Defining "Born Digital"

An Essay by Ricky Erway, OCLC Research

The purpose of this document is to define "born digital" and the various types of born-digital materials. It is intended to improve community discourse by encouraging caretakers of born-digital resources to specify what they mean when they use the term.

Definition: Born-digital resources are items created and managed in digital form.

Types of Born-Digital Materials

Digital photographs

The prevalence of digital cameras is making digital photos one of the fastest growing forms of born-digital content. Custodial emphasis should be on ensuring they are in current, mainstream formats and are copied onto contemporary, durable media. Care needs to be taken concerning color space and compression, which may affect the integrity of the photographs. Much is known about digital photographs and there have been years of experience in curating collections of this nature.

Digital documents

Nearly all documents are currently created in digital form. Whether to maintain them on paper or in digital form is a basic, but important decision. For those maintained in digital form, standard formats such as the Portable Document Format (PDF) should be used to retain formatting, while separating the documents from the software that created them. The many efforts to capture and preserve the intellectual output of a university in an institutional repository are developing expertise in this area.

Harvested Web content

While the Internet Archive captures snapshots of the Web, institutions may take it upon themselves to do more focused archiving in a more thorough manner. A national library may archive its nation's Web sites. A university may archive its own domain. Archives might harvest from Web sites related to a particular subject or event. Open-source tools developed by the Internet Archive can be used to crawl and provide access to the content. The data can be kept in the ISO standard WARC (WebARChive) file format. The approach to Web harvesting is fairly stable.

Digital manuscripts

Personal “papers” can arrive as born-digital manuscripts, which may accumulate while archivists plan what to do with them. A first step is to work with donors so that the problem doesn’t keep growing. Advise them in advance of approaches for weeding, organizing, and naming files. Recommend formats and media. If needed, seek to acquire their equipment as well as their media. Consider periodic accessions from living donors. Update donor forms to reflect policies and practices for born-digital materials.

A very few manuscript collections may merit emulation in order to recreate the workspace of the author. Some high-profile projects are beginning to emerge in this area; they tend to focus on exceptional collections for which grant funding is available to provide exceptional curation. Most collections warrant minimal treatment focused solely on preserving the content.

Electronic records

This category includes government documents and corporate, institutional, and organizational archives. This type of collection might consist mostly of documents in word processing formats or may include an array of e-mail, databases, spreadsheets, presentations, and other types of files, some of which can only be read using proprietary software. In most cases it’s best to get the content out of proprietary formats.

Archivists should be involved in setting policy for their institutions and not just doing clean-up. Fortunately, there are open source tools for ingest and management of electronic records and training is available from the Society of American Archivists and through National Historical Publications and Records Commission funded regional training programs.

Static data sets

Data sets are created in the course of research and can be the basis for future research, but they are often created without consideration for preservation or future access. Some data sets need special software and documentation to make them usable and the system may need to be retained or emulated. Context, including the nature of the sample, data collection approach, and software used, should be retained.

Management of research data is often done outside the library or archives environment. The hundred-million dollar NSF DataNet grant program will establish data curation hubs for the sciences. The humanities are not as well-funded. Not every institution will be able to support every discipline, but some institutions may take on responsibility for the support of a particular field. Multiple universities may join together to create shared facilities.

Dynamic data

This type includes data sets that are added to over time, time-based, or that include genetic sequencing or computer-aided design (CAD). It can include data that is meaningless until it is acted upon—and there may be an infinite number of actions and results. In many cases the software, if not the hardware, environment will need to be retained or emulated.

Dynamic data can also include social environments. The Library of Congress has taken on the Twitter archive. Will someone make a similar arrangement with Facebook? What about discipline-specific social networks? This area requires a lot of one-off solutions, creating challenges for longer-term stability.

Digital art

Digital art may be as simple as digital photography or it may be much more complex in that it could be mixed media, dynamic, or could require recreation of an entire installation to render it effectively. More complex forms of digital art will likely require one-off solutions.

Digital media publications

These are materials that are routinely published in digital form. Commercial publications like music CDs, movies on DVD, and video games are on fairly stable media and when those media are replaced, the content is often rereleased in new formats. Libraries tend to keep up with the formats that their users want.

There is little immediate concern here and licensing and copyright make it difficult for libraries or archives to take action. But, as with early motion pictures, at some point the content will lose its commercial value and, unless someone takes custodial responsibility, it will be lost through obsolescence or decay.

Key Issues

Taking responsibility for born-digital materials has challenges beyond those usually associated with caring for traditional forms of content.

Examples of inherent risks include:

- Bit rot—the files have deteriorated over time
- Obsolete media—the content is on media no longer in use
- Obsolete hardware—due to technological advancements, very few, if any, of the computers or peripherals needed to access the files still exist, and they are difficult to repair or replace

- Obsolete software—software or operating systems needed to make sense of the files is no longer available
- Authenticity—the data have lost their integrity

Familiar considerations take on added dimensions:

- Versions—should unintentionally retained drafts be kept?
- Privacy—should deleted files be recovered?
- Rights—when is something in digital form considered published?
- Licensing—does the original license transfer when software is "inherited"?
- Access—should digital access be subject to the same constraints as analog access?
- Responsibility—is curation of born-digital content part of the core mission of the institution? Is there funding and authority to ensure its on-going programmatic care?
- Users—who will want to access digital content? Are they one of the institution's primary audiences?

The Way Forward

The first step is to establish basic policies and approaches for each type of born-digital material in your care. Then take inventory and assess format and media stability. Find others who are working on similar challenges. There may be already existing standards, tools, or procedures used by another community, such as law enforcement or gamers. Turnkey systems are unlikely, but there are many micro-services to handle various tasks.

While there is much to be done, there are some encouraging signs of progress:

- Increased awareness of many of the issues
- Several instances where born-digital materials are being collected
- Some instances where born-digital materials are being preserved

But there are very few instances where users can access born-digital collections. Through effective communication and collaboration and by taking basic first steps, progress will be made toward that goal.