

## Reflections on Data Management in the NSDL

Mike Wright,  
Technical Director, IIS, UCAR  
Director, NSDL Technical Network Services Project  
mwright@ucar.edu

Brief for the ASIST Summit on Research Data Management for Access and Preservation  
April 9-10, 2010

### About the NSDL

The National Science Digital Library (NSDL) began under NSF funding in 2001 as a program to develop a library for research and education. It was envisioned to hold materials from NSF funded projects, and beyond, to be a central portal to support learning - <http://nsdl.org>. The NSF NSDL program funded a number of grants to develop core infrastructure and to develop a community of collection builders to populate the library. In the early years, a large part of the NSDL collections were references to scholarly articles from a number of publishers (e.g. ACM) and open access journals. In addition to the research material, education and training material was also included, primarily materials from many agency funded activities in Science, Technology, Engineering and Math (STEM).

In 2008, the library focus shifted in response to the changing environment in research and education. In the research area, publishers had developed extensive digital libraries for scholarly work, and services such as Google Scholar were becoming more accessible to finding these works. It was also becoming clear these avenues were the preferred methods for those looking for research articles. The NSF changed the 2008 NSDL program solicitation to refocus on STEM education; moving the library to cater to the STEM education community. The NSF program also changed its name to the National STEM Distributed Learning network to reflect the community that had built up over the years around the NSDL. The library continues on at <http://nsdl.org>, but the collections now focus on STEM education and training with materials that address audiences from K12 through undergraduate and graduate level. A result of the shift in focus has been a major collection deaccession process completed by the end of 2009 that saw much of the research materials removed.

### NSDL: A Network of Repositories

The NSDL is a network of repositories, each managed with a particular collection scope and to meet the needs of a particular audience. There are 114 collections that comprise a total of almost 130,000 items covering the range of core STEM disciplines (Biology, Physics, Chemistry, Math) that have a broad audience (K12, community college, undergraduate and graduate), along with disciplines that have a more specific audience such as Materials Science where the audience is primarily upper-division undergraduate and graduate. In order to manage the breadth of the library, the NSDL program developed an organizational construct of pathways, groups to manage particular discipline areas (e.g. the BEN pathway for Biology, the ComPADRE pathway for Physics), or particular audiences (e.g. the AMSER pathway for community colleges), or a combination (e.g. the MSP2 pathway for middle school

math). A core requirement for the pathways is stewardship of the collections that come under their purview. Currently, there are 16 pathways representing around 60% of the items in the NSDL. Each of the pathways maintains a pathway portal through which their collections can be discovered and viewed. In addition to the pathways, there are two central NSDL organizations; the Resource Center that provides support to the community network, and the Technical Network Services group that provides core infrastructure for the NSDL network.

The central NSDL library (<http://nsdl.org>), managed by the Technical Network Services group, provides a registry of all the materials found in the pathway collections, and other collections not currently associated with a particular pathway. The central registry provides services such as search and browse, collection background and scope information. Other central services relevant to education operated centrally include, for example, the Strand Map Service (providing an API and visualization onto a concept map for STEM), and the Content Assignment Tool (CAT) - a service that indicates what education standards a particular web-based education resource may relate to (this service utilizes machine learning for automated characterization). The current method of interoperability between the distributed repository collections and the central library is through use of OAI-PMH for exchanging metadata about the resources held in the various collections. We should note that some collections are repositories holding the material they manage, others are aggregations of metadata about materials spread across the web (i.e. they are registries themselves).

### **Scientific Data in the NSDL – the Data Life-Cycle**

A key point of the NSDL with respect to data is that data is held, or referenced, in the context of a learning resource, e.g. a training module on a particular meteorological phenomenon in the COMET collection includes specific data needed to provide the users with a concrete use case to explore. As is often the case in the resources held in the NSDL, the data is a copy of that found in major data archives such as the National Climatic Data Center (NCDC) operated by NOAA. In other cases, the data may be small in volume and come from a particular researcher, for example observation data for an ecological field study. The common theme is the way the data is copied from the research context, and the copy then managed with the education or training material as part of that body of work. The data may be massaged for use in these resources so we can consider these derived data in the notion of the data life-cycle. However, it is not common for this data to then be placed back in the management realm of a data repository – in fact it continues now as part of the educational realm.

In noting that data is managed in the context of the educational resource, it's instructive to understand where that resource itself is managed. In many cases, these are managed as web sites (or parts thereof), and not in any specific repository system. We are seeing more use of Content Management Systems for these materials (e.g. the open source Drupal CMS), where the resources (and any included data) are then managed in the CMS data model. We are also seeing the use of learning management systems (LMS) becoming more common place (e.g. Moodle, Blackboard in higher education, Angel and Schoolnet in K12). These systems tend to be closed in that the general use case is to upload your materials into the system to then orchestrate for a particular lesson or curriculum (the primary construct in these systems). This model would seem to be limiting when considering how resources may need to

interoperate with data held in managed repositories (assuming here that we would like to see the data used in educational resources be managed as part of the larger data space from which it comes). If we consider CMS and LMS products as repositories, we will still need to develop the interoperability mechanisms for allowing these to be part of a repository network. Even in the NSDL with its simple use of OAI-PMH, CMS and LMS-based systems often can't participate due to their not supporting OAI.

While we see these educational resources managed on web sites, or in LMS systems, we don't see many of these resources being managed in the growing number of institutional repositories. This isn't surprising given that IR's tend to have a narrow scope with respect to scholarly research material, often just pre- and post-prints of scholarly articles. Yet many of the resources in the NSDL have been developed as part of the research enterprise – often through the mechanism of the broader impact requirement of current funding. This does bring up a question as to why these are not managed in an IR (and corresponding data repositories) along with the related research material. Such a development would be very helpful to the NSDL in providing a means for stability of access to educational resources and associated data, and to preserving the broader impact of the initial research.

### **Partnerships for sustainability**

As mentioned earlier, a primary goal for the NSDL Pathways is to provide sustainable stewardship of resources (and by implication, associated data). To meet the longer term sustainability need of continuing to provide access to resources, many of the pathways have partnered with professional societies in their discipline. Two have partnerships with Public Broadcasting Stations. In all these cases, the partners are bringing infrastructure for longer term storage and access (although to differing degrees of sophistication). These partnerships were encouraged by the NSF as part of the NSDL program, and the emphasis of professional societies was a product of the thinking at the time. What is interesting, however, is the lack of connection to academic institutions or data centers, and in particular to efforts led by institutional libraries around repositories for scholarly work and data. Only two pathways have developed any connections, and mainly to maintain their portal infrastructure. In the future, the NSDL community needs to be more closely aligned with the institutional and data repository communities.