A Newsletter of the UCSD Libraries Digital Library Program

Digital Letters

Spring 2005 Issue Number Seven

Letter from the Editor

This issue of *Digital Letters* showcases two online reference services that are expanding service points for library users and experimenting in collaborative reference. Elisabeth Leonard tells us about CDL's experimental chat reference pilot in which users can access a librarian live, from their desktop and be guided through their browser, to the most relevant resources. Another service, Radical Reference, while not a service exclusively for UCSD users was founded by 2 UCSD librarians—Shinjoung Yeo and James R. Jacobs. They explain to us how this service attempts a significantly new approach to providing reference service in real time on the street. Our main article in this issue is on data services. I interviewed James A. Jacobs about this little known area of librarianship and uncovered why we should be concerned about preservation and access to this type of data in the future.

~ Trish Rose, Image Metadata Librarian, UCAI

The Raw Data Librarian: preserving today's data to be tomorrow's information

TR: Jim, you're a data services librarian and you work in the Data, Government and GIS unit of the Social Science and Humanities Library at UCSD and more specifically you oversee the Social Science Data Collection (SSDC). Forgive my ignorance but I'm not familiar with this area of librarianship. What does a data services librarian do?

JJ: I do similar work to what other librarians do but I work with users who do quantitative analysis of numeric data. "Data" in this context means censuses, public opinion surveys, social surveys, economic time series, and so forth. I help users find raw data for analysis and I do bibliography to the extent that I'm selecting and acquiring data. The library doesn't have a separate fund for data so I work with bibliographers who control specific funds. So for instance, if we have someone in sociology who needs some data that costs money to acquire I'll go to the sociology bibliographer.

TR: You've also used the term "data archivist" to describe what you do and I'm wondering is this the equivalent of a data services librarian?

JJ: Not quite. I can best describe the difference between a data archivist and data services librarian by making an analogy with

(Continued on page 2)

Radical Reference: an open-source organization

There is a common belief that technology is value-free, but many social and political activists are challenging this notion by creating and employing technologies that are imbued with their value system of justice, equality, and community. One such group in the library world is Radical Reference (RR) (http://radicalreference.info).

RR is a volunteer-run collective of library workers (librarians and library staff and students) that provides reference services and information access to independent journalists, activists and the general public via its web site and on the street at various political events (dubbed "street librarians"). RR was initially launched in July, 2004 to assist and support the many activists converging to protest at the 2004 Republican National Convention in New York City. The service was so well received by activist communities that it has continued and expanded its services to include fact-checking workshops, and skill-sharing sessions at the 2005 ALA conference. Today, there are over 150 volunteers across the United States with a variety of professional backgrounds and the ability to provide services in ten languages.

In order to provide its services, RR has consciously chosen to employ open source and/or non-commercial software and web hosting. For example, Drupal, a content management system, is used for managing the web site. The system, written in PHP with a MySQL database backend, includes easy to use tools for blogging, content creation, site design and organization, and user management. Volunteers communicate with each other and our users via email, listserv, GAIM Instant messenger (IM) client, Lightning Bug, a tool to help track online collaboration and reference, and, during political events, Txtmob, a web-based text messaging software, is relied upon to provide synchronous communication between street librarians and home support volunteers. Additionally, RR web hosting is provided gratis by Interactivist, a non-profit organization that supports groups working toward issues of social justice.

Technology is not value neutral; rather it is ideological. In order for RR to achieve its mission of information activism toward social justice, it is crucial to implement technologies that reflect its inherent values as well as the user groups that take advantage of those services. In the name of convenience, libraries often overlook the underlying philosophy and principles of the technologies they employ. RR provides a good example for how an organization can infuse its technologies with its organizational philosophy.

~Shinjoung Yeo, Reference Librarian, SSHL and James R. Jacobs, Government Information Librarian, SSHL

Raw Data Librarian *(continued from page 1)*

the paper world. A data librarian manages a collection of data in the same way libraries manage collections of books – we select, acquire, and organize the collection and provide access to and service for the collection. A data archivist is someone who has the additional responsibility of providing the "copy of last resort" of data files. Historically, the UCSD Libraries began as a data library and was not trying to be a data archive (i.e. we didn't have anything unique in our collection). Increasingly, however, our collection consists of data not available elsewhere so we are beginning to take on the responsibilities of an archive.

TR: Are any of the data sets we archive produced at UCSD?

JJ: We do have some data files produced by faculty, but very little, so far. We're also adding value to local San Diego Census Data by making the files easier to use and that involves producing new versions of data files and those versions don't exist anywhere else. We do have data gathered in San Diego. I would very much like for the library to be able to offer faculty a permanent home for their data and would guess that we will some day. Most of our users are in economics, sociology, political science, and urban studies. We also get users from outside the social sciences, researchers in epidemiology for instance who need demographic or health data.

TR: If someone is providing access to hard science data would they also be considered a data services librarian?

JJ: Yes, though it would depend on what services they provided. Historically, researchers in the social sciences had to preserve and reuse their data because the data were expensive to collect, while those in hard sciences collected their own data rather than using someone else's. But that's changing because of funding policies and the nature of research. This is particularly true with data sets that are large and observational in nature such as space surveys or oceanographic surveys. So to answer your question, yes, whether providing access to social science or biological science data the data services librarian would still be performing the same types of services for quantitative data sets.

Providing service for datasets is a little more like providing service for files in a traditional archive in that they are both collections of primary resources rather than research outcomes. It's also the case that a user often does not need an entire dataset. For example, the government produces data files that record cause of death along with other demographic information such as age, sex, place, etc. A user may not want all of the millions of records in that data file. They may only want 500 records pertaining to a particular cause of death. Unlike statistical summaries you see in books, datasets contain unsummarized raw data suitable for analysis and researchers often need specific subsets for their analyses.

TR: So essentially, you help people pull out subsets of data from these collections?

JJ: My model of service is to remove barriers between our users and the data they need for research. While I do sometimes sit with people and help them create subsets, more of my time is spent providing a web site and other tools that make users self-sufficient and help them create their own subsets.

TR: Jim, what is the technology you're using to do this?

JJ: It's old technology. It's mostly Perl scripts and some commercial software. Together, they enable users to create subsets and convert those subsets into any of more than twenty different formats so they can quickly load the data into the statistical software of their choice.

My model of service is to remove barriers between our users and their research

TR: Where do you get most of your data sets?

JJ: In the past, most of our data came from ICPSR (Inter-university Consortium for Political and Social Research) at the University of Michigan. It's the largest social science data archives in the U.S. UCSD has been a member since the early 1980s and as a member all UCSD students and faculty can download data from ICPSR without charge. In the past, users had to go to the librarian to request the data and ICPSR would send us tapes and we'd add the data to our local collection so users could then get the data easily. Now, the user can download the data directly from ICPSR's web site. In our local collection, we're focusing more on datasets that aren't easily available from ICPSR or elsewhere. For instance, between 1990 and 2000 the San Diego Association of Governments (SANDAG) created population estimates for small areas of San Diego county but when the 2000 census data were released they took those estimates off their website because they assumed no one wanted them. But, for a student in Urban Studies, those data are not just valuable, they are essential to examining the changes in San Diego over time. So we offered to archive the data and SANDAG sent the data to us. Now, we're the only public site providing access to these data. That kind of activity is more typical of what we're doing now.

TR: So this illustrates the importance of the preservation role of a data archive. I guess I was struggling with the analogy of a data repository being like a traditional archive that houses personal papers. Personal items are one-of-a-kind and can only be stored in one place but raw, numeric data seems much more like monographs or serials whose content can easily be duplicated and distributed widely. But I can see how the commitment to store and preserve data for the future is analogous with traditional archives.

(continued on page 3)

Raw Data Librarian *(continued from page 2)*

JJ: Yes. That's exactly it. It is not so much a question of how many copies there are or might be in the world, but a question of who takes responsibility for ensuring long-term access to the "copy of last resort" and who takes responsibility for the completeness and accuracy of the metadata that must accompany the data files.

TR: On the SSDC website, I noticed several pieces of metadata that accompany each collection of data including things like a collection title and information about the data file format. I was curious whether that metadata conforms to any standard?

JJ: That's a really good question. The simple answer is yes. We started doing this at UCSD before there were international standards and we set up our own metadata format. Now, there is an XML standard for data documentation called DDI, which stands for Data Documentation Initiative. We are not using DDI yet, but I've proposed that we do so.

TR: When you get the data collection how much metadata comes with it and how much do you have to create?

JJ: It's basically two packages. First is the data file (simply a text file filled with numbers) and second is the documentation that describes the layout of the data file and how to interpret the numeric codes used in the data file. Ten years ago, that documentation was a printed book called a "codebook." Today, the amount of metadata that come with these files varies widely. Mostly, today we see traditional codebooks but in PDF format accompanied by files that work with specific statistical software packages. We're just beginning to see documentation marked up in XML using DDI. That's good because it is not tied to any software. DDI is designed to last over the lifecycle of the data set. For example, a researcher can use DDI to document a survey instrument (such as a public opinion poll) and the how the data were collected. Then when the data are turned over to a data library, the librarian can document that particular instance of the data and add subjects, keywords, and so forth. The next researcher who uses that data file and combines it with a new data collection can add metadata

(continued on page 4)

CDL's Chat Reference Pilot

UCSD patrons are now able to ask librarians for help via instant messaging software. UCSD is collaborating through a Common Interest Group (CIG) of HOPS to pilot the chat service (http://sshl.ucsd.edu/chat/). The goals of the pilot as established by the CIG are to provide excellent service, to add a service point, to extend reference hours, to show the value of a collaborative reference service, and to determine the usefulness of chat reference to the participating libraries. By the end of the pilot, the CIG will evaluate the effectiveness and usefulness of the service and will recommend to HOPS if the pilot should be extended, altered, or concluded. Participating libraries are: UC Irvine, UCLA, UC Merced, UC Riverside, UCSD, and UC Santa Barbara. The service runs from January 9th through March 24 and is available from 6-9 p.m. Sunday through Thursday.

The pilot service uses software called 24/7 Reference (http://www.247ref.org). The software allows the patron to chat with a librarian without having an instant messaging account, to send and receive web pages, to receive a transcript of the session (if the patron has provided a valid email address), to provide feedback on the pilot through a web survey, and to receive follow-up emails from a librarian when the session has ended.

A chat session is initiated once a patron has filled out the web form (the form asks for affiliation, name, email address, and the question they'd like answered) and clicked on the submit button. The librarian hears an alarm and sees a flashing image on the computer screen, notifying them of the patron's presence. Once the librarian joins a chat session, they can see what web browser a patron is using, what IP address the patron is coming from, and the information from the web form. The software allows librarians to set up scripts, providing for the preloading of frequent responses . This saves the librarian time in searching for answers

The software also permits a librarian to monitor multiple calls from each campus simultaneously, to refer questions to other librarians in the pilot who are live monitoring chat, to email the transcript to another librarian for later follow up, to send files and images to the patron, and to co-browse using the patron's browser in order to guide them through their searches. Administratively, it enables the collaborative effort to be managed effectively. Through the software, for the campuses that have their own individual day time service, the shift is made from their individual service to the pilot project. Heather Tunender of UC Irvine, the project coordinator, gathers usage statistics and forwards the transcripts to the participants for quality control.

So what kinds of questions are being asked? Questions range from "Where in the Library could I find research materials on the Meiji Restoration and its relationship with the Samurai of Japan?" to "I think there's a bird trapped in the air ducts on the 6th floor of Geisel. Can someone help?" Questions are coming mostly from UC affiliates, and while the patrons are mainly undergraduates, graduate students, faculty and staff, are also asking questions. The average length of a call is 10 minutes, but have ranged from one minute (patrons don't always stay to ask the question or have the question answered) to an hour. In the first 2 weeks of the pilot, UCSD patrons accounted for 33% of the callers!

In the few weeks it has been available, the pilot is already demonstrating its value. Not only is it reaping the expected benefits for patrons, who are soundly responding that they find the service useful, but the librarians involved are also reaping unexpected benefits. Because of having input from librarians who are unused to our individual web sites, we are receiving feedback about our web pages – what works and what is confusing. We are learning each other's resources and services, and are making connections with each other. Will the pilot service have enough of a benefit to justify the time staffing it? Only time will tell. Until then, if you have a question and it's at night, give us a call.

~ Elisabeth Leonard, Head, Reference, Outreach & Instruction SSHL

Raw Data Librarian *(continued from page 3)*

about how the new dataset was created. Instead of being static, the documentation starts at the inception of data and goes throughout its lifecycle including its use and reuse.

TR: How do you get the researchers to document their data?

JJ: Data librarians are working with the manufacturers of the primary statistical software packages (e.g. SAS, SPSS, STATA) to incorporate the DDI standard into their software. UCSD Libraries support this through membership in an organization that maintains and promotes DDI, the DDI Alliance. Many statistical software packages have modules for collecting data so we're hoping that creation of metadata will be an almost automatic result of the data collection process. The goal is to get researchers thinking about their data as a long-term project. Some researchers are very proprietary about their data and others just don't see a need to share. But that's changing slowly. One of the goals of the data library community is to promote the sharing and preservation of data.

TR: This idea feeds into the issues you and Charles Humphrey brought up in your article on preserving research data (Jacobs, James A., Humphrey, Charles "Preserving Research Data." Communications of the ACM. Volume 47, Number 9 (2004), Pages 27-29) which is the data community's narrow focus on data publishing and depositing information to the neglect of data preservation and "long-term equitable access and use"

JJ: Yes, that's the point we were trying to make in the article: that the "publishing" model carries with it the idea that "it's my stuff". In other words, the model implies that data is just like a published article and once published no one can plagiarize it. But an article and data are different – an article is your writing, your reporting, your report of your results. It's static. The data file is a collection of observations and can be used and reused – and should be! In the social science community, we have a history of data being shared and reusable.

TR: What about the issue of depositing vs. preserving?

JJ: We wrote the article because we took issue with some of the ideas in an earlier *CACM* article. One was the idea that deposit is the same as preservation; another was the idea that researchers should be responsible for preservation. We believe that depositing doesn't ensure that data will be preserved and that researchers don't want to be librarians. An institutional repository can be a good thing but repositories are not automatic solutions to preservation. If something deposited is not adequately documented, for example, you can get access to all the bits but you won't be able to make sense of them. Archiving is more than just accepting deposits - it requires particular skills. As an academic community and as a society we should take measures to make sure there are repositories and they're well funded and staffed by people with the right skills.

TR: Jim, where do people gain the right kind of training and skills to run the repositories? Are they being trained in library school?

JJ: I don't know about all library schools, but I do know that the newer librarians that take the class I teach don't seem to have gained data preservation skills there. Unlike paper preservation, we don't yet have a body of knowledge about digital preservation that ensures us stuff will be available in five years. I teach a class in data management with Charles Humphrey and Diane Geraci, and in that class we do two things: we teach some specific technical skills but since the tools are always changing we also teach generally what you would want to use the tools for and how you go about preserving something for access. Specifically we cover the role of metadata; migration of data over time; responsibilities of data librarians vs. distributors, researchers, and government agencies; and finally, the connection between service and preservation.

In the digital world, preservation, access, and service are all very interlinked

In the digital world, preservation, access and service are very interlinked. But the terminology we use can be confusing. When we talk about "dark archives" it's easy to think this unlinks access and service from preservation. We can't ignore the access issue when we preserve digital information. If we don't take into consideration how digital information will be found and used, then our preservation won't be successful.

Because we get so many students in our classes from different environments we can't be prescriptive. What we lay out is the range of things that need to be done and we let the students brainstorm ways that might work in their particular environments. Solutions are dependent on funding, user base, what you can do, the politics of your organization, your IT infrastructure, absolutely everything.

TR: Could you elaborate on the recommendation in your article for addressing the future needs of data archives by investing in national archives?

JJ: Institutions have collaborated in preservation through their memberships with ICPSR and other collaborative data archives, but we need to take that model to the next level. Do we have enough archives doing the right thing? Is the funding secure? Where is it coming from? The long-term way to solve these problems is to think of research as a cycle and to think of data use and preservation at the beginning of research. Funders should require not only the collection of data but the sharing of that data as well. NSF is beginning to require this, as is NSDL. We all need to be looking at the bigger picture.