Against the Grain

Volume 21 | Issue 2 Article 13

November 2013

CRL's Long-lived Digital Collections Project: Working to Provide Member Libraries Peace-of-mind

Bernie Reilly
Center for Research Libraries, Reilly@crl.edu

Recommended Citation
Reilly, Bernie (2009) "CRL's Long-lived Digital Collections Project: Working to Provide Member Libraries Peace-of-mind," *Against the Grain*: Vol. 21: Iss. 2, Article 13.
DOI: https://doi.org/10.7771/2380-176X.2556

CRL's Long-lived Digital Collections Project: Working to Provide Member Libraries Peace-of-mind

by **Bernie Reilly** (President, The Center for Research Libraries, 6050 S. Kenwood Ave., Chicago, IL 60637-2804; Ph: 800-621-6044 or 773-955-4545; Fax: 773-955-4339) <Reilly @crl.edu> *http://www.CRL.edu*

The creation and collection of massive amounts of digital data in the humanities, sciences and social sciences today is creating stewardship demands that cannot be met fully by traditional libraries and archiving organizations. During the past three decades large, new repositories of digital data have emerged to meet the needs of the scientists and researchers in the social sciences and humanities. Data stewardship is now undertaken by federal agencies, discipline-based consortia of scientists and researchers, supercomputer centers, universities, institutes, and for-profit corporations like **ProQuest**, **ExxonMobil**, and **Google**.

Some of the emerging data repositories have flourished and persisted; others have not. Funded by the **National Science Foundation** under its **Strategic Technologies for Cyberinfrastructure Program**, **Center for Research Libraries** recently initiated a two-year effort to examine some established, "long-lived" collections of data and digital resources, and to learn how and why it has been possible for some organizations to manage these collections for extended periods. Through a series of eight case studies the **CRL Long-Lived Digital Collections Project** will identify practices, strategies and mechanisms that have enabled those repositories to sustain massive data collections over time and in the face of significant changes in the economic, technology, and legal environments.

The following repositories are subjects of **CRL** case studies:

The Arabidopsis Information Resource

- **The Arabidopsis Information Resource** (**TAIR**) is an open access database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana. Data available from **TAIR** includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.

The Associated Press

- **The Associated Press** (**AP**) serves thousands of daily newspaper, radio, television and online customers with coverage in all media and news in all formats. It is the largest and oldest news organization in the world, its archives holding news text, photos, graphics, audio and video. **AP** operates as a not-for-profit cooperative with more than 4,000 employees working in more than 240 bureaus worldwide. **AP** is owned by its 1,500 U.S. daily newspaper members.

Chemical Abstracts Service

- A division of the **American Chemical Society**, the **Chemical Abstracts Service** (**CAS**) monitors literature from 9,500 chemistry-related journals. **CAS** abstracts and indexes (in English) documents originally published in any of 50 languages, including patents from 29 countries and two international patent organizations. **CAS** databases, accessible through **Sci-Finder** and **STN** interfaces, are used by scientists in over 100 countries.

ProQuest UMI Dissertation Publishing

- **ProQuest UMI Dissertation Publishing** has been gathering and publishing dissertations and theses from universities in the U.S., Canada, and beyond since 1938. Today **UMI** has over 700 university publishing partners, who generate more than 70,000 new graduate works each year. **UMI** also provides Web access to electronic dissertations and theses on both Open Access and subscription bases.

The General Social Survey

- **The General Social Survey's National Opinion Research Center** (**NORC**), founded in 1941 conducts "quality social science research in the public interest." **NORC's** clients include government agencies, educational institutions, foundations, other nonprofit organizations, and private corporations.

The Sloan Digital Sky Survey

- **The Sloan Digital Sky Survey** is a joint project of the **University of Chicago**, **Fermi National Accelerator Laboratory**, the **Institute for Advanced Study**, the **Japan Participation Group**, **Johns Hopkins University**, **Princeton University**, the **United States Naval Observatory** and the **University of Washington**. Funding for the project has been provided by the **Alfred P. Sloan Foundation**, the **SDSS** member institutions, the **National Science Foundation**, **NASA**, and the **US Department of Energy**.

National Center for Atmospheric Research: Earth Observing Laboratories

- The **National Center for Atmospheric Research** (**NCAR**) and the **University Corporation for Atmospheric Research Office of Programs** provide research, facilities, and services for the atmospheric and Earth sciences community.

U.S. Geological Survey

- The **U.S. Geological Survey (USGS)** provides scientific information to describe and understand the Earth; minimize loss of life and property from natural disasters; manage water, biological, energy, and mineral resources; and "enhance and protect our quality of life." The U.S.'s largest water, earth, and biological science and civilian mapping agency, the **USGS** collects, monitors, analyzes, and disseminates scientific information about natural resource conditions, issues, and problems. Much of **USGS**-generated data is maintained in the **National Geologic Map Database** (*http://ngmdb.usgs.gov/*), the **National Biological Information Infrastructure** (*http://www.nbii.gov/portal/server.pt*), and locally by the **USGS**.

The subject of one of the studies, **ProQuest UMI Dissertation Publishing**, has been a basic fixture of academic research for over fifty years. **UMI**, the largest repository of theses and dissertations produced at U.S. and Canadian universities, is a source of essential information to scholars worldwide. It is also, in effect, the official repository of dissertations and theses for the national libraries of Canada and the United States. A draft profile of the **UMI Dissertation Publishing** digital repository effort can be downloaded from the **CRL** site at *http://www.crl.edu/PDF/umi_dissertations.pdf*, and a summary of the report appears in the April 2008 issue of *The Charleston Advisor*.

A second case study subject is the **Associated Press**. The **AP** has long been regarded as the backbone of the American news industry. For over a century the organization's bureaus have produced much of the news reporting that has appeared in American and European newspapers. The **AP** is also a major source of news content for broadcasters, Web news sources, and aggregators like **LexisNexis** and **Factiva**.

Victoria McCargar, a specialist in archiving and electronic archiving, conducted **CRL's** analysis of the **AP** and its archiving efforts. The first draft of this profile is now posted at *http://www.crl.edu/PDF/AP_Profile.pdf*. There **McCargar** characterizes the central role the **AP** plays today in the news media world:

> In an increasingly connected world, the **Associated Press** is refining the meaning of the word "ubiquitous": The latest edition of its venerable *Stylebook* claims that half the world's population has access to **AP's** content every day.

continued on page 36

CRL's Long-lived Digital ...
from page 34

As the oldest and largest news-gathering organization in the world, **AP** content shows up on Websites, **Blackberrys**, cell phones, TV-screen news crawls and street-corner digital tickers — wherever up-to-the-minute breaking news is a sought-after commodity.

Therefore one can surmise that how, and how well, the **AP** organization manages its content is a matter of great consequence to libraries. For on this will be dependent whether or not this important information is available to researchers years and even decades from now.

The **Long-Lived Digital Collections Project** will generate and disseminate models, risk assessment tools, cost data, and metrics that can inform planning and prudent investment in Cyberinfrastructure by the **NSF** and other federal agencies, universities, scientific consortia and institutes, corporations, publishers, and other stakeholders across the spectrum of science, social science, and humanities communities.

The tools and information base developed will also inform **CRL's** continuing assessment and analysis of repositories and collections of digital content of interest to the **CRL** community. Subjects of **CRL** assessment range form dedicated preservation repositories such as **Portico**, **CLOCKSS**, the **Scholars Portal** and **HathiTrust**, to major digital resources maintained by commercial entities like **ProQuest** and **Readex**.

The case studies project is important to **CRL's** mission. A shared repository of primary source materials in paper and microform for U.S. and Canadian universities since 1949, **CRL** is now, in its seventh decade, beginning to actively support its member libraries' efforts to invest wisely and strategically in gaining electronic access to and archiving source materials for research. University libraries today are expected to provide scholars an expanding universe of source materials at a time when the resources available for building and managing collections are dwindling.

Moreover, in serving scholarly needs librarians face an often bewildering array of digital content, collections, and archiving services, with few metrics to guide their investment in those goods and services. With the explosion of the Internet as a channel for the production and delivery of electronic resources, the number and diversity of these products and services will only continue to grow.

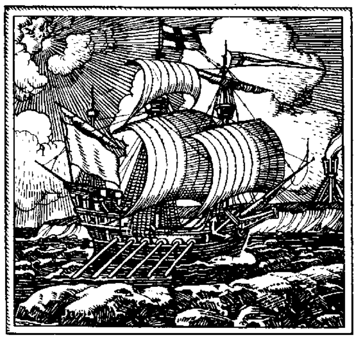
The new knowledge base about digital repositories that **CRL** is now developing will enable libraries to identify reliable, appropriate and affordable digital preservation services, and to ensure their communities persistent electronic access to critical resources. Years ago the **CRL** catalog served a similar purpose: aside from being a means through which researchers could discover **CRL** collections, the catalog functioned as a tool for collection development at member libraries. The catalog was, and to some extent still serves as, a point of reference for individual library decisions on the acquisition and retention of local holdings in the area of journals and newspapers. The **ICON** database, established in 1999, provides a similar registry of the foreign newspaper holdings of **CRL** and a group of partner institutions. This information has enabled **CRL** member libraries to concentrate their human and financial resources on developing and maintaining a larger set of local holdings.

Similarly, the expanding knowledge base generated by **CRL** analyses and assessments of digital resources will enable **CRL** libraries to focus their resources on obtaining access to databases and digital collections that support research and teaching comprehensive and on supporting digital repositories that ensure persistent access to those materials.

During the next few years **CRL** efforts will concentrate largely on news, journals, archives, and other primary source materials that support international studies in the humanities, sciences and social sciences. **CRL** information, analysis and services are intended to enable its members to progressively retire tangible collections in these areas and replace them with secure, affordable and persistent electronic access.

This project is part of the ongoing construction of a **CRL** information base to support investment in digital resources and preservation by its community of member libraries. *The Charleston Advisor*, **Global Resources** workshops and forums, and the **CRL** Web and collaboration spaces are the venues through which **CRL** and its members are building and sharing this information base.

Unchartered Territory: Building a Network for the Archiving of Geospatial Images and Data



by **Julie Sweetkind-Singer** (Head Librarian, GIS & Map Librarian, Branner Earth Sciences Library & Map Collections, Stanford University, 397 Panama Mall, Stanford, CA 94305; Phone: 650-725-1102) <sweetkind@stanford.edu> *http://www.ngda.org*

Librarians working in the realm of geospatial information routinely live in a 20% world. When librarians collectively talk about what systems will be set up to handle content, this typically means books and journals, the 80%. I have found this to be true dealing with either paper-based maps and aerial photography or digital data and imagery. Procedures for paper-based content are well formulated. It is what libraries have learned to do over the last few hundred years. But, lifecycle management of digital content is not fully understood, especially when dealing with that 20% of non-standard content. Over the last four years, librarians and technologists at **Stanford University** (**Stanford**) and the **University of California, Santa Barbara** (**UCSB**) have worked together to learn how to address the challenge of digital lifecycle management, especially focusing on the last component in that cycle, long-term preservation. As is often the case, what we thought we needed to do to understand long-term preservation of digital geospatial data and imagery was the tip of an iceberg that was much larger and more complicated than we imagined.

Funding the Project

In December 2000, the **United States Congress** authorized nearly $100 million to fund a national effort to "set forth a strategy for the **Library of Congress** in collaboration with other federal and non-federal entities, to identify a network of libraries and other organizations with responsibilities for collecting digital materials that will provide access to and maintain those materials."[1] The program was to be administered through the **National Digital Information Infrastructure & Preservation Program** (**NDIIPP**). **Congress** mandated the money be used to develop policies, protocols, and strategies for the long-term preservation of "at-risk" materials. **Stanford** and **UCSB** were in the first round of funding announced in September 2004, which included eight awards totaling nearly $14 million. **Stanford** and **UCSB** proposed the creation of the **National Geospatial Digital Archive** (**NGDA**). The goals of the **NGDA** were to create a national federated network committed to archiving geospatial imagery and data, to investigate preservation strategies, to collect "at-risk" content across a spectrum of formats, and to develop policy agreements governing retention, rights management and obligations of the partners. Along the way, we have had to build two archival storage systems, create collection development policies, content provider agreements, partnership agreements, a format registry, and an interface to federate the materials through an online catalog. This paper will focus on the non-technical parts of the work we have done.

The **NDIIPP** agreement clearly stated that these awards were specifically for archiving

continued on page 38