Cornell University Library Projects of Potential Interest to Computing and Information Science

November 8, 2013

This document provides background information on some of the current projects at the Cornell University Library (CUL) that may be of interest to faculty and researchers in Computing and Information Science. Most of these projects live within two CUL units: Digital Scholarship and Preservation Services (DSPS), which is the service owner for arXiv, leads our collaboration with HathiTrust, and focuses on developing and providing services for Digital Humanities; and the Library IT Service Group (CUL-IT), which provides software development and operational support for all Library IT services, and leads the effort on VIVO and Linked Open Data. Both groups collaborate in support of Research Data Management and Digital Preservation.

In addition to the current specific projects, the Library has previously worked with CIS students in developing library-related software applications, including a selection and ordering tool (http://www.oclc.org/worldcat-selection.en.html) and a Library iPhone app. We are always interested in new opportunities for joint development of library-related software.

Text Mining Corpuses

We have access to two significant text corpuses that could be used for research in data mining: HathiTrust and arXiv.org. CUL is a member of HathiTrust, and we are the fourth largest contributor of digitized print volumes, with 433,870 deposited so far. CUL has been managing operations and support for arXiv since 2001. Beginning in 2010, the Library has been developing a long-term sustainability model for arXiv. We currently have support from over 170 institutions around the world, and a major five-year grant from the Simons Foundation to support the development of a sustainable arXiv business model. arXiv is a major open-source preprint repository for Computing and Information Science, with Joe Halpern serving as the chief disciplinary moderator.

HathiTrust

The HathiTrust Research Center (HTRC) http://www.hathitrust.org/htrc is a collaboration of Indiana University and the University of Illinois, along with the HathiTrust Digital Library. Its mission is to enable “computational access for nonprofit and educational users to published works in the public domain and, in the future, on limited terms to works in-copyright from the HathiTrust.” The research center does not contain any digital books. Instead, books in HathiTrust are indexed and stored in SOLR within a robust computing environment. Currently, there are 3,492,345 public domain volumes available. Computational access is available through a portal that has canned algorithms, and scholars can create collections and then run algorithms on these. Results are most frequently visualized in a tag cloud, but results can also be downloaded and repurposed elsewhere. There are APIs available, documented on the wiki (below). Algorithmic access to works in copyright has not yet been opened.
The HTRC manages a [wiki](http://wiki.htrc.illinois.edu/) for participants. This is a great source of information. Note on the landing page you will see the [elists](#) also managed for users. These are helpful, low-traffic lists that users of the portal can use to notify the HTRC of bugs in the portal and to share code/processes and results with each other. There is also a monthly conference call for portal users. Right now CUL staff are participating in an effort to assemble educational material that helps scholars understand the utility of the portal, hopefully so they can fairly evaluate the ways in which they could envision adopting it for their own research. This is difficult territory; humanists are not necessarily used to thinking algorithmically.

**Cornell and HTRC/computational access:**

- DSPS offered summer internships as part of an emerging Digital Humanities (DH) program run by Mickey Casad. Five graduate scholars in the humanities joined us to learn about various aspects of DH. Michelle Paolillo spent a day with them, first orienting them as to digitization and trying to raise awareness about Optical Character Recognition issues and the various challenges that arise from addressing digital books in an algorithmic fashion, and then exploring the HTRC with concrete examples of their choice related to their work.

- Oya Rieger, Mickey, and Michelle met with David Mimno, a new professor in IS. He is interested in our work and will likely be more in touch as he gets graduate students that are interested in addressing the needs of scholars. We are hoping to match up the scholars with needs with the folks that can enable access.

- Ed Baptist, professor of history, has expressed interest in some in-copyright works that are compilations of slave interviews. Michelle wrote up the case for HTRC. These will likely have to wait for the unveiling of the Sloan cloud for in-copyright material.

**arXiv.org**

arXiv.org has a collection of almost 900,000 articles, with about 70,000 new articles being added each year. The data spans the time period 1991 through to the present and includes almost complete coverage of certain disciplines (especially high-energy physics). Nearly half of the articles have more than one revision and all revisions are retained. The corpus of metadata and articles are freely available for research ([http://arxiv.org/help/bulk_data](http://arxiv.org/help/bulk_data)) although access is somewhat cumbersome. It would be interesting to set up a convenient copy of arXiv data to give Cornell students and researchers easier access.

A complete set of web access logs have been maintained since 1994 and include a huge amount of usage information that has seen only limited investigation. Privacy is an important concern, so this data cannot be made freely available. Cornell investigators Peter Frazier, Paul Ginsparg and Thorsten Joachims have a current NSF project ([http://www.nsf.gov/awardsearch/showAward?AWD_ID=1247637](http://www.nsf.gov/awardsearch/showAward?AWD_ID=1247637)) that is based around mining of arXiv log data combined with machine learning to experiment with new interfaces to the literature ([http://my.arxiv.org/arxiv/](http://my.arxiv.org/arxiv/)).
There has been Cornell work on overlap and plagiarism on the arXiv corpus, starting with graduate work [http://arxiv.org/abs/cs/0702012](http://arxiv.org/abs/cs/0702012) and being continued by Paul Ginsparg with development of tools to detect plagiarized submissions and annotate them in arXiv (e.g. [http://arxiv.org/abs/1311.1764](http://arxiv.org/abs/1311.1764) where the comment "arXiv admin note: text overlap with arXiv:1310.7829 by other authors" was added by Paul Ginsparg as a result of overlap detection).

**Research Data Management Service Group (RDMRG)**

The Research Data Management Service Group (RDMRG) [http://data.research.cornell.edu](http://data.research.cornell.edu) is a collaborative, campus-wide organization that links Cornell University faculty, staff and students with data management services to meet their research needs. Sponsored by the University Librarian and the Vice Provost for Research, the RDMRG is a collaboration of the Cornell University Library, the Center for Advanced Computing (CAC), and the Cornell Institute for Social and Economic Research (CISER). The RDMRG coordinator is Wendy Kozlowski, a CUL metadata librarian.

The RDMRG’s broad range of science, policy, data, and information technology experts provide timely and professional assistance for the creation and implementation of Data Management Plans (DMPs), and help researchers find specialized data management services they require at any stage of the research process, including initial exploration, data gathering, analysis and description, long term preservation, and access.

After the first IT@Cornell day last September, Wendy Kozlowski had a meeting and discussion with David Bindel ([bindel@cs.cornell.edu](mailto:bindel@cs.cornell.edu)) about his experiences with NSF DMP creation and review, pros and cons of providing example DMPs as reference, and specific concerns he feels are relevant to CIS faculty. He specifically mentioned that for computational research, there is a question of how to support reproducibility of results, and whether codes are data to be shared or potentially commercially valuable instruments to be protected. He felt that the RDMRG hosting a discussion of these issues might be helpful in the context of data management.

In spring of 2013, after receiving numerous requests for Electronic Lab Notebook (ELN) cloud software over time for both classroom and research use, Cornell conducted a pilot of the LabArchives cloud-based ELN software. This ELN pilot leveraged a campus collaboration across Academic Technologies, the University Library, research data management, faculty, research labs, and the medical school to evaluate an ELN system. Feedback from users at the end of the pilot indicated that the product should be offered as a regular campus-wide service in the Fall of 2013.

There has been a long standing need for digital tools to manage both science instructional labs and research labs. As the data management and reporting requirements evolve for higher education research, the need for systems to help to manage the information for both instructional and research needs continues to grow. Cornell currently has over 450 registered LabArchives users, and IT@Cornell’s Academic Technologies Group and the Cornell Library are working together to provide technical, instructional and best-practices support for the product.

The RDMRG has also provided Data Literacy training to graduate students at Cornell, both as a specific course, a six-week, one-credit course in Data Management through the Department of Natural
Resources, and through data management workshops for graduate students throughout this semester, on topics such as data organization, data sharing and data documentation.

Web Archiving

The Library is also beginning work on web archiving for support of research and scholarship in specific collecting areas. We have started collecting local websites related to Fracking in New York State: https://archive-it.org/collections/2788. If there are areas of Computing and Information Science where it would be useful to create web archives of important and fragile web resources, we would be happy to discuss this with FCIS faculty.

Digital Humanities and Digital Media Preservation
There are several projects related to digital humanities that may be of interest to CIS faculty. Here are brief reports on three of them: Preserving Digital Art; Digitization Projects; and Improving Access to Publications. In addition to these specific projects, we are also working with Academic Technologies on issues related to MOOCs, including how to make MOOC-created materials available for courses to be taught on campus, whether as videos or as other materials, and choosing texts and other materials from outside sources that we can license for use by students taking MOOCs.

Preserving Digital Art
The Rose Goldsen Archive of New Media Art contains a large collection of digital art, including interactive software, videos, and other digital materials – typically on floppy disks, CDs, DVD, and various video tapes and related media. We have a major effort underway to preserve access to the art itself, even as the hardware on which it originally ran becomes obsolete and unavailable. Here’s a brief summary of the technical work undertaken so far on the project:

The project team has run through all of the disks and made notes as to how they work and render on modern hardware, and on our one iMac (running OS 9). We also pulled from the catalog all of the system requirements as they were notated from the original artwork and took a look to see how many were from the System 7 era, OS 8 era, Windows 95, Windows 98, etc.

From there, we’ve been proceeding on a few parallel tracks. One track is imaging all of the disks in the collection. We’re using ISOBuster, which is a tool that creates RAW disk images for every CD. The majority of the disks imaged without any problem, but there are a minority that had unreadable sectors. We’re considering how best to address these. One way may be to image our copy of the original and see if that had any issues; another is to swap out and try our alternate CD drive. (We acquired Plextor drives that came recommended from AVPS, one of our advisors.)
Another track is to see how these works render in various emulation environments. For the Mac based works, we're working with SheepShaver and BasiliskII; for the Windows-based works, we're using Wine. Initially, we ran a few works through emulation and made note of anything that behaved differently from the original. Some of these issues we might be able to fix; others might not. Our next venture may be to try to take a sampling of works, each spotlighting its own particular challenge, and identify the workable emulation environments for each. One idea we've been considering is what we've been calling "The Jukebox" -- a way to select a work and a corresponding emulation environment so that a user can interact with the work in a contemporaneous emulated computer environment.

Finally, we have started to think about what files go into CUL Archival Repository for each work. For example, the disk image and any attendant emulator files needed to run them, plus technical information about what files were originally on the disk, and other important information.

**Digitization Projects**

We continue to explore issues within Digital Humanities via our grants with the College of Arts & Sciences. One project in particular, [http://warburg.library.cornell.edu/](http://warburg.library.cornell.edu/), pushed the envelope more than the others, and was a collaborative effort with CU Press and external scholars. The challenge in these projects is not just the digitization of the original materials, but transforming them into meaningful online exhibits that allow users to fully engage with the content. You can see a list of all the faculty grant projects here: [http://dcaps.library.cornell.edu/initiatives/asgrants/awards](http://dcaps.library.cornell.edu/initiatives/asgrants/awards).

**Improving Access to Publications**

Currently, all journals in HathiTrust are indexed in such a way that users **cannot** gain easy access directly to an article through a linkable table of contents. There is limited capacity for access through full-text search, but it requires some intermediate steps. HathiTrust currently has no plans to provide direct access at the article level. Our project has the capacity to potentially leverage the scans and metadata that are already available in HathiTrust to provide direct article access in a simple but truly useful way to researchers. By working with the already existing metadata, our goal is to create indexes at the article level and expose the article view in a subject-appropriate delivery site. This is a sustainable repurposing of the valuable content in the HathiTrust, and we believe that it will also contribute code and development to the HathiTrust Research Center. *The Cornell Veterinarian* is a highly respected academic journal that was published from 1911-1994 by the New York State Veterinary College at Cornell University. Currently, *The Cornell Veterinarian* is only available digitally via HathiTrust. A valuable journal for historical research, HathiTrust’s current indexing and access is not conducive to deep scholarship in a single resource like the *Cornell Veterinarian* because it is impossible to view a hyperlinked version of an issue’s table of contents. This introduces obstacles that prevent the scholar from navigating from search results directly to the desired article within the journal.

**Digital Preservation Network**

Ensuring the long-term preservation of digital resources is a huge challenge, due to the technical, organizational, and physical risks involved. As part of our effort to address these issues, the Library has joined together with a number of peer institutions to create the Digital Preservation Network (DPN, pronounced “deepen”) [http://www.dpn.org](http://www.dpn.org). Here’s a brief description from the website: “The Digital
Preservation Network (DPN) was formed to ensure that the complete scholarly record is preserved for future generations. DPN uses a federated approach to preservation. The higher education community has created many digital repositories to provide long-term preservation and access. By replicating multiple dark copies of these collections in diverse nodes, DPN protects against the risk of catastrophic loss due to technology, organizational or natural disasters.”

DPN now has working groups addressing the legal issues, business model issues, and technical issues of creating a network for long-term preservation. While we do not believe that DPN will provide a complete solution for all our long-term preservation needs, it will be an important component of our strategy going forward.

**VIVO and Linked Open Data in Support of Research and Scholarship**

VIVO ([http://vivoweb.org](http://vivoweb.org)) is an open community, an information model, and an open source semantic web application supporting the advancement of research and scholarship by integrating and sharing information about researchers and scholars, their activities, and their outputs both within a single institution and across broad, distributed networks. VIVO is fundamentally interdisciplinary; it enables and promotes the discovery of research and scholarship across traditional boundaries of geography, organization structure and type, academic or clinical or applied domain, technology, language, and culture. There is a diverse set of activities associated with the VIVO project, across federal agencies, academic institutions, professional societies, and data providers, as well as a variety of efforts with the Semantic Web and ontology development communities. VIVO was originally funded by Cornell University and the National Institutes of Health (U24 RR029822) and is currently a community-supported incubator project under the DuraSpace umbrella. You can find out more about the VIVO project at [http://vivoweb.org](http://vivoweb.org).

VIVO was originally developed by the Cornell University Library in support of the Life Sciences Initiative in 2004. Through a $12 million grant from NIH (2009-2012), the VIVO software and ontology were transformed into a robust system for ingesting data about researchers and research from multiple sources within an institution, and making this information available as Linked Open Data on the web. You can see Cornell’s VIVO system at [http://vivo.cornell.edu](http://vivo.cornell.edu).

VIVO is both a software system and a standard way to describe structured information about researchers and their entire academic context (expressed as a formal OWL ontology). Other researcher profiling systems have made their own information available as VIVO-compatible Linked Open Data. This allows the creation of systems to do search or analysis across many institutions. You can see a nice example of this at [http://research.icts.uiowa.edu/polyglot/](http://research.icts.uiowa.edu/polyglot/). Developed by Dave Eichmann at the University of Iowa, CTSASearch allows you to search across roughly 125,000 primarily bio-medical researchers at a number of institutions.

The VIVO software and ontology are now in use by over 100 institutions around the world, including the US Department of Agriculture, Duke, Penn, Virginia Tech, the University of Florida, the University of Colorado Boulder, the National Science Library of Germany, the University of Melbourne, China’s National Science Library, and the Food and Agriculture Organization.
CUL is working in collaboration with a number of other institutions to extend VIVO to provide better support for research data, research resources, characterizing accomplishment in the humanities and performing arts, scholarly information resources, and describing the workflow and process of science.

**OSTP’s Open Access Requirements for Federal Agencies Distributing over $100M in Funding**

The new OSTP policy on open access to both research publications and research data announced this past February (http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research) presents an interesting opportunity for CUL and our peer libraries to be part of an overall solution for ensuring open access to research. The Association of Research Libraries (ARL), the Association of American Universities (AAU), and the Association of Public and Land-grant Universities (APLU) have proposed a Shared Access Research Ecosystem (SHARE), where individual libraries would join together into a federated ecosystem to provide open access to research publications and data (http://www.arl.org/news/arl-news/2773-shared-access-research-ecosystem-proposed-by-aau-aplu-arl). Existing repositories such as Cornell’s eCommons institutional repository and arXiv.org could be part of this federated solution.

CUL has a strong interest in promoting open access to research and scholarship, as demonstrated by our support for efforts like the Cornell Open Access Publication Fund (http://www.library.cornell.edu/compact/coap.html). We will be working with ARL, AAU, APLU, and our peer research libraries to provide fully open and academically friendly solutions like SHARE to the emerging federal mandates for open access.