Building FLOW: Federating Libraries on the Web

Anna Keller Gold[1], Karen S. Baker[2], Jean-Yves LeMeur[3], and Kim Baldridge[4]

1 UCSD Libraries
University of California, San Diego
La Jolla, CA 92093-0175
1.858.534.1214
agold@ucsd.edu

2 Scripps Institution of Oceanography
University of California, San Diego
La Jolla, CA 92093-0218
1.858.534.2350
kbaker@ucsd.edu

3 European Center for Nuclear Research (CERN)
CH-1211 Geneva 23
41.22.76.74745
Jean-Yves.Le.Meur@cern.ch

4Integrative Biosciences
San Diego Supercomputer Center
University of California, San Diego
La Jolla, CA 92093
1.858.534.5149
kimb@sdsc.edu

ABSTRACT

Individuals, teams, organizations, and networks can be thought of as tiers or classes within the complex grid of technology and practice in which research documentation is both consumed and generated. The panoply of possible classes share with the others a common need for document management tools and practices. The distinctive document management tools and practices used within each represent boundaries across which information could flow openly if technology and metadata standards were to provide an accessible digital framework. The CERN Document Server (CDS), implemented by a research partnership at the San Diego Supercomputer Center (SDSC), establishes a prototype tiered repository system for such a panoply. Research suggests modifications to enable cross-domain information flow and is represented as a metadata grid.

Categories and Subject Descriptors

H.3.7 [Information Systems]: Digital Libraries – *collection, dissemination, standards, systems issues, user issues.*

General Terms

Management, Human Factors, Theory

Keywords

Document management system, Open Archives, Bibliographic citations, Eprint repository, Metadata grid, Tier, Class, Domain

1. INTRODUCTION

Solutions for managing documentation generally address the needs of a single tier (class) in the research process (individual, group, organization, or network). Existing practices are deeply embedded in and constrained by everyday workflows. Many individuals have adopted desktop bibliographic citation software packages such as EndNote or Procite to manage citation libraries optimized for personal, not group, use. Individuals wanting to share their citations or documents, as well as research groups wishing to create a common library, often construct static web pages where they can post selected documents and links. Research organizations often use centrally managed spreadsheets, databases, or library catalogs to create ongoing or ad hoc in-house bibliographies in response to administrative or public reporting requirements. Far-flung disciplinary com-munities may have electronic repositories such as eprint libraries

to manage sharing and exchange of research documentation [6]. Until recently, there has been little focus on how to integrate these divergent tools and practices. Two examples include Kepler (open archive software for individuals) [3], and MySRB (a web-based interface to the SRB or Storage Research Broker) [5]. This paper describes a third effort to realize integration across divergent practices and domains.

JCDL '02, July 13-17, 2002, Portland, Oregon, USA.

2. APPROACH

A multi-tier system for information flow in research makes it possible to enhance the participation of scientists at different stages of the research process. This in turn will make it possible to capture a wider range of research products, resulting in further discovery. This system requires protocols that authorize federation and interconnectivity, while preserving personal or institutional differentiation in the form of "streams" both into and out of the federated processes. The World Wide Web serves as a common interface to each layer, enabling participation and management at all levels of organization, regardless of local technical infrastructure. A metadata architecture supporting multiple layers can be thought of as a **metadata grid**, analogous to computational grid architecture, with "a series of layers of different widths" at the center of which are resource and connectivity layers and protocols [2].

Research organizations are well situated within the metadata grid to collect citation data from group members and to act as a sub-station for dispensing access to the data collected. Organizations will need to retrieve this data in multiple ways from a data management system that can be queried; they will also want to create public exposure for data gathered in order to enhance the impact of the research group. The needs of research organizations such as the Integrative Biosciences Program at the San Diego Supercomputer Center (SDSC) therefore have much in common with the requirements of large-scale research networks, such as the national Long Term Ecological Research (LTER) Network, the National Partnership for Advanced Computational Infrastructure (NPACI), and the European Organization for Nuclear Research (CERN). Discussions over many months among researchers and programmers at these institutions and at the research libraries of the University of California, San Diego, revealed the following common needs: (1) to gather (G) structured resource information, both in batch and individual submission modes; (2) to share (S) collections of information within research networks; and (3) to discover (D) relevant research, partners, updated results, and linkages. These

discussions resulted in a decision to form a partnership to create a prototype system that could leverage individual and small-group practices to feed organizational metadata repositories. These in turn would be aggregated across disciplinary and organizational domains, while preserving the ability to represent the individual or their group. For example, Integrative Biosciences (SDSC) wished to use the system to gather together all documents associated with a particular individual, project, or research instrument.

For all of our partners, the targets of these activities (G, S, and D) include the files and metadata representing preprints and technical reports, journal publications and book chapters, presentations, proposals, conference papers, courses taught, and multimedia. This is a much wider range of 'grey literature' than has typically been tracked in open eprint repositories.

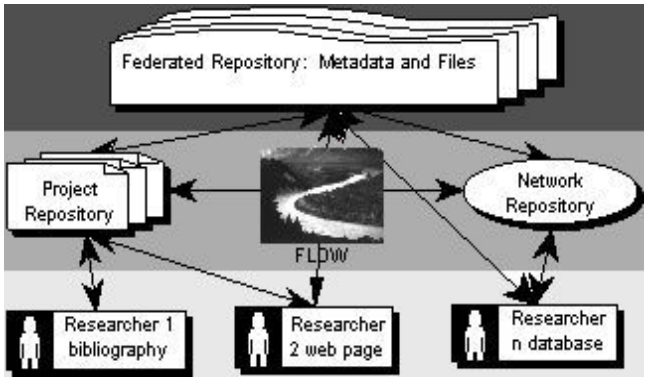

Figure 1: FLOW supporting grid metadata infrastructure.

3. PROTOTYPE SYSTEM

The first goal of the partnership is to develop a prototype system capable of handling G, S, and D functions for this wide range of metadata types. Requirements for the prototype were that it be standards-based, open, flexible, quick to implement, and that it allow search within and across collections. Functional desiderata included the ability to modify and update submissions; provide full text via local or remote files; search by fields or full text; extract and count citations within the system; handle various media and types of data; provide for administrative and peer review processes; permit customization and personalization; and support alerting services. In short, our requirements described the need for a generalized tool easily adapted to the needs of diverse researchers, organizations, and networks.

In selecting a system for the prototype we looked at two systems developed to meet cross-community document repository requirements: OpenEprints software (http://www.eprints.org/) and the CERN Document Server (CDS, http://cdsware.cern.ch). The CDS, under active development by programming staff at CERN, combines the features of an organization's library database with those of an open preprint repository with submit and search capabilities. It is compliant with the OAI protocol, and it includes many service features, including a batch upload module [4], customization features; workflows supporting review processes; alert services, various output and export options, and an automatic citation extraction and tracking capability [1]. CDS is distributed under the GNU General Public License, but technical support is available under a Technology Transfer agreement. SDSC and CERN programming staff are working in partnership with researchers to implement and modify the CDS system at SDSC. They are developing workflows and protocols, and creating new modules designed to support the integration of different layers in the research information process. The first such module will be a protocol supporting easy uploading of bibliographies created and/or maintained in EndNote, as well as the reverse (extraction from the repository to EndNote-compatible file formats). Another module being created is a "person" module that will manage information about individuals, currently handled in a personnel database, and will support associations between individuals and document-like objects managed by the system. While designing for local needs we plan to prototype a set of protocols that are general purpose and use established communication practices in research communities, and yet bridge the needs of individuals, local projects, and research networks and centers.

4. CONCLUSIONS

This partnership provides an arena for discovering and exploring the challenges of cross-domain document work practices. The integration of the three functions (G, S, D) is relevant to all tiers and domains: as an internal need for both small and large organizations, with both public and private types of data; and as a new collaborative way of working across the research domains of individuals, libraries, organizations, or publishers. The single system, supporting these three major axes, can be configured to support G, S, and D functions regardless of the object description, thereby merging a diversity of document objects (public and internal) with other data objects which are traditionally handled outside document repository systems.

5. ACKNOWLEDGEMENTS

The authors wish to acknowledge the generous support of the Integrative Biosciences Program at the San Diego Supercomputer Center (http://biology.sdsc.edu/), and the programming and design support of Frank Sudholt and Joshua Polterock, also of the San Diego Supercomputer Center. This work was supported by NSF Grants DBI-01-11544 and OPP-96-32763.

6. REFERENCES

[1] Claivaz, J., et al. From Fulltext Documents to Structured Citations: CERN's Automated Solution. HEP Libraries Webzine 2001, 5, (November 2001).

[2] Foster, I. The GRID: A New Infrastructure for 21st Century Science. Physics Today 55, 2 (February 2002), 42-47).

[3] Maly, K., et al. Kepler – An OAI Data/Service Provider for the Individual. D-Lib Magazine *7,* 4 *(*April 2001).

[4] Pignard, N., et al. Automated treatment of electronic resources in the Scientific Information Service at CERN. HEP Libraries Webzine 2001, 3 (March 2001).

[5] Rajasekar, A., et al. MySRB & SRB – Components of a Data Grid. Preprint, submitted to JCDL 2002.

[6] Suleman, H., and Fox, E. A. A Framework for Building Open Digital Libraries. D-Lib Magazine *7,* 12 (December 2001).