**George Washington University**

**Gelman Library**

**Special Collections Research Center**

**Jump In, Too/Two: Survey Findings**

by Nick Connizzo

Electronic Media Project Archivist

05.01.2014

**Introduction**

The archivists and librarians at Special Collections Research Center (SCRC) at George Washington University (GWU) have long known that within their widely-varied collections there exists, potentially, a substantial amount of born-digital content. These materials range from decades-old floppy disks and drives to present accessions of DVDs, flash drives, and other media from individuals, university offices, and other organizations. They have also long known, regretfully, that these materials were not always afforded the care and handling that they deserved, often being slotted into series full of miscellanea, remaining unprocessed, or more likely, discarded (with the fate of their contents unknown). Of course, much of these practices were the result of prevailing archival practice, MPLP, or simply a technological, resource-driven, or ignorance-born inability to take all the necessary steps to ensure the proper preservation and continued access to these materials.

We ask now: how can we avoid the mistakes of our forebears and change our practices for the better? What can we do to ensure that born-digital materials and other media are afforded the care and handling to ensure their preservation and the integrity of their contents indefinitely? The Jump In forum’s call for participants aligned perfectly with our needs and we recognized the opportunity share our experiences with others who might be in the same situation.

**Jumping In**

At the beginning of 2014, I was charged with surveying the entirety of SCRC’s holdings for born-digital (and other electronic) media. Within the scope of my survey are all physical and digital formats - any drives, disks, or other storage media plus items that have already been moved off (or never lived on) physical items in the collections. In order to effectively tackle such a broad assignment, I broke down the survey and analysis into four steps:

1. Survey all the collections in the SCRC for born-digital and other electronic media to identify what there is (quantities, formats) and where it is (what collections, what storage sites).

2. Inspect and perform conservation surveys on this material to determine which items are at greatest risk for deterioration or outright loss, inaccessibility (due to fragility, lack of playback devices), and other issues (poor description, misidentification).

3. Craft digitization/reformatting project plans for items deemed most at risk and highest in value, according to available resources. Arrange better care for physical media and digital media on them; ingest of digital media into a secure, long-term storage solution.

4. Rework existing accession, processing, storage, and reference/service workflows to fully incorporate born-digital and other digitized media as a major part of the department’s mission.

Part of the survey also included analyzing accessioning and description practices. While we know that our procedures need to be reworked, or at least refined, it was useful to my survey to understand how materials have been historically accessioned and described, in order to better locate items (or lack thereof).

Obviously, such an effort is expected to take more than a few months, so much of it will fall outside the scope of the Jump In Two survey. However, the first stage (the survey of materials) has already shed a great deal of light on the issues that we (and other institutions, presumably) face immediately, and has proven to reveal a substantial amount of data, shared below.

**Some preliminary results gathered from a survey of a majority of SCRC’s holdings:**

**Collections:** 603 (in four major subsets: Manuscripts, University Archives, the International Brotherhood of Teamsters Records and Labor Archives, and the National Education Association Records)

**Total size of all archival collections:** more than 20,000 linear feet

**Percentage of collections with born-digital material:** more than 10% (mostly University Archives)

**Percentage of collections with other electronic material:** more than 60% (audio, video, digitized surrogates)

**Percentage of born-digital material as a portion of all collections:** less than 1% (see note)[[1]](#footnote-1)

**Formats:** CD, DVD, Hard Drives, Flash Drives, Floppy Disks (3.5” and 5”); predominantly Windows. Upwards of 3 TB of content (the majority being recently accessioned 1 TB HDs). Other substantial sources of born-digital material include CDs with photographs, floppy disks with documents, and audio files on servers.

**Oldest born-digital materials:** hard drive from 1970 and disks from early 1970’s

**Newest born-digital materials:** hard drives from 2013 and born-digital files accessioned in 2013 (and stored on GWU servers)

**Results in finer detail:**

While a strong percentage of collections contained some kind of born-digital material, the bulk of the material is very recent (within the last five years). Additionally, while a somewhat surprising number of collections contained media likely to have born-digital material (disks, CDs with data, hard drives), a majority of these collections contained only a handful of items. Instead, the bulk of the born-digital material is held in a significant minority of collections. This is expected, as many of the SCRC’s manuscript collections date well before the digital age, and those that have received an accession within the last five years are much more likely to have born-digital material than those collections which are no longer growing.

The majority of collections which have born-digital content fall within the purview of the University Archives. While GWU has no defined electronic records policy (though such a policy is currently under review by the University), it stands to reason that organizations that create records electronically as a regular course of business are likely more aware of the need to preserve those records, and we at the Archives know that the University has been creating electronic documents for some time. We have found that even today, some individual donors do not realize and must be convinced that their born-digital creations such as emails, documents, and even tweets and other social media posts are records worth saving. Furthermore, an organization with many members is more likely to have someone among its ranks with the knowledge or forethought to consider saving materials of all kinds. Two of SCRC’s largest collections, the International Brotherhood of Teamsters Records and the National Education Association Records, come from large organizations that are still in operation and are thus actively producing electronic records. Although these collections have proportionally little born-digital material at present, that proportion is expect to skew greatly and irreversibly in the opposite direction within the next few years.

It was often surprisingly challenging, sifting through accession logs and finding aids, to even identify born-digital content. Most of the archival description done by SCRC archivists is not done down to the item level, making it challenging to discover items which have not been segregated by media type. In cases where a hard drive or flash drive was part of a collection, these were typically labeled appropriately or otherwise noted; but in the case of CDs or other disks, whatever label was on that media was added as a line-item to the inventory. Fortunately, conscientious description practice did provide rather thorough scope notes and inventory lists, though I suspect in many cases the final tabulation of how many items with born-digital content will differ from these preliminary results.

In rare cases, materials accessioned by SCRC were immediately transferred or copied from their physical medium into another system. Recent practice has incorporated digital forensics techniques, making disk images, and ingesting the product into a repository. Some materials donated to the SCRC have almost certainly come in on flash drive, on CD’s, or through email or Dropbox; in some cases these were directly added to a repository (or simply put on a server) without any forensics performed and their provenance noted in a log. Since there are relatively few places that files of this type can be located, they were easier to find, but often they were separated from the document outlining their provenance or had no accompanying metadata, making proper identification more difficult.

Anecdotal evidence suggests that there were few differences between the way that SCRC staff treated born-digital materials and how they treated permanent analog items. Often, the container itself was processed, described, and arranged within series, but little or no work was done to determine what kinds of materials were on these containers other than to apply the data gleaned from the label to the finding aid. During the course of the survey I discovered a number of items (a few hundred) that were afforded their own folders or boxes, and in some cases arranged with series of audio-visual and other non-print media; however, trickier to find, though certainly present, were items with no mention (at the folder or box level description) that disks were present, which is almost certainly due to MPLP-based processing procedures.

In summation, many of my expectations going into the survey were confirmed, but I was just as often surprised by my findings. I was able to discover born-digital material in collections that outwardly did not appear to hold any, but also failed to find any electronic records in collections I was sure would have some. Additionally, though many collections did not have born-digital material, some did have the fingerprints thereof - printed out emails or documents. It became clear that the average record-creator probably did not view their files as worthy of long-term retention; more likely, they selected those *official* documents and printed them to include with their papers, a kind of pre-appraisal done before the collection ever reached Special Collections.

**Next Steps and Moving Forward**

Despite some hiccups and limitations historically dealing with born-digital media, we have mandate and the wherewithal from the Library to properly handle them into the future. With that mandate comes support: in addition to training for staff, our LIT unit is standing up new front- and back-end repository systems for the long-term preservation of born-digital content, and we have access to the technology (whether it’s a FRED machine owned by the Washington Research Library Consortium or open-source tools like Archivematica or BitCurator) and the knowledge to implement modern procedures and techniques accepted by the archival community. We hope to go beyond even that, and be a leader in developing and testing new practices in order to contribute our growing expertise to the community.

With the knowledge we’ve gained from this survey, we now know what tools and techniques to which we must have access. We need to be able to handle optical media, floppy disks, hard drives in multiple file systems, and materials in file formats of all types (.doc, .pdf, .mp3, .wav, etc.). We need to be able to support file formats (or normalize items to such formats) that the archival community has not pledged to indefinitely support. The library accessions items from Apple filesystems, reads them on our Windows workstations, and stores them on a Linux server, so we must account for the frequent transfer (and protection through checksums and other integrity checks) of these materials. We are fortunate that there is not a great preponderance of obscure or particularly difficult items such as Zip or MiniDisc, and with that knowledge we can afford to focus on the material types that we do have. Moreover, we can more tightly focus our collecting policies to exclude materials of unsupported formats (or at least to attempt to normalize those media before accession).

More specifically and immediately, the survey turned up a number of collections with inadequately labeled or inventoried materials. These must be inspected to ensure that there is no media with preservation concerns hiding within. Collections with born-digital content have been flagged, and a survey and subsequent reformatting/preservation priority list is already being developed.

**Conclusion**

The road to having a perfect program for the preservation and access of born-digital and other electronic media is long, perhaps infinitely so. Among the first step down that road is surveying current collections and acquiring both broad and narrow views of what we have. The knowledge produced provides the foundation for future policymaking: what technology do we need to support now and into the future? What types of media will we collect in the future? What special care and handling do our current collections need, now and in the future?

Certainly, we feel more confident and knowledgeable now that we have the data at our fingertips. We can make informed decisions about the future - the future of the department, of the community, and, hopefully, the perpetual future of the material that we’re all here to provide and preserve.

1. There is an ironic tension between the typical metrics used to measure the size of archival collections (cubic or linear feet) and the amount of data stored on a CD or a hard-drive. A single linear foot may produce (depending on the file type, encoding scheme, number items, and other factors) anywhere from a few hundred megabytes (low-quality PDF/A scans of documents) to multiple terabytes (high-quality uncompressed video) of digital content. So any comparison of the “amount” of digital material to the physical amount of paper material must be taken with about a half-shaker of salt. [↑](#footnote-ref-1)