creation and management, institutional information and digital collections management, and the Semantic Web data architecture.

We seek to enable low-cost, scalable interoperability among digital research collections, the metadata that describes them, and the services that leverage that metadata and digital material. The project plan includes extending the widely used open-source DSpace digital repository platform currently available at http://dspace.org/. We will extend DSpace to greatly enhance its ability to provide easy-to-use support for arbitrary schemas and metadata, primarily through the use of RDF and Semantic Web techniques, and explore support for distributed research collections. The project will ground its work in focused, well-defined, real-world use cases in the domains of libraries and higher education. We will also collaborate with other significant projects that operate in these domains, (e.g. ARTstor, JSTOR, OCW, Fedora, Sakai, and VUE) to improve understanding and use of Semantic Web technologies to provide scalable data interoperability within and across those platforms

Project Context

The Library Information Management Domain

One of the core competencies of libraries has traditionally been support for the research activities of scholarly information consumers such as university faculty and students. Research activities typically include searching and browsing metadata to find relevant information in many formats including books, journals, maps, visual images, archives, audio/visual material, datasets, computer files, and so on. In supporting search and information retrieval activities, libraries have evolved over time from isolated and eccentric modes of describing their collections and providing access to them towards international standards which support the broad portability of research methods across libraries and collections. Examples of these modern standards in common use are the ISBD (International Standard Book Description), the ISAD-G (International Standard Archival Description), MARC (Machine Readable Cataloging standard), and Z39.50 (a standardized search and retrieval protocol). This degree of standardization across libraries of all types has allowed academic researchers world wide to achieve good efficiency and effectiveness regardless of physical location or type of material sought.

However as scholarly information consumers have become increasingly familiar with and dependent on computer-based techniques for locating and accessing information, especially since the advent of the Web, libraries find themselves under growing pressure to support more flexible and powerful information seeking activities. This demand manifests itself in two ways: the great success of web search engines like Google for doing quick, high recall searches to identify large amounts of potentially useful information, and the demand for richer and more domain-specific description and search

support than was possible with traditional publishing and library standards such as ISBD and MARC. No longer satisfied with compromising accuracy to achieve portability of research methods, researchers now want to have the best of both worlds.

At the same time, managers of information collections that were never well served by traditional library standards (e.g. archival finding aids, scientific project data, and rich digital image descriptions) are becoming empowered to find better solutions for describing and searching their information through constant technological improvements. This is a very useful development for users of this type of previously unsupported information, but is leading to the creation of metadata gulags that are isolated from mainstream library-based search systems thus hiding them from many potential information seekers.

The need to be able to support a wide variety of types and sources of metadata, to integrate them effectively, and to expose them to simple, flexible search and retrieval tools for researchers has become a major challenge for libraries in the Web era. And libraries are an analog of other information-intensive enterprises in that the challenges and opportunities they and their patrons face mirror those faced by cultural heritage organizations, government agencies, and information-intensive businesses around the world.

Current library practices are stressed in several dimensions:

- Historically, libraries have optimized access and discovery of "locally accessible" assets. The pragmatic meaning of "locally accessible" has shifted in the era of networked resources due to the increasing dependence of university-based researchers on licensed digital collections such as ARTstor and JSTOR, driving demand for interoperability across these collections.
- It is too expensive for libraries to support non-standard, community-specific ways to describe or annotate content themselves.
- Even if libraries could afford this, community-specific information is held in distinct information silos or catalogs and does not integrate well with more traditional or generic library catalogues. For example, one can search a library's catalog and the ARTstor image library, but not both at once
- It is difficult for communities to describe and annotate their own digital collections with their own terms in a way that other users can access through the libraries centrally-supported systems. They are effectively invisible.
- Decisions about who should do what are very hard to change. Information initially created and managed by members of the community are difficult to hand off to the library for ongoing maintenance, and vice-versa.

The Semantic Web

Along with email, the Web has become the dominant communication medium for scholars today, and is the primary mode for sharing research and teaching material of all kinds – both formally published and informally posted. Scholars use the Web to find

things, get things, and increasing to use tools that support their research and teaching activities (e.g. Web-based course management systems, digital libraries, e-journals, etc.).

But interoperability on the web today requires too much human mediation. The World Wide Web Consortium, who originally created, and now have responsibility for maintaining and evolving the standards that make the Web work, has offered the following vision for an emergent ext-generation Web, known as the "Semantic Web":

> "The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is the idea of having data on the Web defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications. The Web can reach its full potential if it becomes a place where data can be shared and processed by automated tools as well as by people."[1]

Their vision of where the Web is going aligns very well with our problems in the information technology areas of interoperability across very large, very heterogeneous scholarly data collections. Information is highly distributed, and will remain so, but *must* become more interoperable without reliance on expensive and un-scalable technologies (for example, broadcast searching where a single search is issued to multiple remote catalogs with the results being collated together for the user; this works well for small numbers of catalogs but not at the scale of all relevant catalogs that are Web-accessible).

Challenges and Opportunities

Through the SIMILE project we would like to demonstrate how the current barriers to interoperability across digital libraries can be reduced, thereby creating useful tools that span the individual, the community, and institutional information management roles and requirements, at low cost and with high scalability. To ensure that our work remains grounded, we will conduct the work using libraries and higher education as our source of material, and define the requirements of faculty and students for the technologies they will find sufficiently useful to adopt. This section describes the domain-specific challenges and opportunities in doing so.

We would like to be able to support *arbitrary* metadata for digital library collections in addition to metadata that is based on library standards. These should include very rich and complex metadata such as that used in digital libraries like ARTstor (i.e. the Visual Resource Association (VRA) Core metadata) or OCW (i.e. the IEEE Learning Object Metadata standard). Such simultaneous support has not been possible previously without expensive conversion of the metadata into a common standard such as the library's MARC cataloging standard previously mentioned.

We would like to be able to support the flexible *evolution* of subject-specific metadata. In the past, libraries could control the evolution of their standards to control and limit the

[1] W3C Semantic Web Activity Statement, <http://www.w3.org/2001/sw/Activity>

impact of changes on their practices and systems. But subject-specific metadata (often faculty produced) changes very rapidly and this should be supported since it allows the metadata to reflect the real thinking of the subject experts rather than arbitrary abstractions of it that the library favors.

We would like to be able to support arbitrary, ad hoc *annotations* to digital collections. These might be supplied by the information consumers (e.g. students and faculty) directly, or by external subject experts such as other researchers in the field, or by professional collections managers such as librarians, or even by automated techniques such as collection data mining that discover new connections (e.g. amazon.com's "readers who purchased this book also liked…").

We need to be able to work with an increasing number of *third party service providers*, both of digital collections and of metadata about those collections. Traditionally libraries had exclusive control of the descriptions of their collections, but it is becoming increasingly true that useful metadata is created outside libraries: either by authors themselves or by service bureaus such as publishers or production labs. This is true for all types of metadata, and such metadata should be kept and leveraged whenever possible.

The main opportunity we have with the outcomes of this project is to provide strong subject-appropriate data support while still offering the professional management and services of the institutional library. Ever increasing amounts of digital content are coming from outside the library walls, so that libraries must manage multiple kinds of content with widely varying description. They must also manage interactions and relationships with multiple communities of interest for the same material (i.e. interdisciplinary studies). The kinds of material under management, the fields of study that produced them, and the ways they all have of talking about the materials are not static. They change with time as the communities themselves develop, ebb, and flow.

Another important opportunity is to provide a unified, customizable search interface to all digital library collections that is easy to use. In the new digital domain, libraries would ideally offer "one stop shopping" to all information resources of interest to its constituents, and act as a flexible information clearinghouses. Such technological tools could provide access to the richness of subject-specific metadata while also providing good general interdisciplinary recall across the whole digital library. The interface would provide access to items held "within" the library, as well as to external collections to which the library has negotiated access, and would be configurable by individuals or subjects to emphasize and optimize the presentation of the material that are important to the particular community.

Of course, the advent of search services like Google have changed the role of libraries in serving as the sole source of discovery tools, but we also have good opportunities to work

with and influence Google and other search service providers who have demonstrated an interest in the potential of the Semantic Web data architecture for their own work.[2]

Finally, there is an opportunity to use these technologies to allow the library to become an important source of *new* information about digital collections to aid faculty and students in realizing the utility of digital resources, in addition to simply providing access to those resources. In other words, in addition to simply serving resources, libraries might become data sources themselves, for example by offering recommender services based on individual and/or collective patterns of use and interest in digital collections.

Summary

The proceeding sections discussed the problems of information management, search, and navigation in the library and higher education domain, and how scalable interoperability of data and metadata is of utmost importance to get the full benefit of the large digital collections we depend on. The Semantic Web was proposed as a possible solution to this problem of data interoperability; a way to reduce the cost and increase the scalability of data interoperation. However exactly *how* this should be done remains an unsolved problem – aspects of it are well understood but not yet developed, others require further exploration and testing.

The following section describes the actual work of the project and a set of specific deliverables of the project, centered on very large, heterogeneous digital data collections. SIMILE can provide an effective filter for translating the research of the Semantic Web community over the past several years into practice for the higher education domain. The work of this project will build on prior work of SIMILE, of the digital library community, and of the DSpace community.

[2] The recently announced Yahoo "My Web 2.0" which is a type of "social search" system will support the Semantic Web data model natively, thus enabling a range of added-value services such as community knowledge bases

SIMILE Project Deliverables

The SIMILE project has been underway with separate funding for the past two years, and many aspects of the project are well established and have demonstrated useful results. In particular, the project's core development team is already in place, although we require extension for the next two years to accomplish the deliverables defined below. The work of the project has already attracted international support as well as collaboration on several issues (for example, see the discussion below on recent work to develop the *Fresnel* ontology for displaying RDF on the Web, involving contributions from several international Semantic Web experts working together with the local SIMILE and Haystack teams).

The work of this project over the next two years will include significant further development of a set of prototype and production tools for working with RDF-encoded data from the higher education and library domains and extensions to the DSpace repository system to enhance its support for arbitrary metadata and data schemas. The deliverables fall into six categories: (1) data acquisition, (2) tools for domain experts, (3) tools for data authoring, (4) data storage infrastructure, (5) tools for end-users, and (6) a SIMILE/DSpace metadata engine (see figure 1). Together these categories of work will create a framework of Semantic Web tools and infrastructure that will provide the full functionality to demonstrate the value of this technology, in both prototypes and production systems. We will work throughout the project to assist other projects in the digital library and educational technology domains in using Semantic Web technology and incorporating SIMILE tools wherever appropriate.

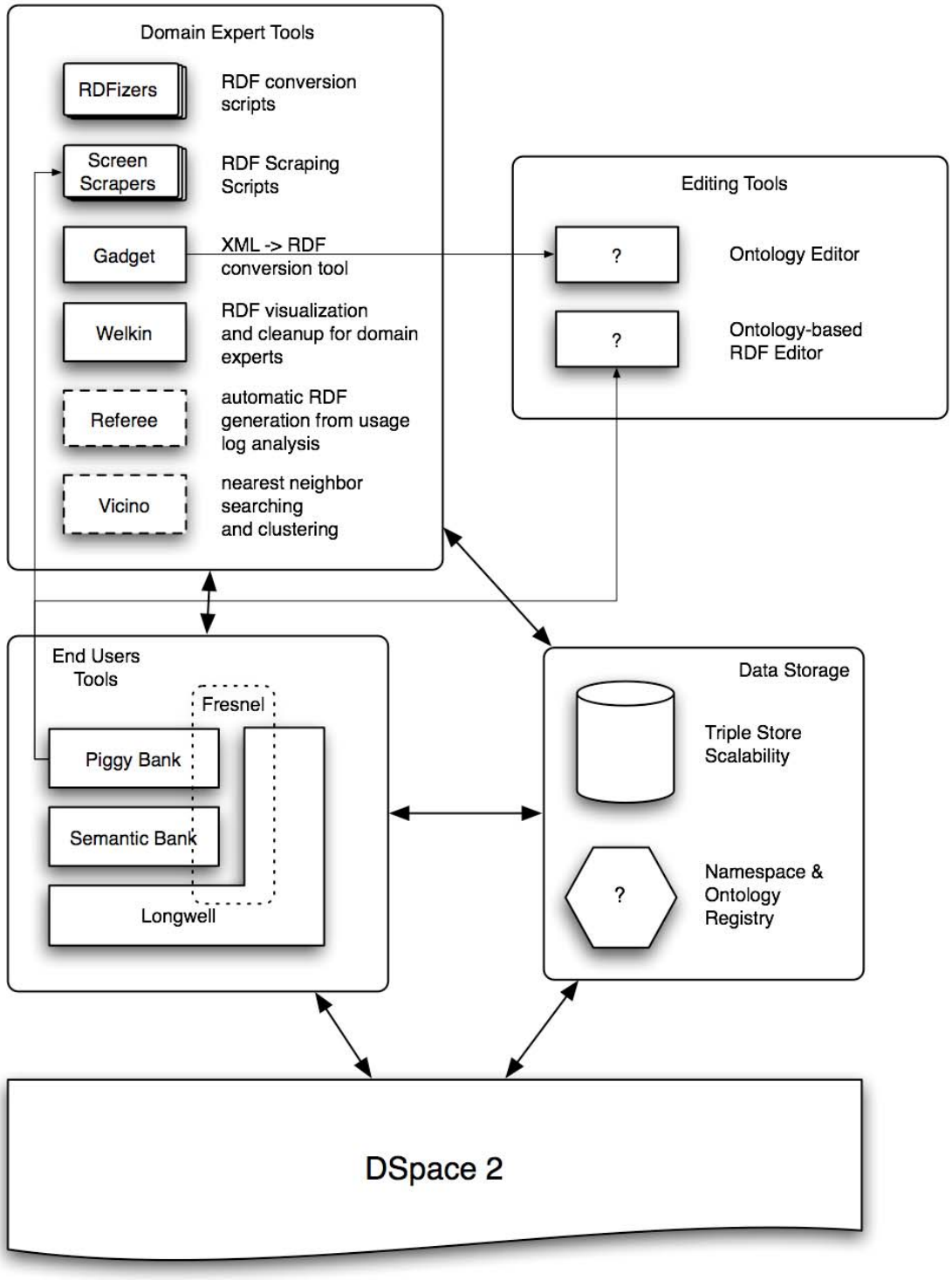

Figure 1 – SIMILE work-in-progress shown in the overall development framework

The SIMILE project methodology is an advanced development process that starts with well-defined problems and existing data collections, develops prototype tools and interfaces that demonstrate how RDF and Semantic Web technologies can help with those problems using the collected data, then refine the prototypes based on feedback from users. We also work with the user community to identify additional tools that would be helpful and design prototype solutions. When a tool or interface cannot be developed due to missing or inadequate standards in the Semantic Web domain, we work with our W3C colleagues to create or fix those standards (e.g. *Fresnel* http://simile.mit.edu/fresnel/ which will become a W3C note when it is complete). Our goal for SIMILE is to produce working, operational prototype tools and a development framework which other institutions, including MIT, can easily develop into operational systems. This project seeks to complete this methodology for work we've done to-date, and for further work that we have identified, for digital content in the higher education and cultural heritage domains.

MIT will implement a subset of these tools and interfaces in its production DSpace repository and will make that functionality available to the entire DSpace community at the end of the project. All prototypes and tools will be released under the BSD open source software license, in keeping with DSpace and SIMILE past practice. Reports will be published on the SIMILE website under a Creative Commons license and also deposited in DSpace@MIT. SIMILE is not a software company, so we have worked from the outset to create a developer community around these tools that can operationalize, maintain, and support them in the future. The SIMILE team has a great deal of experience building such communities and at delivering software at the right stage of development to insure that it is adopted by the community in a sustainable manner.

There are twenty six deliverables described in the sections below. They vary in the level of our understanding of the approach needed for implementation and the complexity of that implementation. We use a star rating system to indicate these two variables for each deliverable: the first star is for our understanding of the approach (i.e. how much further research and analysis will be necessary to begin development), and the second star is for level of difficulty of the actual development. The fewer the stars, the simpler and more straightforward the task. Each deliverable is described in some detail and is also listed on the timeline with an indication of who will be assigned to work on each of them, at what point in the project, and with what dependencies.

Category 1: Data Acquisition

In order to understand and demonstrate how Semantic Web technology can add value to computer systems being developed in the higher education and cultural heritage domains, it is first necessary to capture a significant corpus, or test bed, of relevant data. These include large, representative data collections that have potential cross-relationships which RDF and the Semantic Web can draw out and expose to end users. This category of work involves identifying those key data collections, acquiring them from their owners (with all necessary legal arrangements), and integrating them into the SIMILE test bed.

Prior Work

A large SIMILE test collection has already been accumulated by gathering important datasets that were independently developed by different communities with different schema in the higher education and cultural heritage domains. They include:

- ARTstor. 100,000 metadata records encoded with the VRA Core (XML) schema and associated thumbnail images, originally created at UCSD Libraries and contributed to ARTstor. SIMILE has a legal license with ARTstor to use this data under strict access-control for demonstration purposes only.
- VIA. SIMILE acquired Harvard's VIA collection (fine arts images described with two variants of the VRA Core XML metadata schema) which can be made publicly available.
- OCW. MIT's OpenCourseWare provided a small sample of course websites containing education material of many kinds described with the IEEE Learning Object Metadata (XML) schema.
- JSTOR. A small sample of metadata records for scholarly e-journals in the art and architecture domains, encoded in the TEI header (XML) schema. SIMILE has a legal license with JSTOR for this data, similar to the one negotiated with ARTstor and with similar access constraints.
- W3C Technical Reports. 800 records for technical reports written by staff or members of the W3C and described with a locally defined metadata schema which was already RDF conformant.
- DSpace collections from more than 75 institutions that are using the DSpace software and described with Dublin Core XML metadata. The material consists mainly of scholarly journal articles, white papers, technical reports, or theses.

In addition to these content description data collections, a number of additional "linkage" datasets have been acquired to enhance and extend these data collections by, for example, building linkages from vocabulary used in one schema or domain to another (using OWL equivalency statements). These include:

- OCLC Name Authority File. OCLC provides a Web Service to identify record linkages across variations of personal and corporate names, based on a database of more then 5 million name authority records. This service was developed specifically for DSpace and similar metadata-intensive systems, and OCLC is collaborating with SIMILE to test this functionality with RDF-encoded data (see category 5 on DSpace development for more information)
- Library of Congress Thesaurus for Graphic Materials (a large set of terms for subject indexing of pictorial materials, particularly the large general collections of historical images which are found in many libraries, historical societies, archives, and museums).
- CIA World Factbook. A website created and maintained by the US Central Intelligence Agency containing detailed information about every country

in the world. This provides contextual information when linked to scholarly content from an identifiable place.

Techniques used for data preparation of all these collections have been documented and the RDF ontologies created for them are available to the data owners and the public on the SIMILE website.

Future Deliverables

1a) Identify and acquire additional data collections from the higher education and cultural heritage domains to further demonstrate the value of the SIMILE tools to collection owners and to information seekers (*)

Candidate collections which we have begun negotiations to acquire are:

- ICPSR. Mary Vardigan, the Assistant Director of the ICPSR, has expressed an interest in working with SIMILE to explore how SIMILE tools could provide better data processing and end-user navigation features.
- JSTOR. David Yakimischak, the former CTO of JSTOR, was working with SIMILE to provide data for journals in the art and architecture area. We will work with current JSTOR staff to expand this collection with metadata for journals from other domains, possibly including their entire collection.
- Archival collections from Fedora repositories of the University of Virginia and Tufts University. These repositories include large image collections and very large, faculty-authored websites.

1b) Identify and acquire more "connector" collections to demonstrate the value of inferencing across the content collections (*)

Collections under negotiation to acquire are:

- The Getty Foundations's thesauri
    - Art and Architecture Thesaurus (AAT).
    - Union List of Artist Names (ULAN)
    - Thesaurus of Geographic Names (TGN)

  We have begun communication with Murtha Baca and Karim Boughida at the Getty Foundation to negotiate acquisition of their thesauri. These will be converted into RDF and added to the SIMILE data test bed. The Getty Foundation staff is very supportive of this work.
- Relevant GIS datasets held by the MIT Libraries, to test georeferencing across the content data collections.

1c) Identify and acquire or create additional non-descriptive metadata (e.g., administrative and technical information) about digital assets under local control for collection management (**)

So far, SIMILE tools and interfaces have focused primarily on descriptive metadata for the test bed collections. But many other types of metadata exist for these collections,

including administrative (provenance, rights, technical) metadata and structural (component parts and relationship) metadata. These other types of metadata have enormous potential for leveraging RDF and Semantic Web technology to improve collection and data management operations. Fedora, for example, is using RDF successfully to capture structural metadata and relationships among the complex digital objects stored in that repository system, and other content collections (e.g. OpenCourseWare courses, Sakai learning objects, and DSpace repository items) could similarly benefit from converting such metadata into RDF. As an initial step, a test bed of non-descriptive metadata will be acquired or created in order to define prototype tools for collection managers to edit, search, and navigate that data. At a minimum we will capture this data from OCW and DSpace repositories, but we will also explore what other types of data exist in the other repositories we are working with (i.e. ARTstor and JSTOR, Fedora sites at University of Virginia and Tufts)

1d) Redevelopment of the existing set of RDF translation tools and procedures (*)

First generation programs to convert legacy (non-RDF) data into RDF ontologies involved experimentation that requires further refinement both for the appropriateness of the RDF ontologies and for the efficiency of the conversion scripts. This set of tools includes RDF proxy servers (to traditional HTML-based websites, including DSpace repositories), XSLT scripts, and client-side tools to convert XML/HTML into RDF such as the GRDDL (http://www.w3.org/2004/01/rdxh/spec) functionality in Piggy Bank that automatically and dynamically converts DSpace metadata into RDF. Each RDF ontology developed to-date will be re-examined in light of early prototype feedback, and conversion scripts will be re-factored or re-implemented as necessary.

Category 2: Domain Expert Tools

Leveraging the Semantic Web depends on data and metadata being represented in the W3C's Resource Description Framework (RDF, usually encoded in XML). In order to develop useful tools for processing, searching, browsing, navigation and display, it is currently necessary to create RDF conformant data from native formats of the data, for example, database exports, MARC or other binary formatted records, XML or HTML documents, TeX documents, etc. And production of RDF-encoded data will be necessary to do without requiring the data originator to provide their data already in RDF. The data should, in some cases, also be possible to convert back into its original data format for further processing.

Therefore this set of deliverables includes creation of a set of tools, scripts, and methodologies to translate legacy data into RDF (e.g. translators, proxies, or wrappers of existing content collections) in a form suitable for use by librarians to upload metadata from XML and relational databases. If and when RDF becomes the accepted and dominant data architecture this will no longer be necessary, but that is not the case today or for the foreseeable future.

Prior Work

This category includes the creation of tools (e.g. *Welkin, Gadget*) to support metadata experts such as librarians or other domain experts to explore and understand legacy data collections in order to design RDF ontologies for them or perform other types of analysis that adds value to the data. These sorts of data exploration and ontology-building tools for data experts will be critical for the scalability of Semantic Web approaches to data management and have been specifically identified by ARTstor and other content aggregators as the sort of productivity tools they need and have difficulty designing today, in the XML era. This is an area that SIMILE has not devoted significant effort to in the past, but based on the feedback we have received from early adopters of our tools, this is an area that deserves significant further analysis and development.

Future Deliverables

2a) Design interface for and develop interactive tool for use by domain/community experts to search for and locate pre-existing ontologies (*)

Conversion of legacy data into RDF requires the identification of a suitable *RDF Ontology* for the particular data structure in hand. Identifying existing RDF ontologies that can be reused saves a great deal of time and effort, so that a tool to search for existing ontologies and display them to a data expert would be highly desirable. The domain expert could then select from one or more existing ontologies in mapping the legacy dataset, and proceed with exploring that data, now encoded with RDF, using tools to be developed in other deliverables. The tool should help the expert search and discover existing ontologies as well as make informed decisions about which ones are better.

2b) Design interfaces and develop interactive tools for use by domain experts for legacy data exploration and conversion into RDF (**)

When faced with the task of transforming with legacy datasets in RDF, we found two main obstacles: one is the understanding of the structure of the dataset (the schemas and the meaning that the schema authors intended for the various fields, relations and data values), another (less evident at first) was the understanding of the data itself and how the dataset makes use of the structure of the data in actual practice. While in theory the understanding of a schema should yield enough information to understand the entire dataset, we found this not to be the case. 'Inspection tools' (e.g. Gadget) that are able to perform fast and effective structure and data mining over legacy datasets (mostly from XML repositories or relational databases) are critical to help and guide domain experts during a conversion and/or merging process. We propose to extend and refine Gadget based on our early experiences and test it with data managers who are exploring legacy data (i.e. DSpace@MIT, ARTstor, VUE, etc.).

2c) Develop interactive and batch tools for increasing the utility and quality of RDF datasets (***)

Once data is normalized in RDF, it becomes easily mixable. While this is one of the main benefits of this technology, we also realized how metadata quality tends to degrade with the number of merged datasets. The reason for this is that controlled vocabularies and identifying schemes are locally uniform, but not globally, leading to data of very mixed quality.

This drives the need for a tool to help domain experts draw equivalences and/or augment and adjust metadata to provide the symbolic linkage across incoming data that encodes in a machine-processable way the information that is easily understood by humans (for example, that dc:creator and vra:author are semantically equivalent, or that "Pablo Picasso" and "Picasso, Pablo" are the same person).

The functionality of this tool will include:

- resolving data ambiguities, for example by linking records (e.g., techniques like name authority control)
- suggesting additional metadata extracted from textual descriptions in the data
- using Latent Semantic Indexing and other techniques to suggest relationships between ontological terms
- clustering with near neighbor methods to provide hints for domain experts on possible connections and/or mistakes

It is possible that not all of these functions will prove useful to domain data experts in practice, so prototype versions of the tool can be shown to experts as features are added to get feedback and inform further development.

2d) Design and develop tool for monitoring application usage data in order to prepare suggestions for additional metadata to be reviewed by domain experts (*)

High quality metadata authoring is a very expensive task, but clever and focused analysis of users' behavior might lead to the automatic emergence of metadata.
Examples of such data include:

- referrer logs (pages where users came from) to capture comments by reader (e.g., from blogs) about the data
- user trails (subsequent pages visited by the same users) to indicate relationships between data (e.g., buying trails for "users who read this also read…")
- querying patterns to indicate relative importance of data in the collection (e.g., to indicate future collection policies, or interface design)

We will create a tool to test this functionality, extending the "Referee" library which is currently in the early prototype stage of development. Future work will create a working Web application that will present findings to a data expert for approval to add to the collection.

Category 3: Data Authoring

Prior Work

SIMILE has done no significant work in this category so far but has identified it as a significant and long-standing need to help people make the transition to RDF data modeling over current techniques such as database schemas and XML schemas. It builds on some of the tools that were previously prototyped in section 2 (i.e., the domain expert tools for exploring legacy data and existing RDF ontologies for reuse) but moves much farther towards an environment where RDF is the norm and RDF authoring is common. Our work in this category will be significantly informed by the work of the Haystack project at MIT's CSAIL, which has been exploring requirements and techniques for solving them for many years. Haystack has developed several prototypes that we can use to begin design work on these deliverables, and Haystack PIs will consult with us on the best approaches to take in building these tools.

Future Deliverables

3a) Design interface and develop interactive tools for use by domain/community experts to create new ontologies (*)

The inspection tool described in 2b will be a great time-saver in the ontology creation process, but inevitably situations will arise when no existing RDF ontology exists that maps sufficiently well to the dataset in hand. In that case, the domain data expert will need to create a new RDF ontology from scratch, based on the data model suggested by the dataset. This is equivalent to modern data experts who design database schemas or XML schemas to store their data but with the RDF data model as the underlying technology instead. Rather than forcing data experts to become RDF experts, we propose to design a prototype tool that serves as a WYSIWYG authoring tool for ontologies. While the design of such a tool is straightforward, the implementation of one that is easy to use and effective across domains will be very challenging. No such tool exists today that is specifically aimed at non-programmers who have in-depth knowledge of the data model rather than the encoding technology.

3b) Design and implement a framework for creating RDF authoring Web applications for particular RDF ontologies to allow data experts to input or edit RDF data (*)

In practice, for managed collections (e.g., ARTstor or DSpace@MIT) the data experts will need tools to author and edit RDF metadata over time. SIMILE is not in a position to develop a general purpose RDF authoring tool (which would be an order of magnitude more difficult to develop than general purpose XML authoring tools, of which there are very few). But given reasonably well-designed RDF ontologies, SIMILE can produce a framework for creating sets of HTML forms for data entry and editing which will support a particular RDF ontology and do the conversion between HTML and RDF behind the scenes. Prototypes will be created with the framework for common ontologies such as Dublin Core (DSpace) and VRA Core (ARTstor) to be tested and refined by data experts from those organizations.

Category 4: Data Storage

All Semantic Web applications, including those in development by SIMILE, require effective, scalable storage, management, and navigation of large data sets that are represented in RDF. In order to be successful we will demonstrate the ability to effectively manage and query this data both in-memory as well as over a large persistent store (e.g. an RDF database). The infrastructure available today for storing, managing, querying, and processing RDF data is still rapidly evolving, with several commercial offerings expected this year. We need dedicated resources to evaluate the market and the current tools, to make recommendations for optimization of the tools, and to look for gaps that we can help the vendor community rectify.

Prior Work
An initial report on the scalability of triple-store applications was written in 2004 and published on the SIMILE website http://simile.mit.edu/reports/stores/. That report began the analysis of the current infrastructure tools, but much additional work is needed, and by an expert in RDF data storage technologies. Triple stores that have been evaluated to-date include Jena (from HP), Kowari, and Sesame.

Future Deliverables

4a) Evaluate and demonstrate performance metrics for large corpora, and investigate inferencing scalability to support improved browsing across collection (*)

The SIMILE tools currently use the open source software Sesame RDF database (http://www.openrdf.org/) from Aduna, a Dutch software company that develops information sharing and knowledge management software. We will evaluate the scalability of Sesame with various large datasets (e.g., the entire ARTstor collection and the contents of all the DSpace repositories world-wide, representing hundreds of millions of RDF triples). We will then evaluate other available triple store offerings (i.e. Kowari, 3Store, RDFStore, etc.) with the same data in order to understand which of them have the most potential for scalability and where we are going to encounter technical barriers to scaling. We will further evaluate commercial RDF database products that have been announced but not yet released by IBM and Oracle. Our assessment will focus on both the amount of data we can effectively store and access with each of these triple stores along with the performance associated with the simple inferencing required by the SIMILE project (i.e. metadata equivalence, refinement, etc.). We will produce a detailed report on these findings and on our recommendations for scalable products and best practice in designing applications that use them.

4b) Demonstrate querying and inferencing over a persistent store (e.g. database-backed triple store) store vs. in-memory storage (to demonstrate scalability) (***)

In addition to demonstrating effective storage, query and simple inferencing over data stored in a persistent store (e.g. RDF database), we will provide similar functionality for in-memory storage. Both in-memory and persistent storage of RDF data are useful to support scalability, performance, and the effective integration of federated data (the

ability to pre-fetch and cache various data particular to a user's action). We will produce a report on these findings and guidelines on when to use which approach.

4c) Evaluate the scalability of browsing tools (both for Longwell and for other RDF browsers) with respect to the number of ontologies and very large RDF datasets (*)

Additional work will focus on the development of effective, scalable searching and browsing user interfaces to RDF content. The focus of this work will be on the development of browsers that provide efficient, effective, and coherent navigation of RDF data. Our primary focus will be the development of a web-based "fat client" leveraging JavaScript and XHTML. This work will be draw upon the lessons learned through the development of Hayloft, an Eclipse-based generic Semantic Web browser. To focus on the effective management and navigation of large collections of real-world content, we will demonstrate these capabilities via navigating the entire ARTstor collection and the collection of all items in all DSpace sites world wide.

4d) Demonstrate navigation and inferencing over a persistent store on remote disk rather than a local store (i.e. distributed, or federated, metadata collections) (***)

Additional focus will be on the successful query and integration of data on remote servers in addition to local stores. This deliverable will demonstrate the successful query and integration of data from at least two distributed DSpace repositories into a common navigation interface. This technique will demonstrate the potential benefit of integrating distributed, federated metadata collections and recognizes the social and institutional realities of distributed data management.

Category 5: End-user Tools

This category involves the creation of a variety of interactive tools for browsing, searching, and navigating through and among diverse, distributed collections of digital artifacts. There are two goals for this set of tools: providing prototype user interfaces that allow data owners and end users to see what is possible to do with RDF-encoded data (e.g. cross-collection searching and browsing), and providing production user interfaces that can be added to systems like DSpace as an addition to, or an alternative to the current, more limited user interfaces. Several projects we know of, including ARTstor, SAKAI, and VUE, prefer to develop their own user interfaces for their RDF data but still want to see what is possible, for inspiration, or in order to leverage SIMILE code libraries that are useful for their own interfaces. For example, much of the code in the Longwell system code (described in more detail below) is reusable in other user interfaces without adopting its particular look and feel, as we ourselves have done in the Piggy Bank system (also described below).

Prior Work

The ability to support cross-schema navigation, browsing, inferencing, equivalent-concept folding has been demonstrated with some sample test collections using the Longwell browser (software created by the SIMILE team). This browser demonstrates use of the Core Semantic Web Specification Stack (RDF, RDFS, OWL) in the domain of scholarly descriptive metadata.

In addition, during the past six months the team has created the Piggy Bank (and the related Semantic Bank) Web browser tool to perform local capture, annotation, and management of metadata from heterogeneous sources on the Web. The functionality of Piggy Bank is similar to commercial products such as Endnote™ but accepts any RDF-encoded or encodable metadata automatically, and supports interoperation across it for search, browse, navigate and display. Uploading locally created metadata collections to the Semantic Bank allows for local, community, or world-wide sharing and further annotation of these collections. Recently the team added functionality to Piggy Bank that allows it to capture RDF metadata from non-RDF sources (e.g. the current version of DSpace) via a simple, extensible mechanism that employs common Web technologies, e.g. XSLT to convert from the XML world or javascript to convert from HTML, to the RDF one. This opens up the possibility of using Piggy Bank across a large range of Web-based digital content repositories and search systems, whether or not they currently "speak RDF". It is this sort of simple bridge technology that the SIMILE project is committed to inventing so that the transition to the Semantic Web, with all its potential for increased efficiency and higher functionality, becomes possible right now. Piggy-Bank also now supports georeferencing via the Google Maps API which supports functionality such as displaying physical locations on a map of cultural heritage objects when the holding institution is known.

In addition to developing our own tools and software, SIMILE is leveraging work done at MIT on the *Haystack* project, an extensible "universal information client" that enables users to manage diverse sources of information (e.g., email, calendars, address books, and web pages) by defining whichever arrangements of, connections between, and views of information they find most effective. At times, the interaction offered by the Web browser interface is too limited. The Haystack project is exploring a "rich client" interface that allows RDF data to be manipulated as well as navigated. It might be used by librarians who wish to manage the collections described with SIMILE-produced metadata or by users who want to collect and manage their own subsets of the SIMILE information. Unlike Welkin, which displays information as a graph, the Haystack user interface aims for an end-user oriented presentation of information that is natural for naive end users.

In working on RDF browsing for SIMILE we found that life would be easier if we had a general ontology governing how to display RDF, a kind of stylesheet for RDF that allows us to indicate how we would like to present some abstract data to the user. A primary strength of RDF is its general applicability to any type of data that needs to be represented in a computer application, and RDF is usually further encoded in XML. But designing applications to use XML-encoded RDF can be challenging for the same reason that RDF is powerful – the data model is the same no matter what the data can be used

for. In the same way that XML required a new stylesheet language to specific how it should be rendered on screen, RDF requires a new language to specify how it should be interpreted by receiving computer applications. There can be more than one view of a particular RDF data collection, hence the notion of "lenses" onto the data. Together with other members of the Semantic Web development community, the SIMILE team have defined a way of doing this called *Fresnel* (a type of lens favored by lighthouses), which is a generic ontology for describing how to render RDF in a human-friendly manner. *Fresnel* will be built into the next releases of Longwell and Piggy Bank.

Future Deliverables

5a) Complete the Fresnel vocabulary for describing rules for how to present RDF-encoded information on screen to users, and define heuristics for when to apply a particular data view (*)

The *Fresnel* ontology has been through significant specification work already, but is not yet at the point where it can be released as a W3C Note, the first step on its way to becoming a W3C standard. The remaining work to finish and document the specification is straightforward but will require additional time from the SIMILE expert who has been working on it so far.

Once the Fresnel vocabulary is implemented in Longwell as part of the RDF rendering subsystem, Longwell will be extended to compute which Fresnel view best fits a set of results and to allow user input to override its automated selection. The developer community around Fresnel will use implementation experience in Longwell and other applications to assess whether the vocabulary requires expansion or reduction and will incorporate any changes, if necessary, into the next and likely final version of Fresnel (see also deliverable 5h).

5b) Continue development of the Longwell faceted browser to improve the interface based on usability tests, to increase performance, and to add more features and functions for cross-collection search, browse, navigation and display of heterogeneous metadata (*)

Longwell has proved a source of inspiration for a number of other tools and interfaces and continues to be the locus of most of the code used in SIMILE tools (e.g. Piggy Bank). Longwell is currently in its second release but requires further development, particularly in support of large-scale collections and high volume Inferencing (see also category 4 above). This deliverable is focused specifically on the user interface and usability improvements needed as the tool scales up to much larger collections.

5c) Implementation and improvement of scaling and inferencing in Longwell over large data sets (**)

Longwell requires further development in support of high volume inferencing as described in deliverables 4b and 4c. Utilizing the reports and evaluations produced in 4b will provide a basis for specific implementation decisions in Longwell.

5d) Implementation of scalable, federated collection browsing in Longwell. (**)

Longwell also requires further development in support of browsing across data sources as described in deliverable 4d. Again, architecture reports and evaluations of existing federated systems will provide a basis for an implementation in Longwell.

5e) Enhance Piggy Bank and Semantic Bank to store binary content along with metadata. Broaden the scope of Piggy Bank to allow a greater number of desktop applications to interact with and contribute metadata to it. (**)

Both banks currently work only with textual metadata. They will be enhanced by extending metadata storage to include binary content (images, sounds, etc.).

Piggy Bank currently works with metadata in RDF acquired from the Web, but ideally a user should be able to integrate local data from the desktop into their Piggy Bank repository and manage all their textual or binary data in one place. The closest such tools today include the Mac OS Spotlight tool and the Google desktop tool, but these only work with unstructured text and do not allow you to specify relationships across the different sources of data other than simple word matching. Bringing the scope of data into Piggy Bank that Spotlight and the Google tool supports, but with the power of RDF and Inferencing to blend the data together, will take personal information management to a new level of effectiveness.

5f) Implement new features to improve the usability of Piggy Bank and Semantic Bank, (e.g. improvement to the user interface based on usability studies, new browsing methods, and so on) and improve performance. (**)

New functionality to be developed in Piggy Bank includes:

- Native support for SVG, canvas, and other Web formats for binary content
- Browsing by timeline (akin to browsing by geography)

5g) Integrate Piggy Bank with data cleanup tools for managing personal data (***)

Folding data cleanup tools into the data browsing suite would give users the ability to work directly with their own data, finding equivalencies across their disparate data sets to provide a more unified overall picture while browsing or generating more metadata from text and graph analysis. Integration depends on the completion of cleanup tools, explained in more depth in deliverable 2c.

5h) Develop and publish Fresnel display documents and continue to develop and publish Piggy Bank content harvesters for public use and adaptation (*)

A very straightforward activity of the group will be to continue to create "RDF scrapers" for conversion of HTML in useful websites for inclusion in Piggy Bank. These scrapers (normally in the form of an XSLT stylesheet) are simple to create but require frequent maintenance in cases where the source website managers make changes to their user interface that cause the stylesheets to fail. SIMILE team members can usually create or fix a scraper for a new website in a matter of hours, so we will continue to do this work for the set of scrapers that we depend on to demonstrate and use Piggy Bank, and we will add new scrapers as we identify useful sources of data. We are also developing documentation and instructions so that others can create scrapers for popular websites, and we will provide the infrastructure for people to post new scrapers as they create them (as they are already beginning to do, for sites such as PubMed Central and related scientific databases that are web accessible).

We will also publish the finalized *Fresnel* specification and related documentation, since both Longwell and Piggy Bank will depend on that technology for data sources that are RDF-native rather than dynamically-converted HTML. The Fresnel lenses used by Longwell and Piggy Bank to process RDF data, and the "RDF scrapers" used for legacy HTML data, together form the set of technologies needed to make RDF processing applications such as browsers useful during this period of transition from older, XML-specific data models to an increasingly RDF-centric World Wide Web.

Category 6: DSpace Metadata Engine

Create new DSpace information storage and retrieval components based on the Semantic Web tools and technologies described in the earlier categories.

Prior Work

The SIMILE project has already produced a number of advanced tools that demonstrate the utility of RDF and the Semantic Web to improve interoperability of metadata from heterogeneous sources. These are all available as open source software or reports on the SIMILE project website, and are enjoying a high degree of interest and adoption. But it is our intention to bring these tools, and those to be developed in the next two years, into the DSpace system so that the entire community of DSpace adopters (more than 80 in production at the present time, and more each month) and their users can begin to benefit from this technology. While the SIMILE tools and Semantic Web expertise has a broader audience than the DSpace adopters, we feel that by providing a DSpace outlet for our work we can insure that the higher education and cultural heritage sectors, who are the primary DSpace community today, will see improved functionality as a result of our work.

The evolution of the DSpace architecture is now fairly advanced by its many developers, and particularly by its current set of committers http://wiki.dspace.org/CommitterGroup. Some work on a re-architecture of DSpace has already begun, particularly in the area of scalable storage architectures using data grid middleware technology. The timing is good

to begin to specify how these SIMILE technologies can be integrated into a more modular DSpace design, and to roll out this work to the digital library community.

Future Deliverables

6a) with the DSpace developer community, redesign the DSpace system architecture to support the API required to integrate the SIMILE metadata engine. (**)

Identify gaps in current tools for scalability to larger data sets in the presence of inferencing

- Design DSpace 2.0 API for metadata engine, in the current User Interface layer of the DSpace architecture.
- Develop prototype Longwell implementation that conforms to the newly created DSpace 2.0 metadata engine API.
- Report on modifications, if any, needed in the DSpace 2.0 architecture or metadata engine API.

6b) Design SIMILE-based solutions to current gaps in DSpace functionality to support records linkages, and personal and corporate name normalization (**)

SIMILE's prior work with OCLC on their Name Authority Web Service was described above, and a quote from the OCLC website is illustrative (http://www.oclc.org/research/researchworks/authority/default.htm)

> *"Why we developed this*
>
> We developed this service so that remotely located systems—institutional repository software, for example (DSpace, EPrints UK, CONTENTdm, eprints.org, Fedora)—can offer authority control without having to build full authority control modules. Without this service DSpace does not know, for example, that Mark Twain and Samuel Clemens are the same man; nor does it distinguish well between two authors with the same name. With the OCLC Research name authority service, people entering metadata for preprints can make sure the author names are consistent and well-formed.
>
> DSpace plans to integrate interactive authority checking in the future."

The SIMILE project will allow DSpace to fully integrate OCLC's Web Service into its workflow for metadata authority and enhancement.

Further, OCLC is working with the Library of Congress and Die Deutsche Bibliothek in Germany to create the VIAF (Virtual International Authority File) of more than 11 million personal names initially, and the potential of expanding to every country that maintains such records, both for personal names and other controllable data (e.g. geographic locations or subject terms).

6c) Provide a SIMILE interface for DSpace, suitable for adoption and production deployment by the DSpace community, including (a) distribution-quality code, (b) packaging, (c) required modifications to the DSpace 2.0 platform, and (d) migration strategy from current DSpace platform (**)

Using the DSpace 2.0 metadata engine API developed in 6a, further development will be required to

- Implement a triplestore into the DSpace architecture
- Implement a Longwell faceted browsing interface to the DSpace triplestore
- Develop tools for importing RDF-encoded metadata (produced with tools defined in sections 2 and 3) into the DSpace triplestore

DSpace is a mature open source software system with an active developer community and an effective code governance process. MIT does not "control" the DSpace system code any longer, and so the use of RDF tools to support complex metadata functionality will need to be presented to the community in a way that convinces, rather than forces, the community to adopt this approach to a long-standing limitation of the system.

It is proposed that MIT's current DSpace "committer" (i.e. member of the group who officially maintains the DSpace code base on behalf of the entire community) join the SIMILE team as a bridge between the projects, and to insure that SIMILE outcomes are incorporated into the DSpace platform and that the DSpace Federation membership is kept informed of the benefits of this technology. SIMILE's goals, architecture, and implementation will be communicated to the DSpace development community (and other technical communities in the same general area, including Fedora and EPrints-adopting institutions).

6d) Perform outreach to DSpace and other communities, particularly in the digital library and educational technology domain, on project findings and tools available for use (*)

Ever since the launch of DSpace as an open source software application in 2002, the community of DSpace adopters has been pressing for a solution to the problem of needing support for multiple types of metadata from different domains. Both MIT and HP have consistently advocated for an RDF-based solution to that requirement, but in order to leverage the SIMILE-based tools that we will provide in DSpace, significant education and outreach will be necessary. DSpace has a strong track record for providing useful documentation, training, articles in the media, presentations at DSpace user group meetings and other appropriate venues, and anything else that will help institutions understand what the application is, how it works, and how they can use it to improve their local service portfolio to the academic community. All work done to integrate SIMILE and DSpace will be treated similarly.

Related Projects

VUE
The Visual Understanding Environment project at Tufts' Academic Technology department provides faculty and students with flexible tools to successfully integrate digital resources into their teaching and learning. VUE provides a visual environment for structuring, presenting, and sharing digital information and an OKI-compliant software bridge for connecting to FEDORA-based digital repositories. Using VUE's concept mapping interface, faculty and students design customized semantic networks of digital resources drawing from digital libraries, local files and the Web. The resulting content maps can then be viewed and exchanged online.

Currently VUE supports user annotations of the digital objects available from source repositories, but no other metadata is brought along with the object when it is available. The VUE team would like to extend the concept mapping tool to include metadata about the digital objects when it's available, and to index that metadata as well as the user-added annotations, so that users can find objects and maps in their collection more easily, and use the metadata to assist with identifying links to form addition maps. VUE could then become both a concept mapping tool and one for managing personal collections of maps and their associated digital objects.

But the source repositories do not use consistent metadata standards that can easily be merged together inside VUE, and VUE users cannot be expected to normalize the metadata or map it themselves, on-the-fly, during import. Converting metadata into RDF when a digital object is imported into VUE and using Semantic Web infrastructure (RDF triplestore, SPARQL query language API) would allow VUE to support this functionality easily. Initially VUE would need to use RDF conversion tools for each supported source repository platform and metadata schema, but SIMILE will provide many of the necessary converters, proxy servers, and a toolkit for allowing metadata experts (e.g. librarians or academic computing staff) to create new converters when necessary. SIMILE can also promote techniques for repository platforms to provide simple access to RDF converters directly (for example, using GRDDL – http://www.w3.org/2004/01/rdxh/spec, the new specification from the W3C for getting RDF data out of XML and XHTML documents using explicitly associated transformation algorithms, typically represented in XSLT and made available via the web site).

Fedora
The Fedora digital repository system manages and delivers digital content. It implements a sophisticated digital object model that supports multiple views of each digital object and the relationships among digital objects. Only system metadata to control the digital collections is required, other types of metadata (descriptive, technical) is optional and treated as a component of the digital object and encoded in XML. Fedora implementers can develop local metadata policies and user interfaces that support search and display of that metadata. Fedora supports a Search Interface, via a web form or a custom URL syntax, to both the system metadata and the Dublin Core descriptive metadata (when available via an Implementer-Defined datastream). Fedora uses RDF to track

relationships between digital objects in the repository, but currently uses standard XML for all other metadata.

Supporting interoperability for searching, browsing, and navigating content found in various repository platforms including Fedora, DSpace, EPrints, and others, has long been a goal of repository developers and the tool builders that support the academic community. In discussions with the Fedora project directors, we have agreed that using RDF and Semantic Web infrastructure provides a way forward to achieving that interoperability at very large scale. If the primary repository platforms such as Fedora provide RDF interfaces to descriptive metadata then tools such as VUE and Piggy Bank and the various tools in the Sakai framework. The Fedora project directors are interested in developing this functionality, and in the meantime SIMILE developers can provide XML converters for Fedora repositories that implement Dublin Core metadata. SIMILE's work in RDF conversion and infrastructure (e.g. triplestore scalability) will help Fedora in its adoption of RDF both for all types of metadata.

Sakai

The Sakai Project is a community source software development effort to create a new Collaboration and Learning Environment (CLE) for higher education. The Project's primary goal is to deliver the Sakai application framework and associated Course Management System tools and components that are designed to work together. These components are for course management, and, as an augmentation of the original CMS model, they also support research collaboration. Sakai version 2.0 was recently released, and the initial project will conclude later this year.

Sakai is a very ambitious project that depends on a wide variety of data – course description, course materials, student and teacher personal and preference information, registration data, grades, announcements, security, and much more – most of which has no defined standard schema or data model, and all of which is evolving quickly. The Sakai project leaders have identified the RDF data model as the way forward for Sakai, as the best way to manage that data evolution in an efficiently and sustainably. Using RDF as the Sakai data model will also make interoperability with other critical applications (e.g. archives of course material like DSpace or Fedora) much simpler.

An RDF-enabled SAKAI can take advantage of SIMILE infrastructure like *Fresnel*, the new display ontology for RDF, to support flexible and customizable course web pages, and the Longwell suite of search and browse tools, and the planned RDF ontology building tools. The SIMILE and Sakai projects are already in communication and thinking about ways of leveraging each others work.

ARTstor and JSTOR

SIMILE has already demonstrated the ability of RDF to enable interoperability for searching, browsing, and navigation across multiple significant academic collections, e.g. ARTstor, JSTOR, OCW, and Harvard's VIA catalog of digital images.

Beyond that, ARTstor needs tools to help them improve their ability to explore and manipulate the data they acquire from many, many outside sources which they must then normalize and link for import into ARTstor. ARTstor has a large group of metadata experts who work with metadata they've been supplied by the various sources of their images. They would like tools to help them analyze new data, visualize it properties and relationships (e.g. clusters) and understand its relationships to the metadata already stored. SIMILE has a number of tools (e.g. Welkin, Gadget, Referee, and Vicino) which convert new metadata into RDF and perform these types of analysis and visualization on it, and we plan to further develop these tools and to create new ones specifically designed for use by metadata experts. In addition to supplying us with ideas and requirements for these tools, ARTstor staff can field test them and provide feedback on their usefulness.

OpenCourseWare
SIMILE has already converted a sample of OCW metadata for course materials, originally encoded with the IEEE Learning Object Metadata XML schema, into RDF and imported it into Longwell to demonstrate interoperable searching, browsing and navigation with other collections of teaching materials (e.g. ARTstor, JSTOR, and Harvard's VIA).

While it seems unlikely that OCW at MIT would benefit from introducing RDF directly into their metadata production workflow, as other universities develop new and open systems for OpenCourseWare at their own institutions (as is planned by Sakai and in development at Utah State University), SIMILE's RDF tools will be a natural way to expand OCW's metadata support for additional schemas and with different production workflows. This is similar to the tools that MIT envisions using for metadata production and management in DSpace. Developers at Utah State have already begun to evaluate RDF as the data model for their EduCommons software.

Project Staffing

Project Investigators

- MacKenzie Smith, MIT Libraries. Project PI with overall responsibility for project management and problem definition (15% funding)

  Together with the Technical Project Manager, MacKenzie will provide overall project management for SIMILE, including budget management and reporting, hiring and personnel management, outreach and marketing of findings, coordinating of requirements definition and tools testing, coordination of data collections (e.g. from ARTstor, Getty, etc.), liaising with collaborators from other projects (Sakai, Fedora, etc.) and other high-level management activities as necessary.

- Eric Miller, W3C Semantic Web Activity Lead / MIT CSAIL Scientist (15% funding)

  Together with the Technical Project Manager, Eric will provide the overall technical leadership to the Semantic Web Advanced Developers. Eric's responsibilities will additionally include working with industry to provide supporting tools and technologies that advance this project, along with providing expert guidance on related Web standards and enabling technologies to help ensure the projects success. Additionally, Eric will provide a liaison role between the Simile project and other related Semantic Web Advanced Development activities of the W3C along with a conduit for migrating the pre-standardization work developed by the team into future standards work to help reduce future costs associated with future projects the leverage the Simile infrastructure.

- David Karger, MIT CSAIL (10%)

  Professor Karger will apply his experience researching information management to help guide the development of tools for librarians (and other data managers) and clients to work with the rich semi-structured information corpora our system will hold. Our system posits a relatively lightweight, web-based interface for typical clients, and a more heavyweight rich client to help librarians manage the data. Meeting regularly with the developers, Karger will help transfer ideas from the Haystack research project, which has answered many questions about how to design user interfaces to help users interact with partially structured information.

Appendix 1: Project Timeline