Digitization and Preservation
Elings & Hurley; 4/2/08

Overview:
Digital information resources have entered the scholarly communications chain as both an important information source and product of research activities. This vast, heterogeneous collection of digital content poses significant issues and challenges for the library domain in attempting to manage, preserve and provide access to these materials. How has this situation developed and how to we find our way through it?

Digitization: Digitization of materials held in academic libraries has been underway for over 15 years. These activities have been largely driven by grant funding and, while processes and equipment have become less expensive over time, these endeavors still remain costly. Recent mass digitization efforts (e.g. Google Books, Open Content Alliance) have forced a re-evaluation of digitization processes and standards, encouraging a fresh look at how this work is accomplished. How will these new perspectives enable academic libraries to digitize their holdings more cost-effectively as grant funding diminishes? When will we move away from digital “projects” and begin forming digital programs instead? What do we need to get to this more sustainable position? A recent CLIR report (Preservation in the Age of Large-Scale Digitization, by Oya Y. Rieger February 2008) framed the issues surrounding these activities, stating that libraries need to reassess their digitization and preservation strategies and look to more agile and flexible practices, policies, and partnerships in order to move forward successfully.

Preservation: Academic libraries have embraced access services for “licensed” digital content. Within this area, nascent attempts at addressing digital preservation have been initiated for e-journals (e.g., LOCKSS, Portico). There are, however, large bodies of digital information required for scholarship that fall outside the domain of “licensed” information (e.g. digitized special collections, e-books created in our Google/OCA mass-digitization projects, government & other web-based documents, digital audio/video and scientific/social science datasets created by UC Berkeley researchers).

Traditionally, research libraries have taken on preservation responsibilities for print materials. The question for the Berkeley Library is now “how far do we push digital preservation responsibilities into our core mission?” What practices, policies, and partnerships do we need to pursue in order to preserve our growing digital holdings? What's our role in supporting Portico, LOCKSS and the CDL Digital Preservation Service? Do we need a strategy for preserving e-books? How aggressively do we pursue preserving web sites and their related materials? Should we work directly with UC Berkeley faculty to preserve digital assets (e.g., images, audio, video, and datasets) created in their research and instruction activities? Does the Library have a role in creating and preserving student portfolios?

National Trends
- Academic libraries are re-evaluating digitization processes and standards in light of mass digitization.
There is increasing pressure for reliable preservation repositories for digital content. Efforts to develop and certify “trusted digital repositories” are moving forward but are complicated and will take time.

Digital projects are slowly moving toward digital programs, with permanent funding and staffing.

There are numerous cooperative digital preservation efforts underway both in the US and Europe (e.g. NDIIPP (US), PLANETS (EU), LOCKSS, CLOCKSS, MetaArchive Cooperative, Chronopolis, NEDLIB).

**Local Issues**

Starting facts/assumptions

- Digital collections, including born-digital, are vast, diverse, and growing fast.
- Sustainable & effective digitization and digital preservation efforts require infrastructure.
- The Library has limited in-house image digitization capacity within our Digital Imaging Lab (DIL) and has also outsourced digitization projects for other formats
- UC Berkeley is a member of Portico, LOCKSS and the CDL Digital Preservation Service.
- Other Digital Preservations Services exist (e.g., OCLC’s Digital Archive, Internet Archive’s Archive-It).
- The Library maintains good inventory control over its locally held digital assets in GenDB (our content management system) and they are ready to be “preserved.”
- The CDL has established a nascent digital preservation service
  - It includes a:
    - repository service for “locally managed” digital assets.
    - web archiving service (production release planned for this June).
  - “potential customers” include UC libraries, other campus departments, non-UC organizations.
  - The plan is to recharge (at some level) for preservations services.
    - Recharge prices for various types of customers are not yet decided. CDL will run a pilot to better understand costs before prices are set.
- UC Berkeley’s IS&T Data Group has embarked on a data archiving project called Media Vault Program (MVP) to manage and preserve digital content from faculty and cultural institutions (e.g. Phoebe Hearst Museum, History of Art, Environmental Design, Museum of Paleontology). As of this writing, the MVP has no funding for preservation efforts. The Library is discussing partnership opportunities with MVP.

**Opportunities for New Directions & Starting Points**

- Partnerships and cooperative efforts need to be encouraged and developed to share expertise and costs. This is a theme that runs through many examples in this paper, especially for digital preservation (CDL, Portico, LOCKSS).
  - **Starting Point:** Continue discussions with the Media Vault Project. Identify other partnerships that could benefit the Library.
- Should *creation of digital materials* be a part of the Library’s core mission?
- Do we want to continue to be opportunity driven (supported by grant funds), or create a program in which we direct the selection of materials to be digitized?
- **Starting Point:** Determine the costs & benefits to create a appropriate digitization program where we control selection.

- Should *digital preservation* become part of the library's core mission?
  - **Licensed materials: E-journals:**
    - **Starting Point:** Document the advantages/disadvantages of different preservation services (e.g. CDL, LOCKSS and Portico – what are we getting for our money?).
  - **Licensed materials: E-Books, databases & other licensed content:**
    - **Starting Points:**
      - Find a national forum to discuss the preservation of these materials.
      - For materials that we license, declare our responsibilities (if any) for long term preservation & access. For example, should we license materials for which we do not have “perpetual access” rights? What is our strategy to ensure these rights (e.g., we get copies of the digital materials; they are deposited in an escrow account, (i.e., a preservation service) for us, etc.)?
  - **Web Archiving:**
    - **Starting Point:** Create selection criteria principals and priorities (coordinated with other UCs?) that factor in costs. Assume that we will use the CDL service, but investigate other services to better understand what we should expect from CDL.
  - **Other materials: Library & campus generated digital content:**
    - **Starting Points:**
      - For materials that we digitize ourselves, declare our responsibilities (if any) for long term preservation & access.
      - Declare our role (if any) to help create and preserve digital materials owned by other campus departments.
      - If we have responsibilities for either of the above, create selection criteria principals and priorities that factor in costs.

- Transition our project driven digital activities into a more formal digital library program based in the “life-cycle” of digital materials (selection, creation/acquisition, management, access & preservation).
  - The Library already supports many activities found in a life-cycle based digital library program. We have a capacity to select, create/acquire, manage, provide access to, and preserve digital assets. These capacities are activated on a project basis. It may be better if they were integrated into a *program* that looked at needs across projects, including support of licensed materials (e.g. preservation of e-books).
  - **Starting Point:** Create a Digital Library Program that addresses Library roles and responsibilities for the full lifecycle requirements of digital assets managed or licensed by the Library. That is, address issues listed in this paper, as well as others that will certainly arise in such a discussion.
  - **Smaller Starting Point:** Inventory the roles, responsibilities and capacities that we have and compare these to what would be needed in a Digital Library Program. I.E., not yet commit to implement the program.