Technical Infrastructure Status

Prepared for the NSDL NVC meeting, April 25, 2006

This is a brief report on the status of development work on the new Fedora-based NSDL Data Repository (NDR), NDR-based tools, the search service, and the state of the production metadata harvesting process.

NSDL Data Repository

Since last October, we have encountered and overcome a number of challenges in implementing the NDR. With over a million metadata records, which translate into over two million digital objects in the NDR, this is an order of magnitude larger than previous production Fedora repositories. The initial load took nearly a month, and a system failure then required a complete reload.

Working closely with the Fedora development team, over the last few months we have seen dramatic performance improvements in load and operation times in the NDR. Batch processes that took weeks now take hours. Individual updates and deletes are now two orders of magnitude faster. This stress-testing process has also significantly improved the reliability and verifiability of the repository.

Currently, the pre-production repository contains 2.18 million digital objects, representing over 880,000 STEM resources. We are currently upgrading to the 2.1.1 release of Fedora, at which time we expect to fully synchronize the NDR with the existing Metadata Repository (MR). That process should be complete in the next few weeks. At that time, a fully operational beta version of the NDR will be available for use and testing.

There is one more development effort that must be completed before the NDR can fully replace the MR. Currently, while the underlying repository data can be fully backed up and maintained, the cached versions that are required for actual operation cannot. This would mean that in the unlikely event of a failure of the Fedora cache, the repository information might be unavailable for over two days while all the cached information was rebuilt. This is clearly unacceptable in a production system. To resolve this issue, we are building a transaction journaling system (similar to that used in database systems). This will ensure that we can recover the NDR quickly in the event of a failure. We anticipate completion of the journaling system in June.

Until the new development work is complete, we will plan to run the NDR and MR in parallel. Once the NDR is fully synchronized, we will make it the production system visible through search on nsdl.org, but keep the MR as the backup and be prepared to switch back to the MR in the event of failure.

NDR API

The initial public draft version of the NDR API is almost ready for release. The current working draft is available online at http://ndr.comm.nsdl.org. This is the API that is used by the applications being developed by the CI team (e.g. OAI ingest, Expert Voices, OnRamp). We have recently finished the design of an authentication/authorization system for the API, which will allow collections and the Pathways to directly manage their own resources and metadata in the NDR. We will be releasing the draft API to the NSDL Pathways and Technology Committee for comment and revision within the next 2-3 weeks.

Search

The NSDL REST Search service is fully operational, with complete documentation available at http://search.comm.nsdl.org/cgi-bin/wiki.pl?CornellSearchService. This service is being used by several NSDL portal projects, as well as the main nsdl.org site. The current production service is metadata-centric and indexes the current MR. In preparation for the switch to the NDR, work is nearly complete on resource-centric search. This will combine into a single search result the multiple statements about resources, from many sources, that will be a feature of the NDR.

Work is currently underway to modify the search service to support propagating metadata values from collections to the items that they contain. Since all collection records provide audience-level information, this will support searching by audience level over the entire NSDL. We expect this to be completed and implemented in production by the end of May.

Expert Voices

Expert Voices is an NDR-integrated blogging system, and the first application to build on the new capabilities of the NDR. An initial, pre-alpha version of EV was demonstrated and tested at the NSTA meeting earlier this month. A second pre-alpha test conversation "Boneyard Science: Investigating Forensics" (http://expertvoices.nsdl.org/k12forensics/) is underway and will conclude in about a month with an artifact (reading list or lesson plan, for example) that will be contributed to NSDL.

EV supports incorporating direct references to NSDL resources into blog entries about science. Within the NDR, these references result in annotation relationships being created between the blog entry and the resource, enhancing discovery, selection and use. Carol Minton Morris is currently working to put together 5-10 teams of bloggers to further test the system and start developing science conversations and context for the NDR.

Production OAI Metadata Harvesting

The production metadata harvesting process is going very well, with no significant backlog. Typically, adding a new collection to the NSDL involves extensive back-and-forth with the OAI administrator to resolve problems with OAI implementations, OAI protocol, and data problems. Once a collection has been successfully harvested, reharvests typically have a much lower incidence of problems. Nonetheless, our monthly failure rate for harvests runs between 20-50%. This requires a major ongoing staff commitment to work with the providers to resolve problems and get successful harvests. Our experience is reported in a paper (Metadata aggregation and “automated digital libraries”: A retrospective on the NSDL experience) which has been accepted for the JCDL 2006 conference. A preprint is available at http://arxiv.org/abs/cs.DL/0601125.

We have recently added a major collection, PubMed Central, to the NSDL. We are still working with the PubMed technical staff to resolve harvesting issues for some of their materials, but we have successfully incorporated over 108,000 PubMed resources so far.

Dean Krafft
Cornell PI