

NSF Workshop on Post-Digital Libraries Initiative Directions

Christine L. Borgman, UCLA

Personal digital libraries: Creating individual spaces for innovation

1. Problem to be solved: Enabling individuals to create, manage, and preserve information in personalized, idiosyncratic, innovative environments

People need to seek, use, re-use, create, maintain, and preserve information in support of their work and life activities. In a world of print and hard-copy documents, the process of creating new works involves copying, rewriting, cutting, pasting, and rearranging prior works of one's own and of others, while adding ideas, data, analysis, effort, and other forms of value (Lessig, 2002). In the digital world, creators are dependent upon digital tools to perform tasks that once could be accomplished with scissors and paste, pencil and paper calculations, and later with photocopiers and other reproduction devices (Borgman, 2000, 2003). Individuals need a "place" or a "space" in which to assemble and manipulate information resources for their own purposes, with flexible tools that they can adapt to their practices, skills, habits, and artistry.

Personal DLs represent an alternative framework for digital libraries. Most DL research assumes that a professional team will design, develop, deploy, curate, and manage DLs over the short and long term. DLs will be a primary source of content for scholarship and teaching, and for business, government, and personal activities. Large distributed systems are most effective in providing generalized services, especially as DLs scale upward in the size of repositories and in the size of user communities. In contrast, individual users need flexible, tailored services, and they need to draw content from multiple DLs. Putting DL capabilities in the hands of individuals will allow DL repositories to scale while still providing individuals the tailored tools and services they need to do innovative work.

The real promise of digital libraries, as claimed in the Cyberinfrastructure report (Atkins, et al, 2003), the CLIR Alliance for the 21st Century Library (Henry, 2003), and the NSF proposal for this workshop, is that these information technologies have the potential to transform the conduct of disciplinary research and to foster new areas of investigation at the boundaries of existing disciplines. Fostering such innovation requires that people have a set of flexible tools and services to gather information from multiple sources, including digital libraries, and to manipulate them for their own purposes. The advantage of digital documents over print is that they are "malleable, mutable, and mobile" (Bishop & Star, 1996). Personal digital libraries can be much more than repositories; they can facilitate malleability, mutability, and mobility of information resources. Thus the next research front is to design tools and services that will enable individuals to create and manage their own personal digital libraries (PDLs).

2. Limits on current practice: Monolithic systems, describing data rather than uses

The digital libraries of today (and the near future) tend to be monolithic systems that serve large distributed communities. These are critical mass technologies that become more valuable as their repositories grow in size. Their strengths are also their weakness: by being large and general, they are not easily tailored to individual uses.

Another key limitation of current practice is that metadata in today's DLs is more likely to describe attributes of a document than it is to describe the multiple uses to which it might be put. This is not surprising, as it is generally easier to describe what you have than it is to anticipate how others may choose to use it. Traditional information management approaches do attempt to anticipate uses of the information by describing the objects, providing certain data about their origins (e.g., author, source, date), and by describing what a document is "about." (Baca, 1998; Svenonius, 2000). Archival science approaches are more explicit in recognizing that the uses of information tend to change over time (Gilliland-Swetland, 1998). Records created for one purpose at one time mean something different to their creators than to those viewing them in a later historical context, for example. If DLs are to foster creativity, people need more ways to identify information they might find useful for new purposes.

3. New approaches and evidence for success: Studying users and uses; designing personal digital libraries; improving IR methods based on behavioral models

Research on personal digital libraries can advance the fronts of several core areas of DL research, including user behavior, information management, and information retrieval. Several themes are identified here and some evidence is provided for each.

- Empirical study of the users and uses of digital libraries

Surprisingly little research exists on the users and uses of digital libraries. Two complementary research fronts should be addressed here. One is to examine the long history of research on information needs and uses to determine what theories and methods can be applied to digital libraries. Among the relevant themes in this literature is that people have highly individualized ways of seeking and using information. Another theme is that stage theories of information seeking (e.g., Kuhlthau, 1988a, 1988b, 1991) acknowledge that people do not move monotonically forward through the stages; often they take a step back to reconsider prior moves, and then move forward again. Information needs, queries, or uses, are not static; they evolve over time. As people learn more about a topic, they ask questions differently, following some paths in more depth and abandoning others.

The second research front is to apply these theories and methods to DLs. While considerable research exists on information retrieval systems and online catalogs, most of it concerns text-based systems in library contexts. Today's DLs are larger repositories of more heterogeneous resources and offer more advanced services than the IR systems of recent years. Research on usage of the World Wide Web is growing, but these studies tend to have small samples and are dependent on the vagaries of search engines, making the results difficult to compare or generalize. Now that DLs are being deployed widely in academic and research settings, we need real data on users and uses that can be used to design a new generation of systems.

- Leveraging digital libraries to support multiple user communities for multiple uses.

Digital libraries are very expensive to build and maintain. Making digital libraries more cost-effective requires that they serve multiple users for multiple uses. Doing so is a non-trivial research challenge that we are addressing in two current NSF projects, the Alexandria Digital

Earth Prototype (ADEPT)¹ and the Center for Embedded Networked Sensing (CENS)². User studies are essential to the iterative design process in both of these projects.

The question in ADEPT is how to make information resources that are produced and described for one purpose, such as geographic research, also usable for another purpose such as teaching geography. In research-oriented databases, geographic data may be described by attributes such as location (latitude, longitude), place name, source, date, and instrumentation. But for teaching purposes, an instructor may wish to find an image of any river that exemplifies a certain type of erosion, and set criteria such as the degree of color contrast and size of image (e.g., large enough to display clearly in a 200-seat classroom). Another geography instructor may find that same image useful for explaining a different geographic concept. In this example, the users possess substantial disciplinary knowledge of geography. If we broaden the user community of the DL to include students, then the metadata requirements are even more extensive, as these users have far less knowledge of the subject field.

Making one DL useful to multiple categories of users for multiple types of uses imposes additional design requirements. Additional metadata may be needed to describe documents for multiple applications. Different types of retrieval algorithms may be needed, some to serve sophisticated domain experts and some to serve domain novices such as students.

- Research and development on personal digital libraries

Early research on uses of digital libraries is confirming findings from prior IR studies that individual users are highly idiosyncratic in their information habits. Their expectations from DLs vary widely, as does their use of digital data once obtained. We are finding in the ADEPT project that no matter how rich a repository we might build, users want capabilities to extract resources into a personal space where they can manipulate them. They also want to be able to add resources from their own collections and to combine and manipulate these resources (Borgman, et al, 2000).

- Enhancing information management via personal digital libraries

People already are overwhelmed by the amount of information they have to manage on their personal computers, not to mention what else is available to them on other servers. The hard disks on today's personal computers hold as much data as the large bibliographic databases of only a few years ago. As we gather more non-text files, the scope of personal information management is likely to grow exponentially. Today we often cobble together our bits of content from multiple sources using the software in which the final product will exist (e.g., a web page, a power point file, an MS word file), but we have few good options to manage those manipulated pieces for future use. Often the tools are cumbersome to use and the task is delegated to graduate students or other assistants. Once the assistants depart and the software is updated to a later version, reclaiming the information product may be nearly impossible.

Preserving the original digital content in one or more forms is itself a massive management problem that is beyond the scope of this short position paper. Others at the workshop are likely to address this problem in depth.

¹ National Science Foundation grant no. IIS-9817432, Terence R. Smith, University of California, Santa Barbara, Principal Investigator. <http://is.gseis.ucla.edu/adept/>

² National Science Foundation, Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator. <http://www.cens.ucla.edu>

Personal digital libraries offer multiple opportunities to improve information management:

- Individuals should be able to download content from large repositories into their personal DLs. This should be a focus of interoperability research. PDLs should support the life cycle of information creation, use, re-use, and preservation or disposal.
 - Real innovations occur when people can assemble information from a variety of sources, in a variety of types, often from a range of disciplines, to create their own new ideas, frameworks, models, questions, and so on. PDLs should offer a rich set of tools and services to facilitate this process.
 - PDLs will contain a heterogeneous mix of content from a variety of sources. Some of it will be created by the PDL owner / user, such as authored documents, images, drawings, datasets, weblinks, bookmark files, spreadsheets, powerpoint files for talks and lectures, etc. Other content such as journal articles, texts, or messages may be captured from external sources. The quality of metadata for documents in a PDL is likely to vary widely. Documents captured from external sources may contain rich metadata from multiple metadata schemes, while locally created documents may contain little more than a file name, date, and type (e.g., the software through which it was created). PDLs should allow people to capture available metadata and to add their own metadata that describes their uses for it, no matter how idiosyncratic their practices may be. This will allow them to manage their own resources better and to locate content for re-use.
 - PDLs should enable individuals to upload their metadata to the common DL from which an object came, thus creating community-based metadata descriptions.
- Information retrieval based on recognition rather than recall

Cognitive psychologists distinguish between two fundamental types of memory: recognition and recall. Recognition occurs when you see something familiar, while recall requires that you remember something and are able to articulate it. Most information retrieval depends upon recall skills – the user has to describe what he or she wishes to retrieve. Browsing depends more on recognition skills – looking around until you find something of interest that you recognize as useful. But most browsing still requires that the user describe a starting point.

Recall approaches are most effective with text-based systems because words can be spelled and matched against a corpus of documents. Describing images and sounds is vastly more difficult, both for the indexer and the retriever. Recall approaches also depend on the availability of rich metadata or on sufficient amounts of text to match.

Recognition approaches are likely to be much more effective in large digital libraries of the future and in personal digital libraries. This is true for at least two reasons: One is the proliferation of non-textual documents in digital form (still and moving images, sound). We need ways to summarize non-textual data in ways that people can recognize easily, such as the video “fast forward” experiments reported at the most recent JCDL (Wildemuth, et al, 2003). The second is the lack of metadata on which to base recall algorithms. Metadata is expensive to produce and cannot serve all of the unanticipated uses of any given document. Individuals are unlikely to invest the effort in rich description of everything they add to their own PDLs. They need ways to summarize and to browse their repositories quickly and easily.

Our research on the Science Library Catalog, beginning in the late 1980s, showed that children could learn to use a recognition-based catalog of science materials quickly and easily (Borgman,

Gallagher, Hirsh, & Walter, 1995). While we did not test the system on adults, other communities found the user interface to be very appealing. Our audiences at research presentations invariably asked why other IR systems did not use this approach. In our current ADEPT research, geographers explain the difficulty of describing the images they seek, and how heavily they rely on serendipity and on “knowing it when I see it.”

4. What difference will success make? Innovation, personal productivity, and leveraging DL investments

Successful outcomes from the research agenda for personal digital libraries outlined here will have social and economic payoffs. If individuals can select, organize, use, and re-use digital content in new and effective ways, the promise of digital libraries to foster innovation and creativity may be achieved. Secondly, innovation could be achieved with higher productivity. We spend far too much time today learning to use too many single-purpose tools, none of which truly supports information management. Personal digital libraries will be integrating tools that let us manage our creative resources with less overhead than the tools of today. Thirdly, personal digital libraries have the potential to leverage the substantial economic resources being invested in building large repositories of digital content. Leverage will be achieved in several ways. One is to enable the same content to be used by multiple users for multiple purposes. Another is to make large DLs and PDLs interoperable, such that individuals can download data for local manipulation, and can upload tagged data to share both content and metadata.

5. How much will it cost and what are the milestones? Manageable costs and recognizable milestones

Projecting precise cost figures is as much an art as a science, and I will not attempt to put a dollar figure on this research agenda. The cost should be reasonable, because much of this research can be accomplished in concert with extant research programs. Teams of behavioral science researchers and technologists can study information practices as part of DL-building projects, feeding the results iteratively into the design process, for example.

The design of personal digital libraries can be conducted as independent projects or as part of larger DL projects. Combined projects will be useful in determining interoperability requirements such that content can easily be transferred between PDLs and multiple large, community-based DLs.

Research is needed on metadata requirements to make content usable by multiple audiences and to make content readily re-usable. Related research is required on user-generated metadata. Studies in both of these areas will contribute to larger research questions in metadata, ontologies, knowledge organization, and information management.

Case studies are needed in multiple disciplines to determine what behaviors and requirements can be generalized across user groups and what requirements are individual and group-specific. We know that individual practices are idiosyncratic, but we need to know more about how to support these practices, and which ones are most critical for usability.

Research on evaluation of digital libraries will contribute to the study of user behavior and to the design of personal digital libraries. As identified in a recent NSF-EU DELOS workshop, research is needed on metrics and measures that can be applied in local contexts and also on testbeds to allow comparisons between digital libraries (Borgman et al, 2002).

References

- Atkins, D. E., et. al., *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure*, January 2003, available online at http://www.communitytechnology.org/nsf_ci_report/.
- Baca, M. (ed.). (1998). *Introduction to Metadata: Pathways to Digital Information*. Los Angeles: Getty Information Institute.
- Bishop, A.P.; & Star, S.L. (1996). Social informatics for digital library use and infrastructure. In. M.E. Williams (ed.), *Annual Review of Information Science and Technology*, 31 Medford, NJ; Information Today, pp. 301-401.
- Borgman, C. L. (in press). The invisible library: paradox of the global information infrastructure. *Library Trends*, 51(4), Special Issue on Research Questions for the Field.
- Borgman, C.L. (2002). *Final report to the National Science Foundation. Fourth DELOS Workshop. Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics*. Hungarian Academy of Sciences, Computer and Automation Research Institute (MTA SZTAKI), Budapest, Hungary, 6-7 June 2002. Grant IIS-0225626. http://www.sztaki.hu/conferences/deval/presentations/final_report.html
- Borgman, C. L. (2000). *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: The MIT Press.
- Borgman, C.L., Gilliland-Swetland, A.J., Leazer, G.L., Mayer, R.; Gwynn, D.; Gazan, R.; & Mautone, P. (2000). Evaluating digital libraries for teaching and learning in undergraduate education: a case study of the Alexandria Digital Earth Prototype (ADEPT). *Library Trends*, Special Issue on Assessing and Evaluating Digital Library Services, 49(2), 228-250.
- Borgman, C.L., Gallagher, A.L., Hirsh, S.G., & Walter, V.A. (1995). Children's Searching Behavior On Browsing And Keyword Online Catalogs: The Science Library Catalog Project. *Journal of the American Society for Information Science*, 46(9), 663-684.
- Gilliland-Swetland, A. (1998). Defining Metadata. In M. Baca (ed.), *Introduction to Metadata: Pathways to Digital Information*. Los Angeles: Getty Information Institute.
- Henry, Charles H. May, 2003. Alliance for the 21st Century Library: Proposal to Monitor and Analyze Transformations in Academic Disciplines. Proposal to the Council on Library and Information Resources. <http://www.carnegie.rice.edu/>
- Kuhlthau, C. C. (1988a). Longitudinal case studies of the information search process in libraries. *Library and Information Science Research*, 10(3), 257-304.
- Kuhlthau, C. C. (1988b). Developing a model of the library search process: Cognitive and affective aspects. *RQ*, 28(2) 232-242.
- Kuhlthau, C.C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5), 361-371.

Lessig, L. (2001). *The Future of Ideas*. The future of ideas : the fate of the commons in a connected world. New York: Random House.

Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, MA: MIT Press.

Wildemuth, B., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., & Gruss, R. (2003). *How fast is too fast? Evaluating fast forward surrogates for digital video*. In: C.C. Marshall, G. Henry, & L. Delcambre (eds.), Proceedings of the 2003 Joint Conference on Digital Libraries, May 27-31, 2003, Houston, TX. Los Alamitos, CA: IEEE Computer Society.