OSSArcFlow

Digital Dossier

By Paul Kelly
District of Columbia Public Library

OVERVIEW

DC Public Library employs two Digital Curation Librarian FTEs, and one Digital Projects Intern, whose hours vary but average at around 15 per week. 100% of the DPI's time is spent on digital curation activities. Of their 40 hours per week, DCLs spend roughly 80% of their time on digital curation, although, broadly speaking, the remaining 20% might be considered digital curation- related. Although other employees are somewhat in the orbit of these activities, they are considerably more ancillary to their core duties.

DC Public Library's Information Technology division has recently been phasing out local application support. Special Collections, therefore, relies mostly on a patchwork of hosted services to meet its digital curation goals. That said, IT still plays an important role in that they ensure that both DCLs possess admin-level user accounts on their desktop and laptop computers to allow testing of open source software that does not require a server component. IT of course also provides basic hardware support.

DIGITAL CURATION ACTIVITIES

DC Public Library utilizes three public discovery systems: Sirsi, ArchivesSpace, and CONTENTdm. Although we are currently working to ensure that metadata in one system references corresponding metadata in another (for example, a catalog record would point to a finding aid, and the finding aid would point to digital objects), that project is not yet complete.

EDUCOPIA INSTITUTE

Internal documentation about collections and activities is generally shared via Google Drive, where we can comment, collaborate, and track changes, but is also stored locally on the library intranet.

DC Public Library's digital collections are comprised primarily of digital surrogates, although we are moving toward accessioning more born-digital material. Our non-public Preservica instance (which includes both Glacier and S3), contains 59 digital collections comprised of 357,810 objects, and takes up around 8.2 terabytes of storage. Formats included in preservation storage are TIFF, PCM, WAV, and MOV. These digital collections are mirrored, with some exceptions, in CONTENTdm, where access copies reside in PDF, MP3, and JPEG formats. To compare, current CONTENTdm storage usage sits as 65.05 gigabytes.

The major phases of the digital curation lifecycle at DC Public Library vary depending on whether material is born-digital, or digitized. Born-digital material undergoes pre-acquisition assessment before it is accessioned. Accessioning involves a member of staff, usually a Digital Curation Librarian or Archivist, entering donation data into a standard Google form, which then forms the basis of an ArchivesSpace accession record. For digitized surrogate materials, a new accession is generated from the record that pertains to the source collection. For digital material on physical media, that media is then imaged in the BitCurator environment, and reports are generated. Images and reports are stored in their entirety in Preservica, while selected documents are migrated to a standard format (usually PDF) and made available publicly through CONTENTdm. Digitized materials are almost always created by a vendor. When vendors return material to us, master and access copies are provided. The Digital Curation Librarians either create or approve collection metadata, and ingest masters to Preservica and access files to CONTENTdm.

One exception for born-digital material is web archive data, which is harvested via the Archive-It service, stored by Internet Archive, and described at both Accession and Resource levels in ArchivesSpace, as well as at the website level in Archive-It. Broadly speaking, almost all spreadsheet metadata is edited in both Excel and OpenRefine, and XML is manipulated with a variety of Python scripts. DC Public Library defines its different categories of digital content as audio, video, multipage document, single page document, web archive, and other (which pertains to types that have been processed but for which no standard internal workflow is established, such as disk images or email archives).

GOALS FOR DIGITAL CURATION

DC Public Library's immediate digital curation goals are, one, to increase our internal capacity to digitize or migrate multiple media types, two, to implement a new digital collection management system that handles both access and preservation, and three, to actively seek out more born- digital collections for acquisition.