A Proposed NISO Work Item:
Specification for Resource Synchronization (NISO/OAI Effort)

Proposal for Consideration by the NISO Voting Membership
Approval Ballot Period: November 14 – December 14, 2011

The following proposed work item is submitted by:
Todd Carpenter, NISO; Bernard Hashofer, Cornell University; Nettie Lagace, NISO; Carl Lagoze, Cornell University; Peter Murray, LYRASIS; Michael L. Nelson, Old Dominion University; Robert Sanderson, Los Alamos National Laboratory; Herbert Van de Sompel, Los Alamos National Laboratory; Simeon Warner, Cornell University

Proposal Last Modified: November 9, 2011
Approved by the NISO Discovery to Delivery Topic Committee, November 11, 2011

BACKGROUND AND PROBLEM STATEMENT

This project will result in a new standardized specification for synchronizing web resources. Because of the proliferation of replicated copies of works or data on the Internet, keeping the repositories holdings up-to-date and accurate is an increasingly challenging problem. The specification will save a tremendous amount of time, effort and resources by repository managers, by automating the replication and updating process. By doing this, it will increase the general availability of content available from these repositories. It will also alleviate the variety of problems created by outdated, inaccurate, superseded content that exists on the Internet.

Increasingly, large-scale digital collections are available from multiple hosting locations or are cached at multiple servers. Examples of such collections include the Internet Archive's WayBack Machine, Twitter's collection of tweets, and Linked Data collections such as Freebase or DBpedia. In addition, high profile portals rely on resources originating in many distributed repositories. Examples include the Europeana portal, CiteSeer, HathiTrust, and OAister. The need has arisen for methods to keep these heterogeneous systems that rely on each other’s resources in sync, ways to ensure the freshness of content, and mechanisms to know in real time when and how resources are changing. Although synchronization methods exist, they are generally ad hoc, arranged by the individuals involved, and cannot be universally deployed. There is a pressing need for a well-specified mechanism for resource synchronization that scales up to the existing size of digital collections and their distribution patterns.

Synchronization is especially important for high integrity or essential Web resources. For example, portals that to deliver high quality services pertaining to aggregations of cultural or scholarly resources would clearly benefit from reliable, uniform, and scalable techniques to remain in sync with the collections they build upon. And, in more serious situations, for example the coordination of activities in response to a natural disaster, real-time discovery of updates suddenly becomes a matter of life and death. Finally, as interest grows in studying the dynamics of the web and formally modeling those dynamics, we anticipate that there will be parallel interest in tools and generalizable methods for extracting data about those dynamics.

As we move from a web of documents to a web of data, synchronization becomes even more important: decisions made based on unsynchronized or incoherent scientific or economic data can have serious deleterious impact. As we advance the nature of the content that the web can provide, we must advance the tools necessary for discovering and describing ever-changing collections of web resources.
STATEMENT OF WORK

1. PROJECT GOALS

This proposal plans to research, develop, evaluate, test, and advocate a new specification, tentatively titled ResourceSync, which will be released as an open national standard that will allow for the real-time synchronization of web resources housed in separate repositories. Building on strategies for synchronizing metadata (OAI-PMH), this project will enhance that specification using modern web technologies, but will allow for the synchronization of the objects themselves, not just their metadata.

2. SPECIFIC DELIVERABLES AND OBJECTIVES

ResourceSync will:

- Address the resource synchronization problem at scale;
- Build on existing, widely deployed technologies where possible;
- Provide a capability to express change semantics;
- Provide a transition path for OAI-PMH 2.0 applications;
- Appeal to various communities, including digital libraries, Linked Data, and Web 2.0 environments.

OAI-PMH 2.0 (http://www.openarchives.org/pmh/), developed by members of this proposal team, was designed to address the resource synchronization problem and its approach entailed recurrent metadata harvesting about those resources. As a result, the metadata about the resources can effectively be synchronized with OAI-PMH 2.0 but synchronizing the resources themselves was never specified and hence was dealt with using various ad-hoc approaches.

Furthermore, OAI-PMH 2.0 was specified before certain, now commonly-accepted, web programming design principles were established, most importantly the Representational State Transfer (REST) model. Recent standards such as Atom and its extensions such as the Atom Publishing Protocol, have adopted the REST-based design. Atom is the successor to the RSS (simultaneously “RDF Site Summary” and “Really Simple Syndication”) family of (incompatible) syndication formats that evolved from newsfeeds, blogs, and similar services. Owing to this heritage, both RSS and Atom are biased toward “recent” changes (e.g., “today and yesterday”) and typically lack facilities for describing “historical” changes (e.g., “last year”). This bias is a result of their initial applications: news feeds like CNN.com are designed to let you know the latest stories, but not maintain a list of all stories that have been published at that site.

In addition to OAI-PMH and Atom, the proposed effort will be inspired by other technologies that were developed by the submitters of this work item and that are relevant to the resource synchronization problem domain. These include the OAI Object Reuse and Exchange (ORE) protocol (http://www.openarchives.org/ore/) for describing aggregations of Web resources, the Memento protocol (http://www.mementoweb.org/guide/rfc/ID/) for uniformly accessing time-stamped resource versions on the web, and the DSNotify change detection framework (http://dsnotify.org/) for Linked Data.

3. PROCESS

We propose to research, develop, prototype, test, and deploy mechanisms for the large-scale synchronization of web resources.

Our work will start with a detailed formulation and scoping of the problem, and a related inventory of existing standards and technologies that can play a role in these mechanisms. The end product of our work will be a specification, vetted by experts and reference implementations, code libraries, and tools, which details an approach to synchronize Web resources at scale in an interoperable manner. The project will also provide training opportunities for the community, implementation resources, and documentation for future implementers of the standard.
This will be a joint effort between NISO and the OAI, with a core technology team and an outreach team consisting of the members listed on this proposal, each having extensive prior experience with standardization efforts. We plan to construct a Technical Committee composed of the team members listed here and jointly selected representatives from NISO constituents, and stakeholder communities such as Linked Data, the W3C’s DatasetDynamics, web archiving, and publishing. We have successfully followed this model in the past with OAI-PMH, OAI-ORE, OAC, and other community standards. The Technical Committee will assist the core team by providing additional real-world experience and reacting to proposed specifications. Presuming this work proceeds as a national standard as envisioned, the work of the team will be vetted by the NISO Discovery to Delivery Topic Committee and the NISO membership according to NISO’s Operating Procedures.

In addition, we will engage leading practitioners and researchers and seek their input through an array of coordination mechanisms, such as wikis, email lists, webinars, and workshops collocated with existing conferences and meetings, such as the ACM/IEEE Joint Conference on Digital Libraries (JCDL), Linked Data on the Web (LDOW) Workshops (collocated with the World Wide Web (WWW) Conferences), Coalition for Networked Information (CNI) membership meetings, the Open Repositories conference, and the International Conference on the Theory and Practice of Digital Libraries (TPDL, formerly ECDL).

Throughout the effort, we will work with developers from various communities to ensure the specification is broadly reviewed and evaluated. Working within NISO’s structure, the group will pull together expertise from the various communities. The resulting specification will eventually be made openly available as a standard. In addition, freely available webinars will be presented to the community to explain the rationale, need, and functionality of the standard to the community. Training seminars will also be recorded and made publicly available as well as other published training materials to help speed implementation and adoption of the specification.

A high-level, tentative timeline for project activities that will be accomplished by the date shown is as follows:

- December 2011: Select members of technical committee; establish communication rules, goals, and governance.
- January 2012: in-person meeting of technical committee to select technology candidates that have been circulated and discussed prior to the meeting.
- March 2012: alpha draft of specification; early tools, libraries, and other implementation demonstrations; early quantitative evaluation against OAI-PMH 2.0, http web crawling, and other candidates.
- May 2012: beta draft of specification; public (i.e., outside of the committee) draft standard for trial use, release of demonstrators and review.
- June 2012: workshop at JCDL 2012 in Washington DC for public review and commentary; identify a suitable European venue for the same review and commentary.
- September 2012: respond to public commentary, begin finalizing specifications; TPDL 2012 workshop.
- October-December 2012 – Training webinars
- December 2012: specifications stable; tools and libraries released
- January 2013: Approval by NISO Discovery to Delivery Topic Committee, Release to NISO members for balloting and member input
- February-June 2013: Additional Training events and training implementation calls
- April-June 2013: Response to balloting feedback, finalization of documentation
- July 2013: Final report of grant project and impact assessment written.
- August 2013: Final publication of standard
4. PARTNERS AND PARTICIPATION

NISO and the Open Archives Initiatives have been successful in bringing together teams of experts to develop and test specifications, then vet them through an open process of standards development. A core development team, a monitoring team, led by a NISO management team will oversee the development. The resulting specification will be vetted by the NISO membership prior to publication. Through in-person and web-based training efforts, NISO will provide the community with resources to smooth adoption.