Promoting Digital Scholarship: Building the Environment

A One-Day Seminar on Goals and Priorities
Sponsored by the Council on Library and Information Resources

November 28, 2007

Co-Chairs: Gregory Crane and Amy Friedlander

In 2006, the amount of information created, captured and copied was about 161 exabytes, or roughly 3 million times the information in all the books ever written. And sometime, perhaps very soon, the amount of information will exceed available storage.[1] These estimates have been provided by IDC in a study commissioned by information technology storage giant EMC Corporation, and should therefore be viewed with appropriate skepticism. But it hardly scares the horses to acknowledge a deluge of data. While IDC's study embraces the full range of data, from Voice Over IP (VoIP) and video signals to scholarly communications, report after report from the academic communities acknowledge both the challenge of managing the data and the promise that access to digital information and systems affords present and future teachers, students and researchers. This symposium looks at the promise by focusing on services, data and systems with the investigator at the center: What environment do we need to enable research? What questions can we imagine when information at a scale well beyond any we have known becomes commonplace? And therefore what functionalities should that environment possess to enable investigators to do their work?

We begin with the investigator. How do researchers behave? What are their styles and modes of research? And what questions do they ask and what kinds of analyses do they do now and might they do in a future environment where data is plentiful? But the data are now different. Large scale digitization produces new considerations for data management and other services. The information is voluminous yet relatively undifferentiated, comprising representations primarily of text but also non text works (sound, moving images, statistical and geospatial data). What post-digitization processing is appropriate to support different communities of scholarship, and how is the threshold of processing determined? Are the levels of mark-up really specialized to different communities, and would such data analysis tools be relevant to, for example, analysis of many kinds of digital information serving different communities. Librarians already know that historical sources can be important to multiple communities, not only to traditional historians. How does the need to support multiple communities with a single, albeit enormous, dataset affect the way the data are managed? Finally, having developed requirements based on the investigator and requirements derived from the data, what systems should be built? What are the components of those systems? Where are they situated? And assuming that the systems as well as the resources will be distributed across institutions and organizations, what are the social and governance structures that will also be required to enable information and communication to flow smoothly?

From the perspective of these investigators, the future environment is composed of three broad rubrics: services, the mechanisms or functions that enable investigators to do their work; data, the raw material of research; and systems, the technologies, agreements and structures that

[1] John F. Gantz et al., *The Expanding Digital University*; an IDC White Paper, sponsored by EMC (March 2007),, 3-4.

provide the services. These may be both social and technical. Clearly, there are feedbacks among all three topics. Nevertheless, we begin with the model of the investigator since the enterprise is ultimately devoted to human inquiry and the services, systems and resources required to foster it.

1. Services. In the 1970s and 1980s, applications offered increased analytical capacity, and the expansion of networking in the 1990s offered access to and distributed search across local and non-local collections. Clearly, faculty understand that the digital environment offers them more ability to do more things with fewer intermediaries. Thus, while many librarians believe that their consultative role in the research process will survive into the digital environment, a recent study of faculty attitudes by Roger Schonfeld and Kevin Guthrie of Ithaka, found precisely the opposite: Faculty members appreciate the role of libraries in providing research and teaching resources but expect to become increasingly dependent upon electronic materials – not upon librarians.[2]

    a. How do we model behavior? And what does that process tell us about the requirements for data management
    b. What are the questions and what functionalities are required to address the questions that become possible in the context of massive but relatively undifferentiated data?
    c. What skills are required to exploit these functionalities? And who educates the students?

2. Data and data management: Mass digitization has changed the terms of the discussion. The first generation relied on conversion of texts with careful attention to mark-up so that the resulting object faithfully captured the key attributes of the source. This approach has been powerful and has resulted in substantial progress in standards development and certain types of analysis. Mass digitization, however, has created an enormous reservoir of relatively undifferentiated text without the careful mark-up of the previous generation of conversion but greatly expanding the sheer volume of material potentially available. Reservations about the quality of the conversions have already been voiced; these range from the technical quality of the representations to the utility for sophisticated analysis. So the challenge to the community is three fold:

    a. What is the usability of the data that results from mass digitization projects?
    b. What level of additional mark-up (if any) is necessary to answer those questions? And how specialized would those tools me? For example, would automated tools that marked up an edition of *Hamlet* be relevant to other types of information, perhaps, travelers' accounts that would be relevant to historical reconstructions of prior terrain and landscapes?
    c. What is the level of management that is required to ensure that the data are available now and in the future? Is culling of digital collections to remove redundant copies reasonable? What is preserved? The initial output or the output that has been subjected to greater scrutiny or both?

3. Systems. In a relatively short time, we have moved from mainframes to laptops to a mix of handhelds, notebooks, laptops and desktops with and without connectivity. Powerful applications and increased storage and processing power together with a robust, high

[2] Roger C. Schonfeld and Kevin M. Guthrie, The Changing Information Services Needs of Faculty, *Educause Review*, July/August 2007.

capacity network mean that the end user faces choices and opportunities almost unthinkable a generation ago except, perhaps, among the most starry-eyed techno-enthusiasts. Now, expectations of IT support strain resources on university campuses, and Joel M. Smith and Jared L Cohon, respectively vice president/chief technology officer and president of Carnegie Mellon University, write, "difficult decisions are required to select which expectations to meet and which to disappoint."[3]

a. What systems are required to provide the functionalities?
b. What is the minimum set? And what are specialized to communities of interest or practice?
c. Given distributed systems that enable broad access to diverse resources, what governance systems are required to (i) support the research yet respect rights holders concerns and (ii) minimize the free rider problem?

Many workshops and symposia have already made progress in thinking through aspects of these questions. The purpose of this event is to take stock of what has been done and assemble a set of findings, priorities and recommendations. The first half of the day will be given over to presentations. In the afternoon, we will open up the discussion to all participants with a wrap-up at the end of the day by CLIR's president, Charles Henry.

[3] Joel M. Smith and Jared L. Cohon, Information Technology and the Research University, *Issues in Science and Technology* (Fall 2005).

Program

8:30 am – 9:00 am:	Welcome, Amy Friedlander Introduction: Why is there a problem? Gregory Crane
9:00 am – 10: 00	Services: What do we want to do? What are the functionalities of an environment to support research in texts/research in the humanities? Wendy Pradt Lougee: How do we study researchers? And what behaviors should the environment support? Gregory Crane: What do we do with a million books?
10:00 – 10:15	15-MINUTE BREAK
10:15 – 11:15	Data: What data management issues does mass digitization pose including (i) post processing to support more specialized or advanced analysis; (ii) support for use by multiple communities with multiple needs; and (iii) long term stewardship? Will Thomas: What does data need to be like? Joyce Ray: How are these collections managed?
11:15- 11:30	BREAK
11:30- 12:30	Systems: What systems are required to meet the needs of investigators who seek to use these data? What are the components? And how do they fit together? José-Marie Griffiths: What can libraries do? Okan Kolak: What can industry do?
12:30 – 2:00	LUNCH
2:00 – 4:30	Open Discussion: What are the priorities? What are the requirements? What is shared? Where is the research?
4:30 – 5:00	Wrap up Charles Henry, CLIR