The California Digital Library: *Implications for Social Science Data Files Collections*

by Daniel C. Tsang[*]

The recent establishment of the California Digital Library creates an unprecedented opportunity to bring social science data to users in a more user-friendly format as well as making it available to a much wider audience.

The California Digital Library was established at the University of California in the middle of last fall, in effect creating a tenth campus library, this one entirely virtual. In practice, it is based at the UC Office of the President headquarters in Oakland, California, it's main manifestation in its initial year appears to have been acquiring licensing agreements with computerized databases of academic journals in the fields of science and technology. But with an anticipated infusion of $3 million from state coffers this coming year, plus another $1 million from the University of California itself, the CDL, as it is called, will soon be looking for new disciplines to cover, and new territory to conquer. According to press reports, California has had a budget surplus of $4.4 billion which can hopefully be partially allotted to the libraries.

Social Sciences are a likely future field of attention for the CDL, and UC's data archivists and librarians recently met with the CDL's newly installed Collections Officer to discuss matters of mutual interest, including ways of collaborating in possible future joint endeavors.

This paper, then, is part of a continuing, fairly new process of rethinking how the UC system collects and provides services to social science material in digitized formats.

Since the CDL bills itself as the 10th campus library collecting everything in digitized form, there might have been initial fears that collecting of digitized materials would end on the individual campuses that make up the UC system. While legally and constitutionally the University of California is one state-wide system, in actual practice, there is almost no collection development (in terms of library resources) at the system level, except for licenses to databases negotiated system-wide and made available via Melvyl, the UC catalog and database gateway; and large purchases made in the past through "shared purchases" system-wide, largely made up of big ticket items and large microform sets. Each campus is strongly independent, each headed by its own chancellor and strong faculty senates; campus libraries are no different, each with its own collecting focus, depending in a large part on the research and instructional needs of the faculty on the particular campus.

Despite being three years in the planning stage, the actual creation of the CDL may have caught many campus librarians by surprise. Any initial fears that collecting of digitized materials on campus would be displaced by the CDL collecting the same stuff for all campuses soon gave way to the realization that there was no way individual campus digitized collections would disappear, nor would collection development cease. In fact, since data archives and data collection development vary across campuses, meeting individual needs not necessarily duplicated elsewhere, the likely model that will prevail is one more of collaboration with the CDL than one of the CDL usurping and displacing campus collection patterns. While it is relatively easy (and familiar) for the CDL to negotiate licenses over journals in the social sciences, it is a different kettle of fish for the CDL to enter the field of collecting (and making available) raw data that are the products of social science research. It is not just a matter of licensing, but also one of making it available to users — and which users need to be defined — in an appropriate format.

How the CDL is structuring its own collection development may be instructive here. Currently, within its science and technology collection focus, the CDL has selected biotechnology and computer science as the two areas where acquisition of digitized information in formats other than electronic journals or monographs would take place. This is described in a March 8, 1998 announcement on its listserv (CDLINFO-L) as providing a "laboratory for learning and planning for future CDL collections," and for "gradually [adding] other types of content" beyond electronic journals or monographs. This is an ongoing experiment, and it is too early to report any results. But what is clear is that the CDL intends to acquire datasets in biotechnology and computer science initially, although what it would acquire in particular is not apparent.

In addition, the CDL plans to structure its collection in

three "tiers", namely, Tier 1, material funded, in whole or in part, by the CDL; Tier 2, material funded by two or more campuses; and Tier 3, material that is paid for by only one campus. For the last tier, this implies the CDL would provide a gateway, on its Web site apparently, to material on that campus, even if, as is likely, only users affiliated with that campus can use the material. For tier 2, then, material would likely not be accessible to those outside the campuses that funded it; and for tier 1, since the CDL funded the material, it would presumably be made available to most or all campuses (a campus can opt out of a particular acquisition).

Its current collection principles are that priority should be given to digital format acquisition of those resources which offer economies of scale by benefiting the most faculty and/or students both locally and system wide. Also, electronic materials should be selected based on increase of access to the installed base of UC library collections and build on the investments already made by the university in digital resources. If and when the CDL enters the field of social sciences collecting and social science data in particular, a major issue appears to me to be the one of defining who would be the potential user population. On a traditional campus archive or data library, UC data archivists and librarians have generally acquired material only for the use of faculty and students on that campus. In addition, as I argued in an earlier IASSIST conference paper[1], collection development, in UC's as elsewhere, has generally, been reactive rather than proactive; i.e., datasets are acquired in response to specific requests, rather than, as with books, collected in advance of specific request, and often without anticipation of actual immediate use (e.g., through book approval plans). Since that paper, however, with the advent of the main source of social science data now distributing, in effect, its entire newly acquired collection every quarter or so in the from of Periodic Release CD-ROMs, libraries are acquiring data without any selectivity, in actual practice, and certainly not in response to any specific request. In addition, with the ICPSR's entire archive much easier to retrieve (with the demise of distribution on round tapes), data archivists can potentially replicate much of the collection on their own campuses, although as of yet, there is no mirror site to the ICPSR archive yet established. Such a mirror site might be one area the CDL might fruitfully consider, especially since its mission appears not only to serve the campus-based academic community but also the community at large, although it is still unclear how that would work in practice.

The CDL, after all, is called the California Digital Library, not the University of California Digital Library, strongly implying it is collecting digital material to serve the entire state, i.e., all the people of the state of California. That was the quid pro quo that apparently was necessary for the state legislature to fund the CDL. In announcing the creation of the CDL last fall, UC President Richard C. Atkinson spoke of creating "UC's library without walls." It would be a library allowing "scholars of all ages and interests to range worldwide in their quest for knowledge, using the Internet, the World Wide Web and a computer."

If the target population is the scholarly community within UC and beyond, that would be one thing. Libraries are used to collecting for scholars; the CDL could very well mirror ICPSR's archive (after becoming a full-fledged member like the other UC's that are members) by paying them enough money so that they would not think that they are losing money. But since there is no physical campus associated with the CDL, what does this mean? Or is the CDL membership to take the place of the individual campus memberships? That does not seem likely, given that existing campus archives and data collections have constituencies they have nurtured and served for years; they are not likely to disappear, at least not without a fight. But if the CDL promised access to all ICPSR data, and provided front-end interfaces that facilitated the extraction of variable-level data (it would need to write the software etc.) of selected datasets, would that not make local service points less necessary, or even impractical? The countervailing argument, of course, is that with all secondary analysis of data, it is not sufficient merely to make the data available; there has to be a variety of services (metadata access; interpretation of metadata; statistical consulting; etc.) that, because computer setups vary across campuses, and even within, are best handled at the local level.

In addition, the CDL can perhaps better negotiate licenses at the system-level with data vendors that provide data, such as economic time series, of interest across the UC system.

But given its mandate to serve the state, the CDL and its pioneering vision, the CDL could very well do more than just provide access, even improved access, to what currently exists. The CDL could well get involved in one or more large-scale digitization projects, funded by industry and government, to archive and make accessible datasets previously not readily available, such as government data at both the state, county and municipal levels. The challenge here is for joint partnerships with communities local and state; what is sorely needed is a collaborative effort to make sure that government data does not go the way of the main frame and that they are archived and preserved, and eventually made accessible to users.

As the CDL develops, one potentially controversial area involves intellectual property rights. Who owns the research that faculty have invested their time in? The University is now arguing that it does, and that faculty who sign away rights to articles, for example, are just making commercial journal publishers much more rich. Administrators are now wondering why a university should

have to pay exorbitant fees to access the journal output of their own faculty, just because a commercial journal published it. Richard Lucier, the CDL's head, argues that scholarly publishing must change, and that universities must on their own, compete with the industry and "publish" on line the scholarly output.

With data however, is it likely that individual faculty, even at UC, will be willing to deposit their research data at the CDL, if the University insists that it must? The University recently revised its policy on research; now, even if faculty use a dataset gathered elsewhere for secondary analysis, it must register with local Institutional Review Boards. Local boards could well insist that faculty deposit data thus gathered (or originally collected data for that matter) with the CDL, or the campus data archive. Right now there is no such provision or mandate, not least because researchers are unlikely to be willing to part with their data, and there is no common understanding of who owns what. If a faculty member leaves, he or she is allowed to take research data he or she has gathered; thus far, I am not aware that the university has insisted on ownership. But a case could well be made, given that every employee is, upon hire, made to sign away most of his or her patent rights to the University.

Involvement in large-scale projects is especially likely, and necessary, if indeed, the user community stretches beyond the scholarly community. If indeed the vision is to let anyone access the information (and since the CDL is called the California Digital Library, not the University of California Digital Library, as originally envisioned), that suggests that anyone, "of any age", would have access to the library. That would well be a mammoth task, devising a dataset (or more) that would be useful to such a mythical user.

There are many more issues one could raise, not least that of digital archive maintenance, authentication, and dataset updating, as well as version control. At a minimum, the CDL could provide a union list of what exists in existing campus archives and collections, providing some bibliographic control to an existing situation that is anarchistic at best. But I see it as doing more than that; it can best make the use of data more appealing, by providing the necessary tools to access data from the hard-to-find to those most popularly requested. As such the CDL can help those of us in the data archive community by making, and educating, more people to be sophisticated data users and consumers.

In conclusion, the CDL is unlikely to be the sole digital repository for California of social science data, but its creation and expansion will likely spur collaborative efforts with existing collections and archives as well as create new ways to provide improved access to these collections.

1. Daniel C. Tsang, "Academic Libraries and Collection Development of Nonbibliographic Machine Readable Data Files," IASSIST Quarterly, volume 12, number 3, Fall 1988, pp. 47-55.

* Paper presented at the IASSIST Conference, May 21, 1998, at Yale University, New Haven, Connecticut.