A Proposal to the Andrew W. Mellon Foundations

Title	Repositories Interoperability Framework: Augmenting Interoperability across Scholarly Repositories
Period	October 1, 2006 – September 30, 2008
Amount	$555.413
Submitter	Carl Lagoze Senior Research Associate Cornell Computing and Information Science 301 College Ave. Ithaca, NY USA lagoze@cs.cornell.edu

Attachments:

- Proposal Body
- Curriculum Vita – Carl Lagoze
- Curriculum Vita – Herbert Van de Sompel
- Position Description – Software Engineer
- *Rethinking Scholarly Communication: Building the System that Scholars Deserve*, D-Lib Magazine, September 2004.
- *Pathways: Augmenting interoperability across scholarly repositories,* submission to Journal of Digital Libraries

Repositories Interoperability Framework:
Augmenting Interoperability across Scholarly Repositories
A proposal to the Andrew W. Mellon Foundation

1 Introduction

We are in the midst of radical changes in the way that scholars produce, share, and access the results of their work and that of their colleagues. In the past several years we have seen dramatic growth in the deployment of scholarly repositories including institutional repositories [15], data set repositories, and others. These increasingly populated repositories are interconnected by a nascent interoperability framework based on standards such as OAI-PMH [23] and OpenURL [19], which are the result of previous work by the authors of this proposal.

Yet, the developments we have seen thus far are certainly just the tip of the iceberg, and the technical infrastructure is still rather primitive. In an earlier publication [24] we proposed a more advanced notion of a scholarly communication system that would fully compliment the way that scholars work. Such a system would allow flexible composition of information units – text, data sets, images, etc. – and distributed services for the formation of new types of published results and new methods of collaborating. This would be a loosely coupled system based on an interoperability fabric where the units of scholarly communication could follow a variety of scholarly value chains. In such a proposed system each hub provides a service such as registering results, certifying their validity, alerting scholars to new claims and findings, preserving the scholarly record, and ultimately rewarding scholars for their work.

The opportunities that such changes offer for radically changing the nature of scholarship have captured international attention. Terms like *cyberinfrastructure* [6], *e-scholarship*, and *e-science* all describe a concept of *data-driven scholarship* where researchers access shared data sets for analysis, reuse, and recombination with other network-available resources. Interest in this new scholarship is not limited to the physical and life sciences. Increasingly, social scientists [9] and humanists [10] are recognizing the potential of networked digital scholarship.

A core component of this vision is a new notion of the scholarly *document* or *publication*. Rather than being static and text-based, this scholarly artifact flexibly combines data, text, images, and services in multiple ways regardless of their location and genre. Examples of such complex scholarly artifacts extend across disciplines as described by Don Waters in his introduction to a panel on this subject at the recent Joint Conference on Digital Libraries. His examples included complex digital objects created by archaeologists that combine artifacts, images, maps, and charts, and creations by art historians that combine museum artifacts, library references, and representations from scholarly service bureaus like ARTstor. He also cited numerous examples in the sciences including new digital objects based on genomic databases, chemical structures, and astronomic observations.

This vision requires an interoperability fabric that is considerably richer than that provided by OAI-PMH. Rather than just allowing exchange of structured descriptive metadata, it needs to represent and exchange information about *complex digital objects:* their structure, lineage, and persistent identity. By representing these new forms of information at its core, the infrastructure will support the development of applications that fundamentally change the way the scholars produce and share information.

We propose here a two-year project for the definition and deployment of an interoperability fabric that supports these capabilities. We call this the *Repositories Interoperability Framework*. The components of this proposed project are:

1. The formation and management of an international working group to develop a set of specifications for the Repositories Interoperability Framework. These specifications will describe common data models and interfaces for exchange of information based on these data models.
2. The establishment and management of an experimental deployment community that will exercise the interoperability fabric in a variety of milieus, with the goal of empirically proving the interoperability fabric before wide-scale deployment efforts.
3. The establishment of a sustainable community to support the widespread deployment and management of the standards fabric, and thereby make real and substantial changes in the scholarly communication system
4. The development and publication of reference implementations of the standards built upon common repository packages such as aDORe [22], DSpace [21], ePrints [2], and Fedora [17]. These implementations will be made available on the project web page as Open Source software under the terms of the Educational Community License [1].

This plan conforms closely to the highly successful model [18] we followed to develop and establish the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). We believe that the careful combination of strong leadership, organized community participation and feedback, and widespread experimentation were important components of a process that led to the de facto standardization of OAI-PMH on a very wide scale. We are confident that we can extend that model into the development of this admittedly more ambitious interoperability effort. We are also confident that the two years of work proposed here will have an effect on scholarly communication and research far beyond that achieved by our earlier work.

2 Context of this Proposal

The work proposed here builds on a strong foundation of over two years of work. This work both provides some proof of the technical feasibility of the infrastructure we propose and the interest in the community for such an infrastructure. This existing work is described below.

2.1 NSF Pathways Project

Since September, 2004 Cornell University and the Los Alamos Research Library (PI's Lagoze and Van de Sompel) have been funded by the National Science Foundation through grant number IIS-0430906 to investigate infrastructure for a distributed, loosely-coupled, service-based scholarly communication system [5]. In this context we have been examining a number of areas, in particular:

- The nature of a graph-based information model that expresses semantic and service-based linkages among information units.
- Techniques for matching semantic-based services to sub-graphs in the information model (i.e. matching complex document graphs to appropriate services for preservation, and other tasks).
- Machine-learning based techniques for fuzzy pattern matching in labeled graphs and recognition of complex document structures in those graphs [11, 12].

One major result of this work is the development of an interoperability framework for sharing information about complex digital objects among heterogeneous repositories and prototype demonstration of that framework. The details of this framework are beyond the scope of this proposal and described elsewhere [8], but in summary include the following components:

- A data model called the *Pathways Core* that provides an abstract way to describe complex digital objects. The key features of this model are a simple repository-centric identifier scheme, semantic typing, the ability to express lineage links among digital objects, linkages to concrete representations, and recursion to express arbitrary digital object nesting. Many aspects of this data model are inspired by ideas originally described in the seminal Kahn/Wilensky digital object paper [16]. It is also influenced by other complex object formats including METS [3], DIDL [14], and others.
- A format for serializing instances of the data model as *surrogates*. These surrogates provide representations of digital objects according to the model, and can be transferred between services, clients, and repositories. The experimental serialization is based on RDF/XML [7], but we anticipate that others are possible.
- A set of service requests for exchanging surrogates. Specifically, these service requests are *obtain* for requesting a specific surrogate from a repository, *harvest* for batch request of a set of surrogates according to filters such as date range, and *put* for requesting deposit of a digital object described by a surrogate.

We have built a prototype that demonstrates the utility of the model and interfaces for implementing an *overlay journal* [25]; a virtual aggregation of "articles" homed in various repositories. This prototype implements the Pathways interfaces on an aDORe, DSpace, and Fedora repository. It makes use of the Live Clipboard [20] technology to initiate movement of digital object representations among repositories. Building the prototype demonstrated that relatively advanced functionality could be built on top of relatively simple interfaces and models. We are optimistic that this work provides a solid basis for the work proposed here.

We note that the work proposed here *leverages* but does not *overlap* with the work in the NSF-funded Pathways project. The NSF funding specifically does not support the community level activities and specification writing proposed here. It also does not support the production and dissemination of reference software. Furthermore, we expect that the proposed Repository Interoperability Framework will be a suite of standards that will not simply be wholesale "technology transfer" from the NSF-funded Pathways work. We imagine that the activity proposed here – specifically the creation of a suite of community standards for cross-repository interoperability – may result in significant departures from the technical and terminological aspects of the NSF Pathways work and may result in decisions to defer functionality of the NSF-funded work to a later time. As noted in the next section, our preliminary efforts to expose the Pathways work to a broader audience already surfaced decisions made within Pathways, such as the *Put* interface and terms such as *surrogate,* that need to be reconsidered to achieve acceptance outside the research context.

We believe that this proposal is in the spirit of other Mellon-funded activities that support the transition of core research to wide community dissemination and understanding the mechanisms for enacting that transition. The first year of the funding proposed here will also overlap with the final year of NSF funding, offering a rich opportunity for further technical research supported by the NSF that can provide input for the proposed Mellon-funded community activities. We are planning to use this year of NSF funding to address some of the technical issues that arose in our prototype work, including refining the data model definition, and the nature of the interfaces. We also plan to extend our NSF-funded work with demonstrations of automated matching of web-available services to the complex objects described by the surrogates.

2.2 Community Workshops, Panels, and Presentations

The ultimate success of this project depends on a substantial level of community participation and buy-in. Thus, a significant amount of the proposed work involves community building and management.

We have already engaged in the initial phases of this community building and we note three activities in the sections below.

2.2.1 Augmenting interoperability across scholarly repositories Workshop [4]

This invitation-only workshop took place in April, 2006 and was jointly sponsored by Microsoft, the Andrew W. Mellon Foundation, the Coalition for Networked Information, the Digital Library Federation, and JISC. The two authors of this proposal were on the organizing committee. The results of the NSF Pathways work were presented at the beginning of the workshop as a strawman. Not surprisingly the meeting transitioned to critical discussion of the strawman, and participants raised a number of issues needing attention. A brief summary of these issues is as follows:

- *Deep vs. Shallow Copy* – We presented a data model and surrogate that intentionally avoided asset transfer, and instead focused on abstract digital object structure with contained references to the concrete data streams. There are a number of reasons for this, most notably an attempt to avoid intellectual property

complexities, but the workshop participants noted that a spectrum of asset transfer scenarios surfaced across applications. The mechanism for flexibly accommodating these scenarios needs to be resolved.

- *Identifiers* – We proposed a repository-centric identifier scheme, as an attempt to avoid lock-in to one of many competing object-centric identifiers. Identification is a notably complex and competitive issue that has bedeviled numerous interoperability efforts. The workshop discussions on this issue make us aware that any final results will require collaboration with the identifier communities to ensure that the Repositories Interoperability Framework is compatible with their schemes.
- *Persistence* – It is unrealistic for any infrastructure to assume absolute persistence of networked resources (especially when their control is decentralized). We proposed a declarative rather than an imposed method, which for some applications may not hold enough assurances. This approach needs to be examined, especially given that persistence of information objects is vital to the viability of the scholarly communication system.
- *Object Relationships* – Our proposed model expressed containment relationships and lineage relationships. There was general consensus that these two relationships are important but exist in a larger taxonomy of inter-object relationships.
- *Harvest Functionality* – OAI-PMH is a highly successful, and widespread interoperability standard. Our strawman included harvest functionality, but questions arose about how the proposed framework subsumes and/or compliments OAI-PMH. In particular, it is not clear whether the *ListRecords* is necessary, and could be replaced by *ListIdentfiers* combined with *obtain* functionality.
- *Put Functionality* – Perhaps the most contentious aspect of our proposal at the meeting was the functionality related to deposit of digital objects in repositories. As noted by Clifford Lynch, some of the confusion was the result of conflating two quite different goals: populating institutional repositories and transfer of information among repositories and services. Furthermore, it was clear that "deposit" should really be "deposit request", to avoid any sense that repositories "must" accept objects submitted via a put interface. It may be, as noted by Lynch, that an initial release of the Repositories Interoperability Framework would not include "put" functionality.
- *Terminology* – The choice of terms is always important in the politics inherent in any standards activity. We noted in the workshop that some of the terms we chose in our Pathways prototyping such as *harvest*, *surrogate*, *identifier*, and *put* caused confusion in a manner that drew attention away from the underlying substance.

There are no "right" and "wrong" answers to any of these issues that arose. Instead, like any infrastructure effort, the correctness of the ultimate decision derives from a subtle combination of technical feasibility and process. This proposal defines a package of work that aims to meet this measure of correctness. It balances focused work by an experienced technical work group with mechanisms for frequent community feedback and exposure. This is the technique that worked well for OAI-PMH.

We note that the workshop ended with substantial interest in carrying on activities in the area of interoperability across scholarly repositories. We are confident that a number of the attendees of the workshop would be good additions to the technical working group proposed here.

2.2.2 Augmenting interoperability across scholarly repositories Panel [13].

As a follow-on to the workshop, the authors of this proposal participated in a panel at JCDL 2006. The purpose of this panel was to report the results of the April workshop and gauge community interest. While the results of the panel are only anecdotal, it is our judgment and that of other attending parties with whom we have talked that the attendees of the panel showed ample enthusiasm for this interoperability work. This impression has also been evidenced in subsequent communications after the end of JCDL.

2.2.3 Additional presentations.

We have also presented these ideas at a number of international venues including IATUL 2006 (Porto, Portugal), ElPub 2006 (Bansko, Bulgaria), and the 2006 JISC/CNI Meeting (York, UK). Each of these presentations raised considerable community interest.

3 Proposal Work-plan

We propose to the Andrew W. Mellon Foundation the following schedule for work. Dates are, of course, proposed and may change due to expediency and other scheduling issues. Dates also may overlap to indicate parallel activities.

3.1 Year 1 – October 1, 2006 – September 2007

The goal of the first year is to develop an initial alpha set of specifications and build a working group and experimental community to develop and test those specifications. The preliminary schedule of activities is shown in the following table.

Proposed Dates	***Activity***
10/1/06 – 11/30/06	Project Groups Organization
10/1/06 – 11/30/06	Meeting organization & project setup
12/06	1st Working group meeting
12/15/06 – 1/15/07	Issue resolution
1/15/07 – 2/28/07	Initial specification draft
3/07	Specification feedback
3/07	Initial implement to specification
4/07	2nd Working group meeting
4/15/07 – 5/31/07	Stabilization of alpha specification
6/1/07 – 8/31/07	Controlled experimentation, implement to specification
6/1/07 – 9/30/07	Reporting and year 2 preparation

Further details on the activities in the table above are as follows:

- *Project groups organization* – Our experience has shown that the composition of groups involved in the technical, organizational, and political aspects of the project is essential to the success of the project. This phase will concentrate on

composing groups. Following a model that was successful in OAI-PMH, we plan three major groups, in which membership may overlap:

  - *Steering Committee* – The members of this group will provide guidance in setting the overall goals of the project and requirements of the interoperability suite. Its membership will be recruited from high-level representatives of the library, publishing, and Internet community who have a stake in the broad goal of cross-repository interoperability. We hope to achieve representation from a diversity of communities, to promote wide acceptance of the results of the project. We expect that meetings of this group will take place via quarterly conference calls. This group will have access to all internal communication and developing documents within the project.
  - *Technical Working Group* – This will be a group of approximately 10-12 technically focused people, whose main task will be the production and review of the suite of specifications that are produced by the project. Our expectation is that this group will consist of recognized technical leaders of the digital library and network standards community for whom a reasonably high level of commitment to this work conforms to their professional responsibilities. This group will meet face-to-face three times during the life of the project, and communicate regularly via email lists and conference calls.
  - *Specification writing group* – This will be a 4-5 person subset of the technical working group who will be responsible for editorial tasks (writing, rewriting, formatting, etc.) of technical specifications arising from the project.

- *Meeting organization & project setup* – This phase involves organizing the web presence for the project, the intra-project communication tools (e.g., Wikis), and compiling proper documentation and organization for the working group meeting.
- *1st Working group meeting* – This will be a two-day face-to-face meeting with the following goals:
  - Establish the goals and requirements for the technical task
  - Establish the process and communication model for the group
  - Define major issues that need to be resolved before specification draft
  - Partition the technical tasks and define sub-groups
- *Issue resolution* – During this period the working group will communicate electronically to resolve outstanding questions before specification draft. This resembles the model we used in the OAI-PMH process.
- *Initial specification draft* –During this period, the 4-5 person specification writing group will work intensively to draft the specification. This may involve a face-to-face meeting, which in the past has proven most productive.
- *Specification feedback* – During this period the first-draft specification will be shared with the working group (no external exposure). The specification will also be visible to the steering committee. The purpose of this review period is to prepare for the 2nd working group meeting by raising issues and developing counter-proposals.

- *Initial implement to specification* – A very small controlled sub-group of the working group will attempt to implement according to the draft specification. The goal of this is basic sanity checking and experiential testing in preparation for the working group meeting.
- *2nd working group meeting* – This will be another two-day meeting of the technical working group to settle remaining questions with the specification.
- *Stabilization of alpha specification* – The specification will be edited to stabilize technical and editorial content, and prepare for public exposure. At the end of this period, the specification will be published on the public web site as an alpha draft, and a mechanism for wider feedback will be implemented.
- *Controlled experimentation, implement to specification* – A wider, but still internal, experimentation group will use the specification for a set of experiments. These experiments will follow usage scenarios like the "overlay journal" used in our initial prototype.
- *Reporting and year 2 preparation* – As outlined below, year 2 is critical to the ultimate success of the project. This period will be spent developing a community-wide plan for the year 2 activities, and writing complete reporting of the year 1 activities. A key component of year 2 planning will be the identification of potential experimentation partners for year 2 and contact with those partners to solicit their interest. A number of other projects, such as the Library of Congress NDIIP program [4], have recognized the value of carefully soliciting participants in an experimentation phase.

3.2 Year 2 – October 1, 2007 – September 30, 2008

The goal of the second year is two-fold. First, we will work with Mellon, other funders, and with the Steering Committee to organize a community around the specifications and organize a deployment context. Second, we will graduate the specification from its stable alpha state (the result of first-year work) to beta and finally production status. The preliminary schedule of activities is shown in the following table.

Proposed Dates	***Activity***
10/1/07 – 11/30/07	Meeting organization and publicity
10/1/07 – 11/1/07	Specification review and editing
11/1/07 – 12/15/07	Community meetings (US and Europe)
12/15/07 – 2/15/08	Solicitation for community experiments
2/15/08 – 5/15/08	Community experiments
6/08	3rd working group meeting
6/15/08 – 7/15/08	Beta specification stabilization
7/15/08 – 8/15/08	Final public review
8/15/08 – 9/30/08	Final specification and project wrap-up
10/1/07 – 9/30/08	Reference implementation development

Further details on the activities in the table above are as follows:

- *Meeting organization* – A successful model in the OAI-PMH context was the sponsoring of two public meetings (US and Europe) to publicize the work, raise

interest, and field questions. During this period we will organize (publicize, find venues, find co-sponsors) for these meetings.

- *Specification review and editing* – This represents a transition of the specifications to a "post-alpha" phase, cleaning problems raised during the first year of work.
- *Community meetings* – These meetings (two meetings, one in a US location and another in a European location) will provide widespread public exposure for the work and provide the venue for developing an experimental community.
- *Solicitation for community experiments* – We will work with Mellon and other funders (e.g., Microsoft, JISC, NSF cyberinfrastructure, IMLS) to select communities of interest for experiments in a variety of targeted communities and hopefully establish a relatively small-scale grant structure for moving those large-scale experiments forward. This will build on information gathered in the first year of the project.
- *Community experiments* – Hopefully with some seed funding, we will coordinate and manage fairly significant representative experiments involving parties from a variety of disciplines and contexts. Our hope is that this will include not only the institutional repository community, but publishers, cultural institutions, and organizations like ArtSTOR. We believe that careful and consistent management is crucial to the success of these experiments.
- *3rd working group meeting* – This will be the final meeting of the technical working group to resolve any remaining issues with the specifications and fully evaluate the two years of work.
- *Beta specification stabilization* – This phase will produce a penultimate version of the protocol for public review.
- *Final public review* – This phase will provide a period for any public feedback on the specifications.
- *Final specification and project wrap-up* – This phase will produce and disseminate the final specifications and necessary reports on the two years of work.
- *Reference implementation development* – In addition to producing a suite of specifications, we hope to demonstrate the practicality of these specifications with publicly available Open Source (ECL) reference implementation plug-ins for a number of the popular open source repository implementations. The second-year budget includes funding for a software engineer, resident at Cornell, to develop and/or coordinate development of this software. We recognize that another plausible mechanism for this activity is contracting out these implementations to development teams within the individual repository communities. We note, however, that the existence of such teams is not certain (i.e. Fedora project funding ends in 3rd quarter 2007 and DSpace funding is currently uncertain). We propose, therefore, to leave this item in the budget and then review with Mellon at the end of the first project year the best means of acting on the substance of the line (the actual reference implementation development).

4 References

1. The Educational Community License, Open Source Initiative, 2006, http://www.opensource.org/licenses/ecl1.php.
2. eprints.org, 2005, http://www.eprints.org.
3. METS, Metadata Encoding and Transmission Standard, Network Development and MARC Standards Office, the Library of Congress, 2002, http://www.loc.gov/standards/mets/>.
4. The National Digital Information Infrastructure and Preservation Program, Library of Congress, 2006, http://www.digitalpreservation.gov/.
5. Pathways: Lifecycles for Information Integration in Distributed Scholarly Communication, 2005.
6. Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., Messina, P., Ostriker, J.P. and Right, M.H. Revolutionizing Science and Engineering Through Cyberinfrastructure, National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure, 2003.
7. Beckett, D. and McBride, B. RDF/XML Syntax Specification (Revised), W3C, 2004, http://www.w3.org/TR/rdf-syntax-grammar/.
8. Bekaert, J., Liu, X., Van de Sompel, H., Lagoze, C., Payette, S. and Warner, S., Pathways Core: A Content Model for Cross-Repository Services. in *Joint Conference on Digital Libraries*, (Chapel Hill, NC, 2006), ACM/IEEE.
9. Berman, F. and Brady, H. Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences, 2005, http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf.
10. Crane, G. What Do you Do with a Million Books? *D-Lib Magazine*, *12* (3).
11. Dmitriev, P. and Lagoze, C., Automatically Constructing Descriptive Site Maps. in *Eighth Asia Pacific Web Conference*, (Harbin, China, 2006).
12. Dmitriev, P., Lagoze, C. and Suchkov, B., As WE May Perceive: Inferring Logical Documents from Hypertext. in *HT 2005 - Sixteenth ACM Conference on Hypertext and Hypermedia*, (Salzburg, Austria, 2005).
13. Hey, T., Waters, D.J., Lynch, C.A., Van de Sompel, H. and Lagoze, C., Augmenting Interoperability Across Scholarly Repositories. in *JCDL 2006*, (Chapel Hill, NC, 2006), ACM/IEEE.
14. Iverson, V., Song, Y.-W., Van de Walle, R., Rowe, M., Doim Chang, Santos, E. and Schwartz, T. MPEG-21 Digital Item Declaration, International Organization for Standardization, 2000.
15. Johnson, R.K. Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication. *D-Lib Magazine*, *8* (11).
16. Kahn, R. and Wilensky, R. A Framework for Distributed Digital Object Services, Corporation for National Research Initiatives, Reston, 1995.
17. Lagoze, C., Payette, S., Shin, E. and Wilper, C. Fedora: An Architecture for Complex Objects and their Relationships. *International Journal of Digital Libraries*, *6* (2). 124-138.
18. Lagoze, C. and Van de Sompel, H. The Making of the Open Archives Initiative Protocol for Metadata Harvesting. *Library Hi Tech*, *21* (2).
19. National information Standards Organization (U.S.). The OpenURL Framework for Context-Sensitive Services, 2003, http://www.niso.org/standards/resources/Z39_88_2004.pdf.

20. Ozzie, R. Live Clipboard - Wiring the Web, 2006, http://liveclipboard.org/.
21. Smith, M., Bass, M., McClellan, G., Tansley, R., Barton, M., Branschofsky, M., Stuve, D. and Walker, J.H. DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine*, *9* (1).
22. Van de Sompel, H., Bekaert, J., Liu, X., Balakireva, L. and Schwander, T. aDORe: a modular, standard-based Digital Object Repository, 2005, http://www.arxiv.org/abs/cs.DL/0502028.
23. Van de Sompel, H. and Lagoze, C. The Open Archives Initiative Protocol for Metadata Harvesting, Open Archives Initiative, 2001, http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html.
24. Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C. and Warner, S. Rethinking Scholarly Communication: Building the System that Scholars Deserve. *D-Lib Magazine* (September).
25. Warner, S., Overlay Journals. in *CERN Workshop Series on Innovations in Scholarly Communication: Implementing the benefits of OAI (OAI3)*, (CERN, 2004).