NSDL Library Architecture: An Overview

The National Science Digital Library (NSDL) was created by the National Science Foundation to provide organized access to high quality resources and tools that support innovations in teaching and learning at all levels of science, technology, engineering, and mathematics education. To support the development of a library potentially containing millions of records and offering a wide range of services, the Core Integration (CI) team is designing and developing core infrastructure to support long-term evolution and growth. Given the heterogeneous community of participants and technologies, the library is being developed with two key notions: a *spectrum of interoperability* and *one library, many portals*.

The NSDL must accommodate heterogeneous participants, content, and users through a spectrum of inter-operability that provides a low cost of entry into the library in order to support the broadest acceptance by participants. To support this notion, the architectural design for this library is based on sharing of human and machine-generated information about resources and exploitation of that information for the deployment of core services (e.g., search and discovery, and archiving).

NSDL users will be very diverse, including students, instructors, the public at all levels, librarians, NSDL federated partners, and community interest groups. A general portal to NSDL is available at NSDL.org. However, to best serve such a diverse audience, the library is designed to support the notion of one library, many portals. The goal is to provide many different views of the library but with user interfaces that convey the sense of a single library. In FY05, the first specialized portal was deployed serving middle school teachers. In addition, NSF funded Pathways Projects to provide stewardship to specific audiences, and their portals and services will also take advantage of the core infrastructure to provide additional audience and discipline-specific views of NSDL.

Current Architecture

The initial release of the NSDL technical architecture was made available to the public in December 2002. Follow-on activities concentrated on stabilizing the core infrastructure, making improvements to the user interface based on community feedback, automating manual steps of metadata harvesting, and automating metadata creation. FY05 developments included replacing the metadata-centric repository with a Fedora-based architecture, which includes the full functionality of the old repository, while adding features to model relationships between resources, services related to specific types of resources and multiple information types. Figure 1 shows the principal features of the core NSDL architecture. A description of the major components follows.

Major System Components

The NSDL Architecture consists of several components that interact via web service interfaces and protocols.

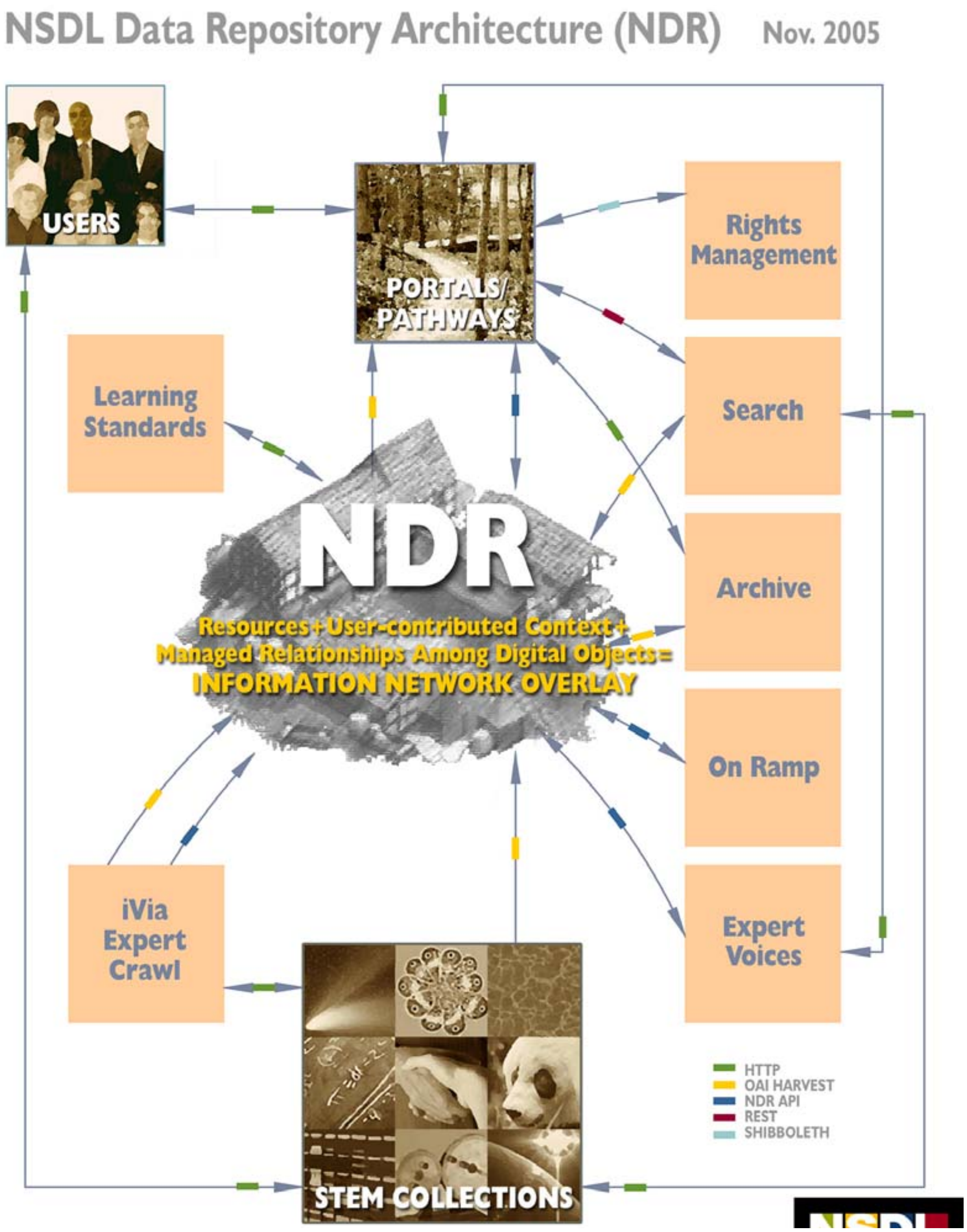

Figure 1. NSDL Architecture

***NSDL Data Repository (NDR)* –** Initially, the NSDL Repository stored only metadata ingested from participating projects, as well as metadata gathered from open-access Web resources. As CI gained experience with NSDL providers and users, and mapped out new potential services for NSDL, it became apparent that the repository needed to represent a more resource-centric view, including the need to support a) content, such as annotations, reviews, or information on structuring a set of resources in a lesson plan; b) explicit relationships among resources in the repository; and c) information about who or what organization provided a particular piece of information about a resource.

To support this view, CI created an ***information network overlay*** that represents the digital library as a graph of typed nodes, corresponding to the information units (documents, data, services, agents) within the library, and semantic edges representing the contextual relationships among those units. It expresses the complex relationships among information objects, agents, services, and meta-information (such as ontologies), and thereby represents information resources in context, rather than as the result of stand-alone web access. For more detailed information, see the November 2005 D-Lib article, ***What is a Digital Library Anymore, Anyway?*** *Beyond Search and Access in the NSDL* by Lagoze, et al.

CI has created the NDR using a Fedora digital object repository to support this information network overlay. The NDR represents resources as digital objects, and associates with them multiple metadata records from different sources. It represents the organizations and individuals that provide metadata or select resources, and relates them to the appropriate metadata and resources. Since Fedora is fundamentally a content repository, it can also represent content such as annotations or reviews. Finally, Fedora provides an RDF-based flexible relationship structure which supports arbitrary relationships among resources, for example relating all those that match a particular educational standard, or structuring the resources that are assembled into a lesson plan.

Metadata Harvesting - Metadata, in both raw and normalized forms, can be provided for harvesting by NSDL via the Open Archives Initiative (OAI) protocol or the NDR API .

- The OAI metadata harvesting protocol (http://www.openarchives.org/) uses the XML format is to ingest metadata. Information on metadata standards can be found at http://metamanagement.comm.nsdl.org/cgi-bin/wiki.pl.
- The NDR API uses REST calls to interact with the repository. Information on the NDR API can be found at http://nsdl-fedora.comm.nsdl.org/cgi-bin/wiki.pl?NsdlApi.

Search Service – The search and discovery component of NSDL provides fundamental capabilities for locating resources and collections within the library. Search services allow any item represented in the MR to be found, but in reality, the metadata provided by collections vary dramatically in formats, quality, and comprehensiveness.

To address this challenge, the search service combines indexing of metadata from the MR with indexing of full text content using open network protocols (e.g., HTTP or FTP) where the content is linked via the identifier in the metadata record and freely available. The underlying technology uses the Jakarta Lucene search engine, an open source text search engine (information at http://jakarta.apache.org/). Search services are directly accessible to both library portals and service providers who wish to search the contents of the MR directly via a REST interface. For more information, see http://search.comm.nsdl.org/cgi-bin/wiki.pl?CornellSearchService.

Access Management Service – NSDL core access management must accommodate a widely varying set of requirements from the users of the library and for ***rights management***, from the providers of intellectual property, i.e. the NSDL collections, and the providers of other services to the library. While many items in the library will be freely available and anonymous user access is permitted, access to some materials is restricted.

The NSDL core access management system relies on the Shibboleth protocol [http://middleware.internet2.edu/shibboleth/] to distribute identity verification (authentication) and cohort membership (authorization) to the administrators of distinct communities of users. In other words, the user's "home" institution performs user identity and capability management. Federated communities performing user identity and capability management can easily tie-in to this system using standard protocols (e.g., Kerberos and LDAP).

CI advocates using Shibboleth for NSDL collections' login systems. By logging into a participating service provider with Shibboleth, users potentially gain single-sign-on capabilities to that and other participating service providers, requiring only one login when navigating between them during a complete browser session. This eases the burden on individuals and organizations that would otherwise have to handle separate usernames, passwords, and related information. Shibboleth can enable personalization of services, and it allows user organizations to agree in a federated manner about what type of user information is appropriate to share with target websites that require log in. For more information, see http://www.columbia.edu/dlc/nsdl/shibtech/pathways2005/.

Main User Interface – Users of the NSDL access collections and services through portals. Because library users will be very diverse, NSDL will eventually offer several portals built by the Pathways Projects to support the unique needs of each user community. CI maintains the main library portal at NSDL.org using PHP, MySQL, and Internet Scout Portal Toolkit running on Apache Servers.

Archive Service – A basic requirement of national libraries is stewardship of the materials assembled within those libraries. The CI team is building persistent archive services to retrieve materials represented in the MR from public sites (with a crawl depth of 10 levels) and archive both the metadata and content for future retrieval. Web materials that are deleted or "lost" will be recoverable through archive services, and users will be provided options in NSDL search results to retrieve prior versions of resources.

iVia – Over the past several years NSDL CI has relied on technology that has been largely dependent on human-generated metadata. Despite its success, this work has demonstrated many issues with harvesting metadata from distributed sources. These include wide variations in metadata quality, scalability constraints of managing and validating multi-sourced metadata harvesting, and the expectation gap between current google-influenced search techniques (full-text and keywords) and search interfaces based on structured metadata. Consequently, a fundamental goal of CI is to move towards more automated tools for collection growth and development of core services such as search, aligning NSDL more closely with cutting-edge digital library technology.

iVia provides focused crawling, an automated mechanism for including resources in the NSDL collection. In the near-term, this will allow CI to include the resources in human expert-selected web sites, essentially "seeding" the crawler with the home pages of those web sites. iVia work also includes developing and deployment mechanisms for automatically generating basic Dublin Core records from textual resources. The iVia tool also provides automatic assignment of Library of Congress Classification to NSDL resources. Currently the iVia tool is being used by a small number of collection providers under the direction of CI staff before the tool is released more widely.

User Help – Text-based user help is available to users through NSDL.org. For reasons of scalability, our user-support and help services will be enhanced via automation—the CI team cannot provide direct personal assistance to individual users. CI provides Virtual Reference Desk (VRD) capabilities (information at http://www.vrd.org/) using AskNSDL (http://nsdl.org/asknsdl/) as a central component of the online NSDL help system. This system aids users in gaining assistance from one another, e.g., from more experienced colleagues who have expertise in using a specific collection or can answer discipline-specific pedagogy questions for example.

The VRD service is e-mail based and patron questions are distributed to a large group of experts. Responses are not intended to be immediate in this model, as they are in the traditional face-to-face reference interaction. However, we believe this approach will help establish a culture of sharing expert knowledge, and a true sense of community among NSDL users and contributors. In addition, providing direct personal assistance is an expensive resource endeavor, and the VRD approach is the only one that can scale to meet NSDL needs under the current funding model.

Next Steps

To support broader education and communication goals, current development areas include:

- ***Expert Voices*** – CI is developing a customized open-source blogging service that provides a communication avenue for experts, teachers, and library builders to share their knowledge. The blog styles will be both informative and journalistic. *Story starters* and AskNSDL questions will be used to spark discussions for the scheduled experts. The blogging system will allow and encourage tagging of NSDL resources and submission of recommended resources and simple structured metadata. This will create a system for establishing relationships between blog entries and resources. The entries will be searchable and available via RSS. More information can be found at http://blogging.comm.nsdl.org/.

- ***On Ramp*** – NSDL manages content in a variety of formats for use in multiple publications. Publication in this arena is a loose term that can mean publishing web content, print publications, email distributions, among others. To address the issues of large scale content management, On Ramp is attempting to address: workflow for the creation of content; content management; distribution of content to multiple publications (targets); and dynamic formatting conversions of content prior to delivery to targets. More information can be found at http://nccs.comm.nsdl.org/.

In the next development cycle, CI will concentrate on overlapping Pathways requirements that are designed to provide specific beneficial educational outcomes. Each Pathway supports audiences facing unique pedagogical and technological challenges.

By distilling overlapping requirements, CI aims for the following outcomes:

- To understand and refine the educational utility of NSDL by partnering with multiple projects, professional groups, and publishers with strong understanding of specific audiences and disciplines. This work includes understanding the interoperability requirements needed to integrate the work of the Pathways Projects, and ultimately realizes the “one library, many portals” concept.

- To develop technologies, processes, and standards that are broadly applicable to additional audiences and the entire range of NSDL activities, and test those developments in the context of the Pathways and specific partner’s sites.

A phased approach gives CI and our partners the opportunity to experiment, observe and document the educational uses of digital libraries, in the context of focused and significant audiences. Choosing technologies, standards and mechanisms for the core infrastructure is a complex process in which the needs of the NSDL, the availability of the technology, the costs, the adoption rate of end library users, and the wishes of our partners are all important. Our challenge will be to interpret trends and adopt the technical approaches that will best stimulate the growth of the NSDL while keeping it broadly accessible. We will make these choices in conjunction with the assistance of the broader NSDL and digital library communities.

For questions or further information, please contact Karen Henry, NSDL Technical Project Manager.