October 30 Meeting Notes

Agenda

- review webinar on name identifiers
- discuss approaches to implementation
    - how much to implement
    - possible models
    - possible collaborations – see just below
    - scalability issues

From Dean's 10/21 agenda:

> One item people might want to take a look at was brought to my attention by MacKenzie Smith during my visit to MIT. Apparently OCLC is working on a follow-on to the original PURLs. This is still a work in progress, but some info is available at: http://prototypo.blogspot.com/2008/02/beyond-redirection-rich-and-active.html

Notes

present: Adam, Bill, Brian, Jon, Marty, Rick, Simeon

Marty will circulate a link to the 10/29 NISO names identifier webinar. The Thomson ResearcherId and ISNI identifiers for individuals are further along than organizational identifiers; while of interest, we do not plan to collect or maintain external identifiers for individuals or organizations as part of this project.

LANL

Simeon reviewed the Los Alamos National Lab Id 2-step locator system for their approximately 10,000 aDORe repositories, many of which correspond to an individual ARC file. Resolution speed (~10 milliseconds) is essential since a single page view may require multiple (~10) separate id resolution actions. They maintain some 600 million identifiers, which are converted into a 16-byte md5 hash as the index of a very large MySQL table. The repository id is the only other field in the main resolution table; collisions are resolved via a much smaller 2nd table of repository ids, which also contains the URL of the repository. The system is powered by 4 IBM blade servers having collectively 16 Gb of memory, and was implemented in Java. Hashing the identifiers, while resulting in some collisions that have to be resolved, does support load balancing by the simple mechanism of the leading digits of the hash values.

We concluded this system was too closely linked to the aDORe federated repository design (which had not been implemented at the time Bill and Adam started using aDORe here at Cornell) to be of use for this project. While capable of scaling and very fast, we do not anticipate wanting a 2-stage approach or needing so large an identifier space.

OCLC PURL Service

Marty had researched the second generation PURL service software project at OCLC (see Purlz Initiative (OCLC and Zepheira)). A February, 2008 blog post by David Wood () characterizes as the most significant improvement the typing of the URLs and the potential to combine PURL resolution with other services. Initially intended to allow returning a variety of status codes (301, 302, 303, 310, 404 and 410), the new PURL service could "combine strong identifiers with rich metadata, providing the building blocks for other semantic applications."

The conceptual design for the new PURLZ service also goes further to suggest an "active PURL" capable of returning an RDF graph describing a web service, either to provide a more nuanced response or to redirect to different service(s).

At the time of the meeting Marty had not been able to confirm whether the new PURLZ service is still actively under development. Since the meeting an email from Thom Hickey Nov. 3 indicates it has been: "The PURLz project is very active. We are currently in the midst of getting the new code running here so we can migrate purl.org to it. Developments beyond the current PURL functionality are on hold until we get past that hurdle."

It's not clear whether we could have access to the source code. If contact can be made and looks promising, Simeon and Brian will look at the structure of the new implementation and assess whether it makes sense for Brian to do a local test implementation.

Another thread of discussion has surfaced since the meeting in the form of an email from Eric Neumann posted May 30th to the public-semweb-lifesci@w3.org list following a presentation by Eric Miller to a SemTech 2008 session, including:

> Even though PURLs offer re-direction (even when NCBI finally publishes its own RDF), the URI would remain as purl.org/../ncbi/entrez-23111. Some of us find this troubling, since most users of such data would expect the URI to state clearly it comes from http://ncbi.nlm.nih.gov or http://ebi.ac.uk/ . Some of you know this has been a long (and difficult) discussion, and we still have no clear proposal of how to do this. And here is where we have lost momentum-- no clear path forward that most of us would be happy with...
>
> Enter Active Purls: In Eric Miller's SemTech session on "PURLZ", he described a new model for PURLs coming out this summer called Active PURLs. These PURLS are associated with services that are defined the source. I asked him about one specific service: if one starts with a PURL URI to Entrez data converted to RDF by (let's say) the HCLS community, can one include a service to say that if NCBI now has created its own RDF Entrez version with URI, that the PURL should be permanently re-directed to the NCBI location? In other words, can one ask a bunch of PURLs "what are your direct, and permanent URIs to the authoritative source?" Eric's reply seems to suggest this is quite easy to set up..
>
> SO what is gained? Well if you have a large KB with lots of purl.org//ncbi/... records, there would be now a mechanism to replace all these wholesale to the new NCBI URIs. In other words, a global replace of older initial PURLs to proper URIs from authoritative spaces. This seems to make a lot of life science folks asking how to begin working with SW and URIs happy-- a clean and basic transition model can be offered to the community for getting started with SW via Active PURLS.

We also briefly discussed Pete Hoyt's local PURL implementation for use in assigning an actionable URL in the 856 field when cataloging records; it was not designed to scale so should not be considered a candidate for this project.

Harvard Name Resolution Service

We looked briefly at the Harvard Name Resolution Service Guide, notable for its description of an apparently complete set of administration tools. Bill questioned the continued use of URNs (as names without a specified location) vs. actionable URLs, but otherwise the group felt we should learn more about the implementation and availability of the source code – perhaps Dean has a contact from his recent visit to Harvard.

Handles

Bill asked whether the group had ruled out consideration of the Handle System at a previous meeting; while we have been moving to think of the resolution service without a set of richer accompanying metadata and services, the Handle service has worked well for eCommons and LSDI.
Bill has wanted to host a mirror of the local Handle server and Mann would be a logical place to do so. He will talk to Chris Manly from DLIT to see when it might be possible for Chris to work with Bill Klinko and John Cline at Mann to do a second installation for redundant resolution of Cornell handles.
A simple Handle record (vanilla configuration, no extra metadata) uses around 100 bytes. Our current Handle server contains 43 thousand Handles. To evaluate the Handle System further it would be helpful to create a test suite to load ~1 million records and conduct load tests on resolution performance.

Timetable

Given current projects underway in DLIT and ITS, completing an evaluation/load test of the Handle System and equivalent testing of the PURLZ and/or Harvard NRS system by the end of the calendar year is the most optimistic scenario for implementation. If the PURLZ and Harvard projects have data on scalability and load testing that would be helpful in our evaluation.