

## **Access Tools: Bridging Individuals to Information**

Linda Frueh, Internet Archive

We are here to discuss the enabling factors for individuals to participate in data-driven research. The goal from the Internet Archive's perspective is to provide researchers and the general public with powerful and flexible access to stored information. As a community, we have brought several links between information and individual access into production; with the successes we have come new opportunities to complete the data-to-researcher bridge and new directions in which to fund research.

### **1. Information Collection: In Production**

The workshop participants represent significant success stories in gathering information into large collections. In order to create meaningful archives, libraries and repositories we have dealt with rights issues, the cost of collecting and organizing data and digitization for easy access. We also are managing the challenges of working across multiple institutions to create collections.

Tools have been created to enable large-scale information gathering, including on-site digitization, digital database platforms and web harvesting. The Internet Archive has used these tools to create collections numbering in the billions of web pages, in hundreds of thousands of digitized books, movies and audio recordings. Others have assembled quantitative data sets in the millions of items. These information-gathering tools are most powerful when built on open-source principles; they result in some standards across collections, which will later facilitate information sharing. Many such tools are being adopted, shared and further developed by the research community – but new opportunities exist to facilitate digital information collection.

### **2. Petabyte Storage: In Production**

Storage of vast quantities of data is the second challenge taken on by this group. Petabyte scale storage is now readily available, with stable, replicable architecture. Storage facilities are becoming networked to enable information sharing and exchange. As a group, we're taking on the management of offsite storage and adoption of standards to enable internetworking of multiple locations. Further challenges include technology migration and development of much larger scale storage capability – perhaps to exabyte scale.

But researchers should not be limited to any single storage facility or organization in plumbing data for new discoveries. Hence, interoperability of storage structures is an opportunity to strengthen the information-to-individual bridge.

### **3. Preservation: Research Still Needed**

Peter Murray-Rust has cited very high figures – 80% and up - for the loss of primary research data post-publication. Preservation of information is an essential task in the information bridge - enabling scientists and scholars to find unexpected and novel associations without having to generate new primary data.

The Internet Archive and others are implementing policies to ensure sustainable, archival preservation by keeping duplicate copies on separate devices and storing multiple copies on different continents, with different administrators. LOCKSS systems and the Internet Archive's mirror facilities in Amsterdam, Alexandria Egypt and San Francisco are examples of these policies in action. The goal is to protect information from the failure or policy changes of any individual system, organization or administrating body. More sophisticated preservation policies, addressing file format issues and technology standards, are sure to be a new frontier for leadership by JISC and NSF.

#### **4. Access Tools: Research Needed**

Ultimately, researchers must be able to get at stored information. Access tools are the last, critical step in supporting individual participation in data-driven research. These tools can be for finding, associating, tagging and downloading information – one node at a time or in bulk. Cornell University is pushing forward this frontier with its link structure analysis of the web, and its access tools overlaid on Internet Archive collections.

Opportunities abound to encourage and enable better and more powerful access to archived collections. For example, there are approximately 5,000 living, written languages today – how will we support language-independent research in this new digital world? The Internet Archive has 25,000 digitized texts in Arabic; how will a Portuguese-speaking scholar gain useful access to them? Our Television Archive has one million hours of television – none of them indexed. How will scholars and historians be able to review and analyze them?

The Internet Archive has hundreds of thousands of users each day, but we believe that with more sophisticated access tools it can be millions. Some observations from the Archive's experience with access tools so far:

- Offering programmer access to our storage machines has not attracted much activity, particularly not from researchers and scholars themselves.
- We have been successful at boosting access with APIs and user-friendly interfaces: e.g. the Wayback Machine, fully searchable text and web collections, and flipbook readers.
- The Internet field is bursting with community-oriented sites and organizing tools - Wikipedia, YouTube, GoogleMaps, and LibraryThing are some of the successes at attracting tremendous traffic because of their novel interfaces. Their success offers insights into features of interest to individual users – including tagging, annotation, user-generated content and community discussion.
- To design great access tools we need to understand and incorporate end-user workflow issues; this suggests applications tailored to specific research communities

#### **What can NSF/JISC do to support Individual participation in data-driven research?**

Digital collections are coming together in enough quantity to support data-driven research. The limiting factor now is tools that provide great, interactive access to the materials. Opportunities for development in this area include:

- > **Sets of open source software that can index collections and make them easily and flexibly findable**
- > **Library-scale machine translation**
  - Many languages to many languages
  - Hundreds of thousands to millions of books
- > **Library-scale universal OCR**
  - Language-independent OCR
  - Multiple languages in many typefaces
- > **Massive/bulk digital indexing of video content**
- > **An open source web search system**
- > **Tools to enable time based studies for trend analysis and sociological change**

This group can help us to, in the words of David MacArthur (NSDL) “move from the era of developers to the era of end-users.”