Adding Metadata and Ingesting Large Born-Digital Archives with Archivematica

JANE GORJEVSKY AND DINA SOKOLOVA

Abstract: Columbia University Libraries is working on a large-scale project, funded by the Ford Foundation grant, to permanently preserve and make accessible the archives of the International Fellowships Program. Active in 2001-2013, the IFP offered fellowships for post-graduate study to social justice leaders from underserved communities in Asia, Africa, Latin America, Russia, and the Middle East. The IFP records included a substantial digital component: 3.6 TB from 22 countries, in 245 file formats, 10 languages and 7 non-Roman character sets. This presentation focuses on metadata and ingest issues we faced when processing this major born-digital acquisition, and on procedural and technological solutions we adopted.

The only descriptive metadata on the file level was contained in file names and directory paths, so these were retained as an originalName metadata element in AIP METS file. Files from each office were sorted into three groups by desired access level (online, reading room, and embargoed until 2075).

Archivematica software was used to create the Submission Information Packages (SIPs) and subsequently transform them into Archival Information Packages (AIPs). One or more SIPs were created for each access group, depending on the directory size. We developed a formula to calculate if a group of files was small enough to fit in one SIP. Audiovisual materials, databases, emails, and compressed files were addressed separately. Processing included character conversion and format normalization. Access restrictions and SIP-specific descriptive metadata for each package were entered manually. AIPs were transferred to preservation storage in BagIt format.

About the authors:

Jane Gorjevsky holds a newly created position of a Digital Assets Archivist at Columbia University Rare Books and Manuscript Library since January 2012. Her duties include working with record creators, the Columbia Libraries staff and other stakeholders to plan and manage new electronic acquisitions as well as preservation and reformatting of legacy digital media. Jane is also responsible for developing policies and workflows pertaining to the RBML digital holdings and hybrid collections. Jane joined the staff of Columbia University Libraries in 2001 as the grant-funded processing archivist, and in 2003-2011, served as the Curator of the RBML Carnegie Collections unit, dedicated to collecting, organizing, preserving and providing access and reference to the archival records of four major philanthropic institutions founded by Andrew Carnegie. Prior to her work at Columbia, Jane was an institutional archivist for the March of Dimes foundation. Jane has an M.A. in History and M.A. in Archival Management and Historical Editing, both from New York University, and B.S. in Applied Mathematics and Statistics from SUNY, Stony Brook.

Dina Sokolova has been a Digital Preservation Librarian at Columbia University Libraries Digital Program Division since March 2012. Her role incorporates acquisition, assessment, and long-term management of born-digital and digitized materials from library's special and general collections, as well as development of digital preservation workflows, policies and procedures. Dina has been employed with Columbia University Libraries since 2001 in several positions, working on preservation-related digitization projects. In the last two years, she has actively participated in WikiProject Digital Preservation, editing and contributing content to Wikipedia articles related to digital preservation. Dina holds a B.S./M.S. in Microelectronic Engineering, a Masters in Library and Information Science from Long Island University, and a certificate in Database Application Development and Design from Columbia University.