Engaging and Connecting Faculty: Research Discovery, Access, Re-use, and Archiving

Jon Corson-Rikert and Janet McCue, Albert R. Mann Library, Cornell University
{jc55,jam7}@cornell.edu

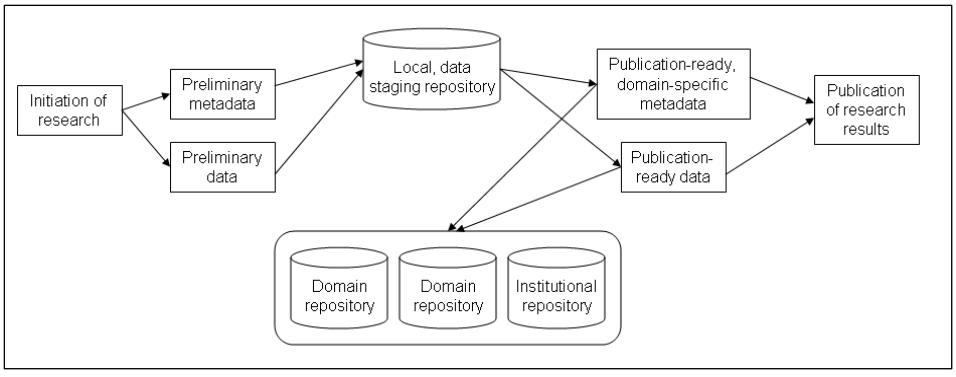
Introduction and Overview

A lack of *organized* data archiving may be just as serious a problem for scholarship and research as the outright disappearance of data. Perhaps the most significant contribution the library can make to the persistent discovery of scholarly information is to fix what we might call the "first mile" problem: the adequate identification and description of information before it is discovered by search engines and disseminated across the Web.

Addressing this issue by forging richer relationships between information resources and related content will help improve the precision of searching and provide more useful context on retrieval. This briefing discusses some options for solving this first mile problem by exploring the issues surrounding a new type of collaboration between scientists and research library staff. Mann Library has initiated activities related to the preservation, discovery, and sharing of primary linguistic and ecological research data through an NSF Small Grant for Exploratory Research. Concepts and standards from the Semantic Web community are also being applied to categorize and connect related data elements about people and their research drawn from central IT data warehouses, department data marts, and direct faculty input as a service to promote the discovery of faculty, their research activities, and project outcomes (vivo.cornell.edu).

Collaborations between Scientists and Research Library Staff

Two years ago, Cornell University received an NSF Small Grant for Exploratory Research to explore a new type of collaboration between scientists and research library staff. This grant focused on activities related to the preservation, discovery, and sharing of primary linguistic and ecological research data, especially at the level of individual research labs and departments. The initial planning grant involved Cornell's Language Acquisition Laboratory and the data collected over a period of more than 20 years from 20 different language groups. A supplemental grant allowed us to generalize our investigations by working with 20 investigators affiliated with a second research group, the Upper Susquehanna Applied Ecology Program, who were building additional datasets and models to leverage a 30-year history of climate and soil data from two major project sites in New York State.



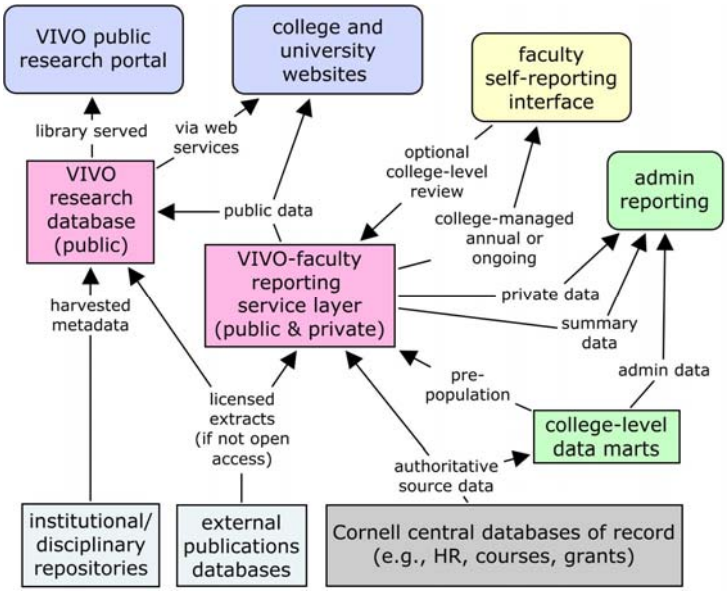
A recent proposal to NSF outlines a strategy to address research data access and archiving across additional disciplines, again focusing on the preparation as much as the preservation stage of the information lifecycle. The new proposal also emphasizes the need for data staging repositories with different access restrictions from the ultimate disciplinary or institutional repositories housing published data. Researchers need help managing datasets before publication, and in many cases the raw data need to be preserved to enable results to be replicated or later analyses to be conducted based on later tools or new hypotheses. Raw data may have additional restrictions due to confidentiality, incomplete analysis, or intellectual property reasons. A staging repository implies a limited scope and time frame for storage, and the library would offer such a service as part of an overall plan for publication and longer-term preservation.

VIVO: the Virtual Research Community

In addition to providing services to encourage metadata preparation and support collaboration among researchers, libraries can also help improve the discovery of research outcomes. Institutional repositories such as the Cornell Open Access Repository (dspace.library.cornell.edu) typically expose metadata to search engines such as Google, and faculty members' individual or departmental web pages are similarly indexed as isolated text information. However, by providing coordinated information about individual researchers, their grants and research projects, talks, courses, research descriptions, and impact statements, the library can provide users a richer landing page from which to explore and access research outcomes.

At Cornell, a specialized library service dubbed VIVO (vivo.cornell.edu) promotes the ubiquitous discovery of people and their research activities. VIVO leverages concepts and standards from the Semantic Web community to provide seamless access to diverse resources on the Web, supplementing the discovery power of search engines by presenting each displayed page in a rich context of related information and enabling the user to navigate across the research activities of Cornell unencumbered by administrative boundaries. Using an underlying, evolving ontology model, new information is categorized and connected to existing information while also indexed for local search and exposed to Google and other search engines. Information is stored as independent data elements – people, seminars, publications, departments, research areas -- rather than pages, supporting presentation not only as traditional faculty profile pages on department sites but also in different combinations in many other contexts. Direct relationships such as membership in an academic department, authorship of publications, or participation in a grant can be used to build up derivative "collections" by inference, providing useful targeted aggregations such as recent publications for a research area, grants by funding agency, or campus-wide directories of international activities or domain expertise – visible in VIVO or deliverable as web services to portals anywhere at Cornell.

The Library's strategy is to leverage automated sources of data, especially central university databases of record. We also coordinate with Cornell's individual colleges to query faculty for updated information on their academic and research activities, supplemented by tapping into PubMed, Biosis, and the ISI Web of Science citation databases. The effort has recently received funding support from the Cornell administration to extend beyond the life sciences to include all research, with an initial focus on gathering information on physical science and social science activities. Most important, we have developed close working relationships with central and college-level IT and administrative staff, who recognize the potential for VIVO leveraging their own work and providing additional visibility (and return on investment) for the information they develop and maintain.



For projects such as VIVO, the Library brings to the table credibility as a neutral and a competent information broker, in the game for the long haul. As data stewardship and full lifecycle information management become essential to research competitiveness and even mandated by federal funding agencies, research libraries have an important leadership role to play. Libraries will need new skills and above all new partners, within and beyond our own universities. The tasks will challenge both our own and our partners' traditional thinking, but in many ways the future of the research library will depend on acquiring, preserving, and delivering the data and knowledge essential to the research enterprise today.