Digital preservation: 7 steps to get started

Jody L. DeRidder | Feb. 8, 2017 | Reynolds Journalism Institute

There are a number of practical steps publishers at news agencies can do to lay the groundwork for preserving our historical record. While this is not a complete set of instructions for digital preservation, it will prepare the way for more advanced efforts to follow.

1. Reach out

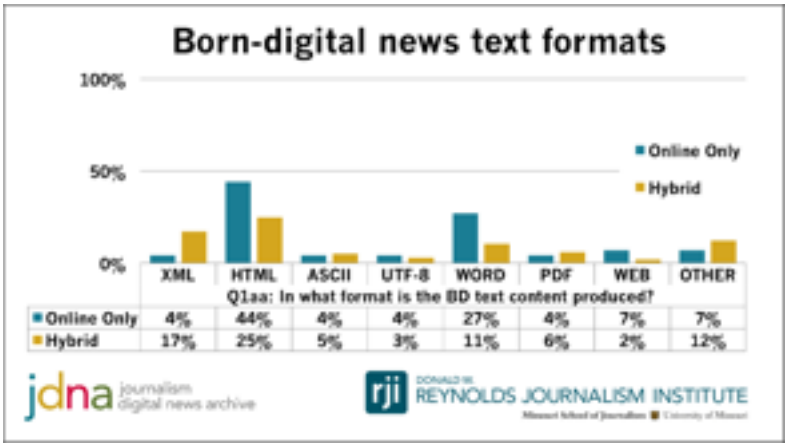
One of the best first steps to digital preservation is to reach out to others who are concerned about retaining access to news content. A great place to start is at the Donald W. Reynolds Journalism Institute conference November 10-11, 2014, entitled "Dodging the Memory Hole: Saving Born-digital News Content". Here, concerned stakeholders from a variety of institutions and agencies are gathering to address the issues of digital preservation specific to the news industry, make connections and begin to establish viable plans of action.

2. Learn more

A growing number of freely available online resources can provide a wealth of information and instruction. Check out the Digital Preservation Tutorial hosted by Cornell and the Inter-university Consortium for Political and Social Research. Specifically for those preserving newspapers, Educopia published "Guidelines for Digital Newspaper Preservation Readiness", which distills steps into incremental processes that institutions of almost any size or type can begin to address. It references the National Digital Stewardship Alliance Levels of Preservation, which begins with protecting your data and provides useful resources such as an inventory readiness checklist.

3. Consider how new content is being managed

What formats of materials are being created, where are they stored, and how are they organized? The fewer the types of formats, the better the quality and the more widely supported they are in the archival community, which ultimately means a better the outlook for digital preservation. Proprietary content such as Microsoft Word, can be difficult to support long-term because the software continually changes and is not backwards-compatible. Yet in a 2014 phone survey by the Reynolds Journalism Institute, 27 percent of born-digital news text created by online news agencies was in Microsoft Word. Knowing the amount of content generated in problematic formats is key to planning ahead.



Large amounts of content are being produced in HTML and Microsoft Word formats, according to a 2014 RJI survey.

The industry must ensure that processes are in place to avoid short-term loss until long-term procedures have been established. If all content is delivered online, how are older issues stored once taken offline? If all new content is already submitted somewhere regularly, such as for continued access to public and legal notices, that's a prime time to also share that content with an institution interested in long-term archiving and access.

4. Examine current storage methods

Organization is a must for storing media so it's usable, and most media require cool, dry storage conditions, protected from light and handling. Writing or labeling directly on media such as CDs and DVDs can cause irreparable damage, so media should be cased with useful descriptions added to the outside casing. If content is tossed into closets, boxes, drawers and shelves without organization, access to the materials is already compromised by a lack of organization. If the materials are stored in an attic or cellar or any other area where they can be exposed to heat, light, humidity or sudden changes in temperature, the rate of content loss will be accelerated.

Other questions to ask include: How often are backups made of content, and how often are they checked for viability? One of the serious problems with backups is that bit-level corruption can happen anywhere, so verification of content viability before backups occur becomes key. Are backup copies stored offsite in safe locations? Are any copies stored in geographically distant safe rooms? If not, make note of these as issues to be addressed. If possible, at least relocate stored content to a place where temperature, humidity, light and access are controlled. What security procedures to you have in place to avoid tampering or loss of content?

5. Organize and categorize content

Over the past few years the types of media and formats of files used for news capture and delivery have changed radically. It's extremely likely that old hard drives, zip disks, floppy disks, old server racks, and other varieties of media have stacked up in the back room untended. Sorting those and documenting how much of what kind of media is stored will help establish priorities and determine costs for digital preservation. Having difficulty identifying some of it? The Digital Preservation Management Workshops developed a "Chamber of Horrors" describing various obsolete and endangered media that might help.

Image via http://www.dpworkshop.org/dpm-eng/oldmedia/disks.html.

Examples of obsolete media from which content needs to be extracted and updated.

Once this level of organization has been completed, document what issues, titles, and years are where as well as any other helpful information that can be gleaned before opening the media itself. This is the beginning of an inventory, which will be developed further in later steps, so be sure to use a spreadsheet that can be expanded and later exported to a form useable by computers.

6. Determine whether to train staff or outsource the work

In the past few years staffing levels have fallen tremendously in news agencies across the country. Only the largest and most successful news publishers still have archivists or news librarians on staff. If such staff is available, how current is their training, and have they the time and resources to take on the next steps of digital preservation? If it is in the agency's best interests to manage the preparation of content for preservation, consider the costs of continual training, specialized hardware and software for accessing older media. Other elements to consider include the resources necessary to create metadata, normalize file formats and extract checksums for existing content. One of the best training programs in the country right now is "DigCCurr Professional Institute" at the University of North Carolina, Chapel Hill.

7. Identify next steps

Reviewing the results of the initial inventory will point out some of the next steps that need to be taken. For currently accessible content, the next steps would be to improve storage, backups, organization and metada-

ta. This might involve negotiating a collaboration with a memory institution that already has the infrastructure in place to manage news content long term.

For legacy content, the next steps would be to determine what will be required to make it accessible and usable by current hardware and software systems. This is a serious research project, as the variety of media and formats can vary immensely. After locating and obtaining the hardware and software necessary to access the media, extracting it without modification and collecting initial metadata are critical. Write blockers should be used when opening media, and checksums should be captured along with creation date, size, file structure and other information before extraction has even begun.

Some research institutions have established digital forensics labs, such as this one at Stanford, in which a variety of media can be safely opened, content extracted, analyzed, catalogued and reformatted for use in current computer systems. The BitCurator Project provides training and software tools for extraction, analysis and collection of metadata. This open-source software is designed to run on a Linux system, which can be mounted virtually onto a Windows station.

Next, identifying the files will be necessary in order to access them. Open source software such as FITS, JHOVE and DROID can be used to identify files. Anyone can search on various file extensions in the UK National Archives, which provides an online technical registry, PRONOM, to help identify file types and the software necessary to access them. After identification, legacy files need to be migrated to current archivally viable file formats for storage and ongoing management.

It is possible to outsource all this work to vendors but first ensure that the vendor involved is fully competent in digital preservation measures and protocols.

Preparing content for preservation clearly involves technical expertise as well as library and archival knowledge, so the next steps for publishers has come full circle. Now is the time to reach out to other concerned stakeholders and establish protocols and agreements for how news content can be managed long-term for continued access and use. Several institutions have already developed models for transfer and management of born-digital news content, such as:

- Texas Digital Newspaper Program
- Kentucky PaperVault Project
- Digital Library of the Caribbean
- Florida Digital Newspaper Program
- Minnesota Digital Newspaper Project
- Born Digital Project of the California Digital Newspaper Collection.

Contact state archives and local memory institutions to locate others who can assist in preserving born-digital content. Connect with concerned leaders such as Edward McCain at the Reynolds Journalism Institute. Join in the discussion on how best to manage and preserve the history of America.