## Background

The Manuscripts Division, within Princeton University’s Department of Rare Books and Special Collections, holds over 8,500 linear feet of materials covering five thousand years of recorded history and all parts of the world, with special strengths in Western Europe, the Near East, the United States, and Latin America. Collecting strengths include: Ancient History, Art and Photography, Ethiopic Manuscript Collections, Islamic Manuscripts, Literature, Medieval and Renaissance Manuscripts, and Theater and Film.

Although most collections in the Manuscripts Division pre-date the 20th century, the number of more contemporary collections is rapidly increasing, particularly Latin American literary collections that include the papers of still-active authors and cultural figures. Coming to terms with this reality has led our division to begin taking the necessary steps to properly manage the born-digital materials that are increasingly part of these collections.

Participating in this survey project has provided us the opportunity to tackle the first step of managing our born-digital materials, which involves establishing intellectual control over what we already have in our collections. Those involved in the project are Faith Charlton, Processing Archivist of Americana Collections, Kelly Bolding, Manuscripts Processor, and Chloe Pfendler, Curatorial Assistant for the Manuscripts Division.

## Survey Process

With the assumption that we currently have a minimal amount of digital media, we decided to survey all of our manuscript collections, which number over 1500, by first surveying our EAD finding aids for description indicative of possible born-digital materials.

In the past, descriptive practices regarding digital media were not done in a consistent manner if at all, largely because the division received so few collections with born-digital material; as a result, we anticipated that existing descriptions of digital media would vary from finding aid to finding aid. This meant that our survey tool would need to capture descriptions of digital materials located in various EAD elements. With the help of Regine Heberlein, Principal Cataloger and Metadata Analyst, we wrote a simple XQuery (see Appendix A) that looks across all descriptive components of our EAD finding aids to locate text strings that match a regular expression based on list of words, including variant spellings, that we determined would indicate the likely presence of born-digital materials, such as disk, floppy, CD, DVD, drive, digital, electronic, etc. The XQuery generated a list of matching components in XML, which we then imported into Microsoft Excel, creating a table design in which each row represents a component that the XQuery located, and each column, an EAD element within that component.

Along with assistance from student workers, we then meticulously examined and revised the spreadsheet, removing any irrelevant, extraneous, or redundant information (for example, false positives that resulted from other text strings matching our regular expression, or duplicate records for the same item resulting from multi-level description). We then extracted and transferred the most relevant data, including media types, file formats, and quantities, into additional columns, in order to provide more structure for this technical data, as well as to compute estimated totals. We also created columns to track our attempts to determine from the EAD data whether or not an item was likely born-digital or contained files of materials that had been digitized, keeping in mind that these columns will be updated when we physically examine the materials at a later date.

Findings

With the understanding that we are relying on imperfect metadata from older finding aids, our survey has determined that the Manuscripts Division holds approximately 232 (born) digital media totaling 394 GB. Formats discovered and estimated quantities of each format are listed in the chart below.

|  |  |
| --- | --- |
| **Media Type** | **Quantity** |
| 3.5" disk | 5 |
| 5.25" disk | 6 |
| Floppy disk | 17 |
| CD | 103 |
| DVD | 32 |
| CD or DVD | 60 |
| Zip file | 4 |
| Flash drive | 1 |
| Unidentified | 4 |
| **Total** | **232** |

In order to store two preservation copies and one access copy of all of the files on the digital media we found, we would need a total of 1.2 TB of storage space. While this number is likely somewhat inflated due to the fact that all media are probably not filled to full capacity, we prefer to err on the side of overestimating our storage needs, especially due to the anticipated presence of additional digital materials in our holdings for which we have no description.

Furthermore, because many of the items we discovered have dates that reflect the *content* on the media and not the date of the media itself, it is difficult to note with certainty the dates of the material; though with that in mind, we can say that the date range is approximately 1985-2014. We also observed that our EAD data only includes information about the file formats and operating systems on which files were created for a handful of items.

Next Steps

The immediate next step in our survey process is to enhance the metadata that we gathered from querying our EAD files, as many of the media we identified are not sufficiently described in our finding aids for us to complete a thorough survey and prioritize materials for further processing. This step will entail physically looking at the items, in particular those items whose media type is unknown or whose descriptions are lacking specificity, such as those broadly described as “floppy disks” or “diskettes,” and adding to our data any annotations on the items that may exist. We’d also like to continue to determine, if possible, which items are actually born-digital as opposed to those that contain files of materials that were *digitized*,and to note whether or not the latter are duplicates of original paper copies. Until we revise our accessioning workflow, our temporary plan is to keep our Excel spreadsheet up-to-date in order to keep track of digital media that we receive as part of future accessions. Once we gather additional information to fill out our descriptions of digital media, we intend to add that information to our EAD finding aids in a structured way, and to document descriptive workflows in order to ensure consistent description of born-digital materials moving forward.

Once physically surveying our digital media is complete and we have a more solid understanding of what we have in our collections, we intend to use this information to help us determine how to prioritize the treatment of these materials. Comparing our list of collections that contain digital media to frequency of use statistics from Aeon, our Public Services request and workflow management system, we plan to strategically prioritize born-digital materials from highly used collections for additional processing, which will include the analysis, possible disk-imaging, and file capture of digital media using BitCurator and the Forensic Recovery of Evidence Device (FRED) at Mudd Manuscript Library. Simultaneously, we will be developing policies and procedures for born-digital materials that will address new acquisitions, preservation, and access. We are extremely fortunate in that a solid foundation of policy and infrastructure for managing born-digital materials has already been laid by our colleagues who oversee the University Archives and Public Policy Papers at Mudd Manuscript Library. We hope to begin to apply what they have developed to collections within the Manuscripts Division in the very near future.

## Appendix A: XQuery Survey

**xquery version** "1.0";
**declare namespace ead** = "urn:isbn:1-931666-22-9";
**declare default element namespace** "urn:isbn:1-931666-22-9";
**declare copy-namespaces no-preserve**, **inherit**;

**import module namespace functx**="http://www.functx.com"
 **at** "http://www.xqueryfunctions.com/xq/functx-1.0-doc-2007-01.xq";

**declare variable $COLL** **as** *document-node*()+ := *collection*("path/to/EAD/directory");

**let $contains\_media** := **$COLL**//**ead:c**/**ead:\***[*not*(self::**ead:c**) and *matches*(*string*(.), '(\s|^)flopp(y|ies)(\s|$)|(\s|^)dis(k|c|kette)s?(\s|$)|(\s|^)cd(-rom)?s?(\s|$)|(\s|^)dvds?(\s|$)|(\s|^)digital(\s|$)|(\s|^)(usb|hard|flash)\sdrives?(\s|$)', 'i')]

**return**
<results **xmlns**=**"urn:isbn:1-931666-22-9"**>
{
**for $media** **in $contains\_media**/parent::**ead:c**
 **return**
 <c **level**=**"**{**$media**/***@level***}**"** **id**=**"**{**$media**/***@id***}**"**>
 {**$media**/\*[*not*(self::**ead:c**)]}
 </c>
}
</results>

## Appendix B: Photographs of Digital Media



From the Mario Vargas Llosa Papers (C0641) and the Stanley Kunitz Papers (C0837).



From the Douglas Kent Hall Papers (C1384).