Acquisition of Digital Records:

Lessons of the Ford Foundation International Fellowships Program Project

Jane Gorjevsky and Dina Sokolova
Columbia University

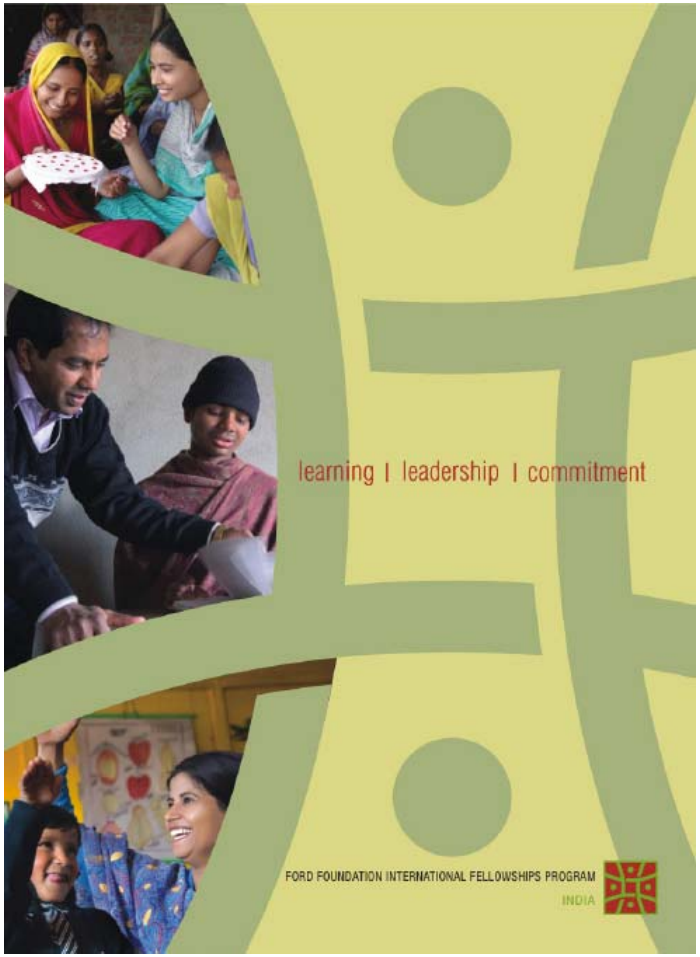
E-Archive Pilot Project

- Digital content acquisition procedures
- Hardware and software needs
- Sorting and weeding parameters and workflow
- Metadata creation or capture
- Preservation routines
- Access restrictions (tiered access)
- Finding aids and tools to view digital assets

Ford Foundation International Fellowships Program

offered fellowships for post-graduate study to more than 4,300 people via offices in 22 countries with an overall program management by Secretariat in New York in 2001 – 2013

Ford Foundation International Fellowships Program Archive



- Permanently preserve IFP paper and electronic records
- Provide access to IFP digital archives based on three types of user access:
  - publicly accessible online
  - viewable onsite only
  - embargoed until 2075
- Make IFP materials discoverable via OPAC, EAD finding aid, custom project interface.

Funded by Ford Foundation grant, October 2011

Records Scope and Content

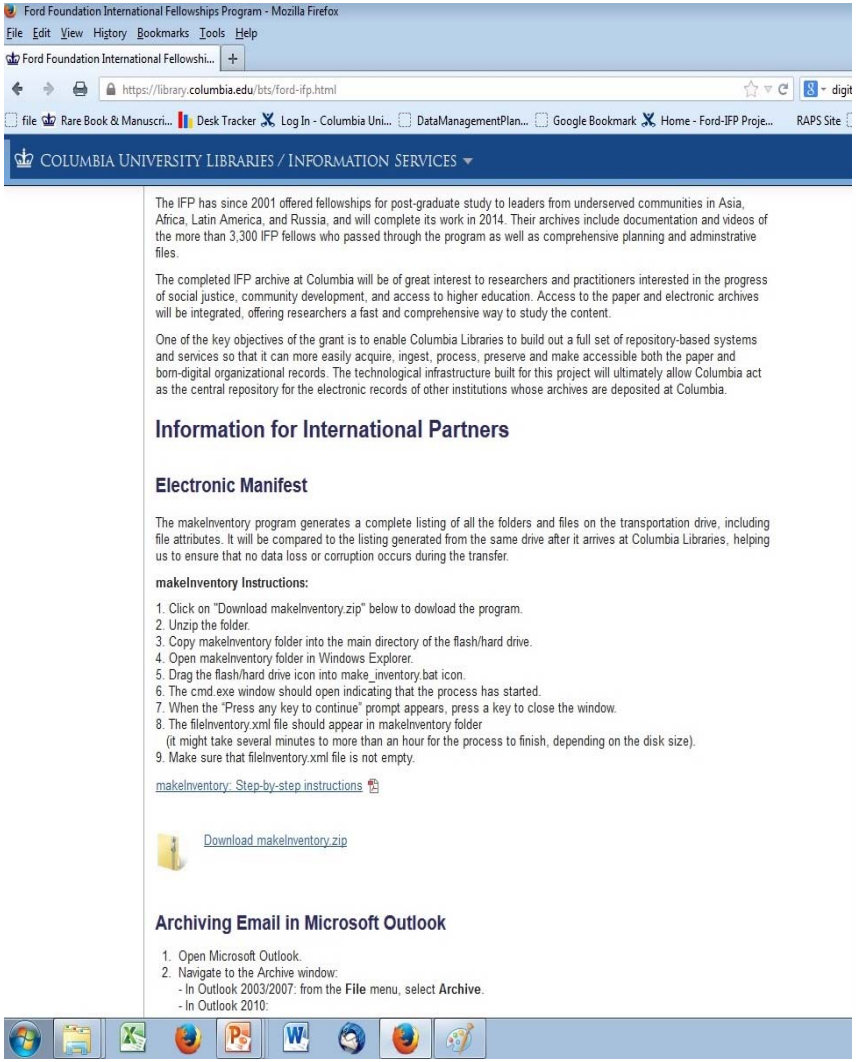
- Paper and digital records from 22 International partner organizations, New York Secretariat and CHEPS (Center for Higher Education Policy Studies)
- Materials include:
  - Office documents
  - Time-based (audio and video) materials
  - Databases
  - Email correspondence
  - Websites
  - Academic and personal records of fellows
  - Surveys, interviews and statistical reports
  - Datasets
- 3.6 TB of electronic materials in PC and Mac formats

Initial Assumptions

- Most materials in English
- Records arrive pre-selected and sorted into 3 access categories
- “Embargoed files” not accessible until 2075
- Full list of fellows and their consent status provided
- Limited number of file formats
- Sensitive information in paper format only
- No obsolete media

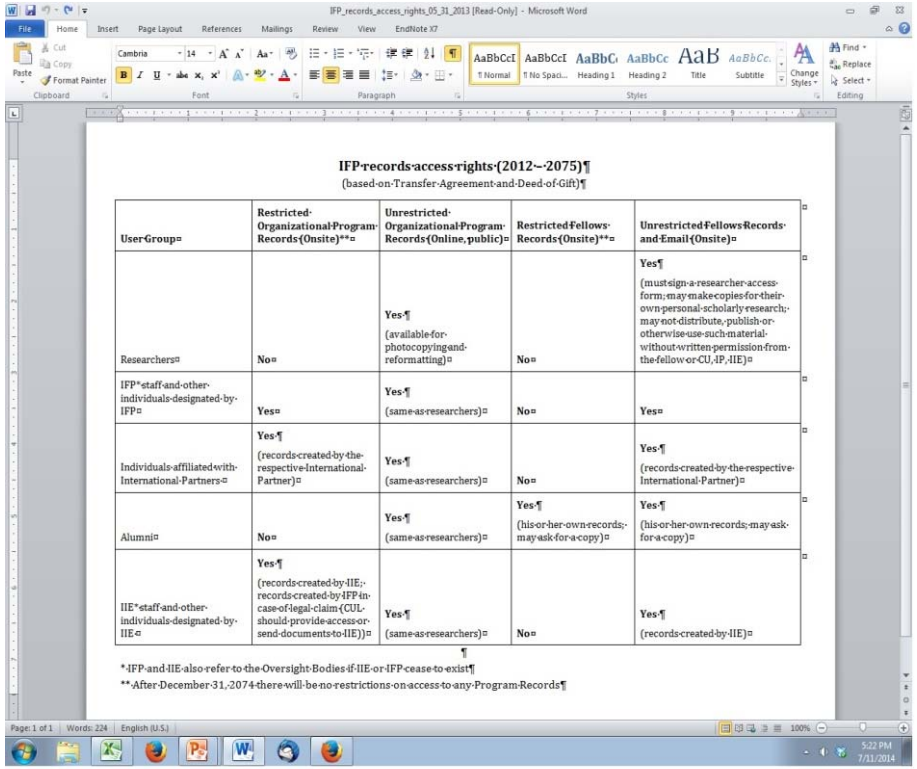
Acquiring Materials: First Steps

- Record surveys (2010, 2012) and samples
- Selection, sorting, format and file naming guidelines
- Transfer instructions and tools on Behind the Scenes section of CUL Website
- Archiving Web Resources via existing CUL program using archive.org toolset
- Internal documentation and templates on Wiki: pre-acquisition surveys, record transfer routines, inventories, accessioning, pre-processing and ingest workflows...



Content Challenges

- Selection and sorting by creators proves unreliable
- Personally Identifiable Information
- Privacy and confidentiality concerns vary by country
- Growing complexity of access needs



IFP·records·access·rights·(2012~·2075)

(based·on·Transfer·Agreement·and·Deed·of·Gift)

User·Group	Restricted·Organizational·Program·Records·(Onsite)**	Unrestricted·Organizational·Program·Records·(Online,·public)	Restricted·Fellows·Records·(Onsite)**	Unrestricted·Fellows·Records·and·Email·(Onsite)
Researchers	No	Yes (available·for·photocopying·and·reformatting)	No	Yes (must·sign·a·researcher·access·form;·may·make·copies·for·their·own·personal·scholarly·research;·may·not·distribute,·publish·or·otherwise·use·such·material·without·written·permission·from·the·fellow·or·CU,·IP,·IIE)
IFP*·staff·and·other·individuals·designated·by·IFP	Yes	Yes (same·as·researchers)	No	Yes
Individuals·affiliated·with·International·Partners	Yes (records·created·by·the·respective·International·Partner)	Yes (same·as·researchers)	No	Yes (records·created·by·the·respective·International·Partner)
Alumni	No	Yes (same·as·researchers)	Yes (his·or·her·own·records;·may·ask·for·a·copy)	Yes (his·or·her·own·records;·may·ask·for·a·copy)
IIE*·staff·and·other·individuals·designated·by·IIE	Yes (records·created·by·IIE;·records·created·by·IFP·in·case·of·legal·claim·(CUL·should·provide·access·or·send·documents·to·IIE))	Yes (same·as·researchers)	No	Yes (records·created·by·IIE)

*·IFP·and·IIE·also·refer·to·the·Oversight·Bodies·if·IIE·or·IFP·cease·to·exist

**·After·December·31,·2074·there·will·be·no·restrictions·on·access·to·any·Program·Records

Manual item-level content appraisal for unrestricted category
Initial access assumptions insufficiently restrictive

Format Challenges

- About 350,000 files in 245 formats, 10 languages, 7 non-roman character sets
- Long filenames/file paths (> 260 characters)
- Compressed and password-protected files
- Variety of transfer media (hard and flash drives, DVDs, floppy disks, ZIP disks, DV tapes) in need of conversion

Metadata Challenges

File/directory names -the only source of descriptive item-level metadata:

Non-roman character sets:

IFP\...\??? ???????\??????? ?????.jpg
IFP\...__________.doc

Long filenames/file paths:

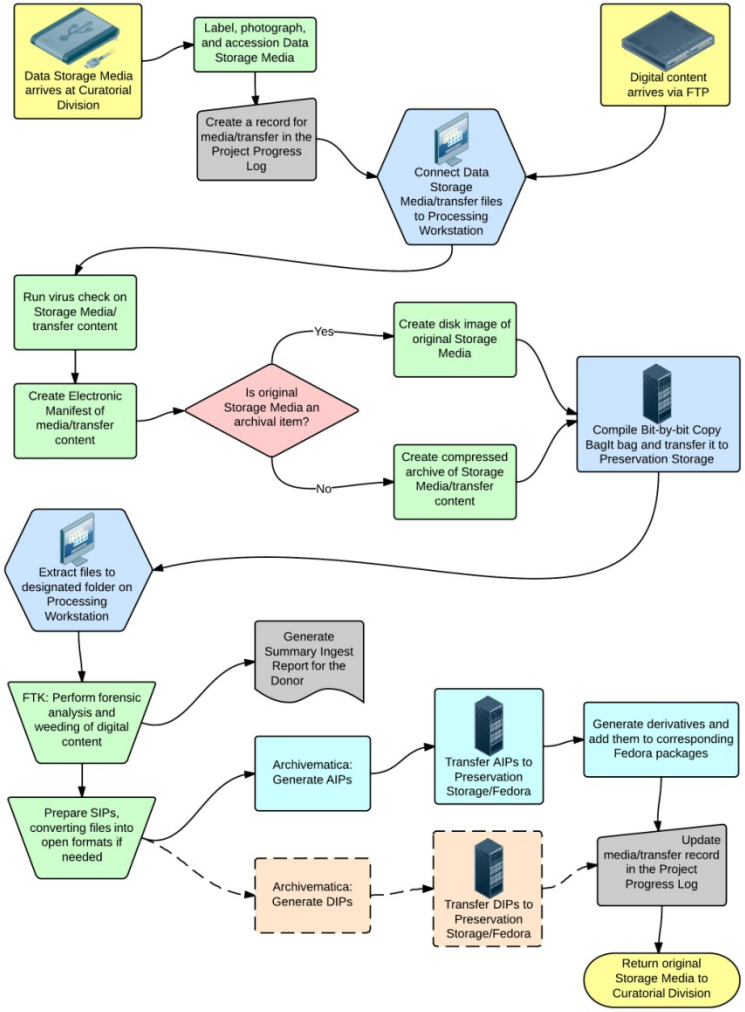
IFP\Newsletter\Alumni Meeting\... \...\...\Fifth meeting October 23-28, 2008\Agenda\IFP Assembly\Other\07.jpg

Foreign languages:

IFP\...\...\Foto bersama usai sidang kongres Perhimpunan Pelajar Indonesia Australia di Balai Kartini Gedung KBRI Canberra, 2012.jpg (A group photograph of Indonesian students taken after the congress in front of the Indonesian Embassy in Canberra, Australia, 2012)

Digital Preservation Workflow

- Preservation of bit-by-bit copy of the original transfer and related documentation (media photograph, virus check report, file inventories)
- Content appraisal, selection, and arrangement
- Processing of selected content with Digital Preservation software
- Transfer to local Preservation Storage System



Technological Tools

- Processing workstation: Forensic Recovery of Evidence Device (FRED) and Apple Mac computer

Hashdeep

- makeInventory program

- Forensic Toolkit (FTK)

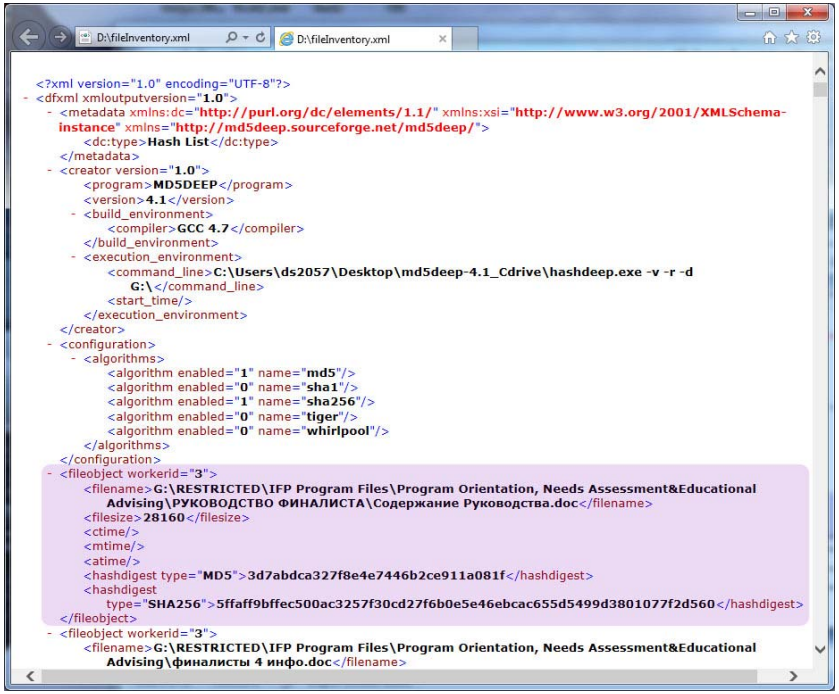
archivematica

- Archivematica

Processing Workstation

- **FRED**:
  - Create bit-by-bit copy of the original transfer and metadata using write-blocking device and external disk drives (PC-formatted storage media)
  - Perform content analysis and selection using Forensic Toolkit
- **Mac computer**:
  - Create bit-by-bit copy of the original transfer and metadata (Mac-formatted storage media)
  - Transfer bit-by-bit copies of original transfers to Preservation Storage
  - Transfer Submission Information Packages (SIPs) to staging area for processing with Archivematica

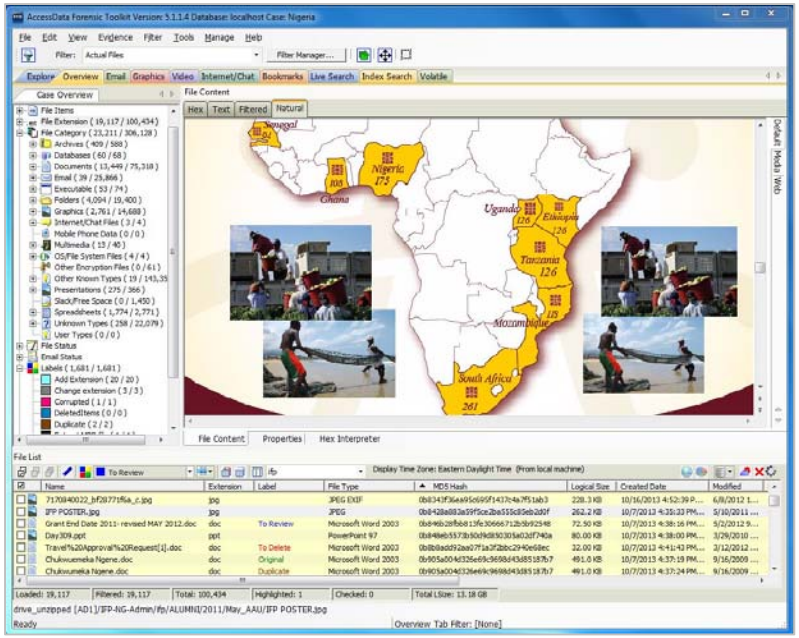
makeInventory

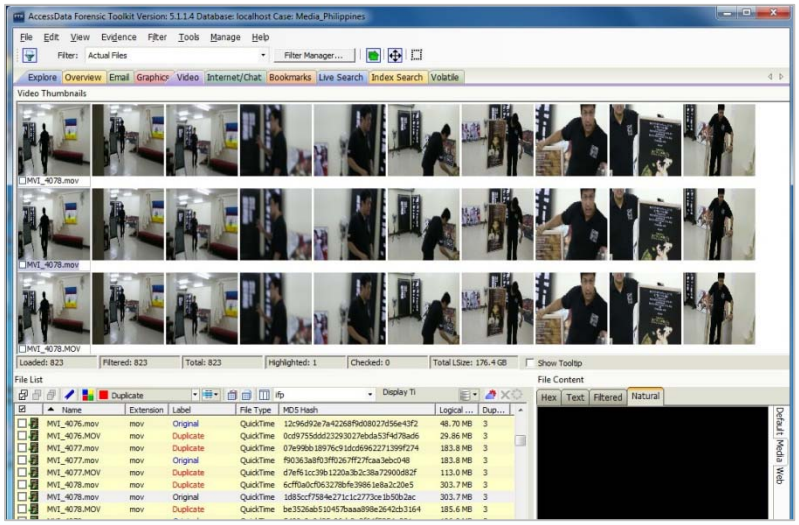


- Windows program based on Hashdeep
- Records filenames/paths, file sizes, checksums in MD5 and SHA formats
- Retains filenames in their original languages
- Run on transfer media by both content donors and Columbia Libraries
- Inventories are compared to ensure content integrity

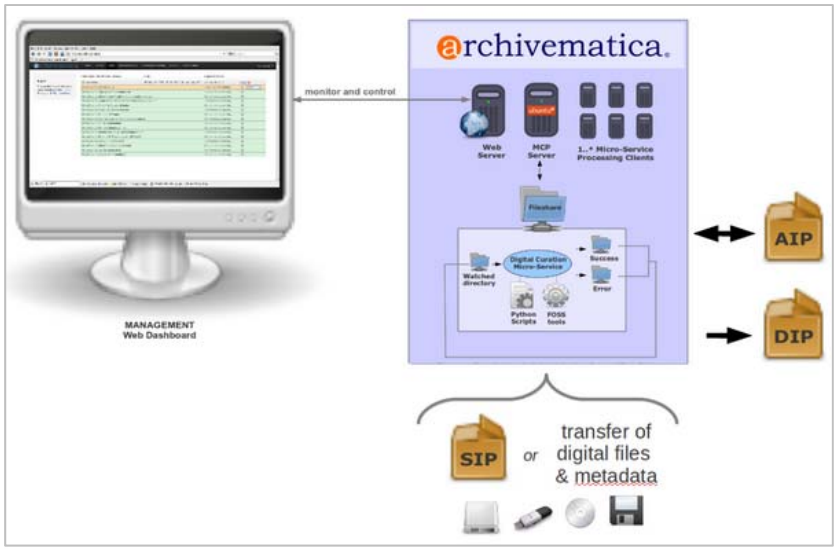
Forensic Toolkit

- Displays number and types of files
- Displays the file content and metadata
- Identifies system, password-protected, and duplicate files
- Restores corrupted files
- Allows searching for Personally Identifiable Information
- Creates periodic thumbnails for videos
- Allows assigning labels to individual files or groups of files
- Generates customizable reports





Archivematica: Overview



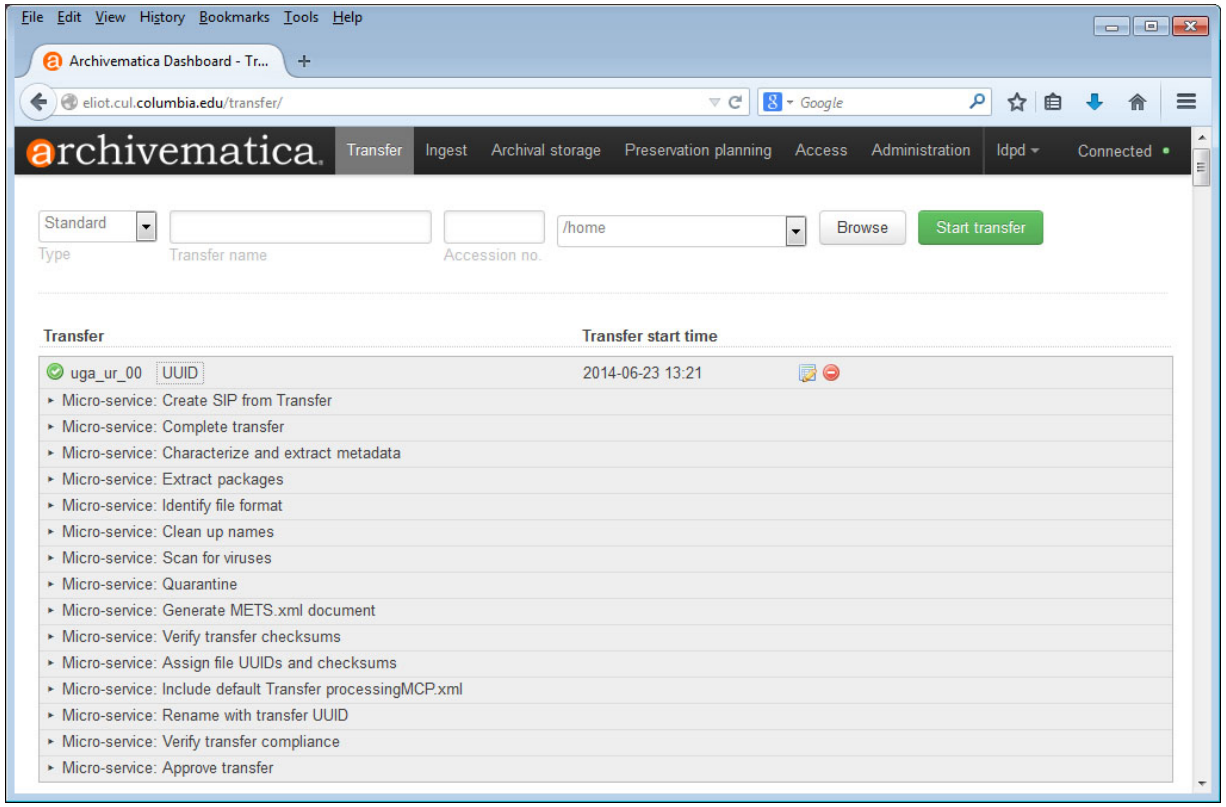
- Open-source OAIS-compliant digital preservation system
- Compiles SIPs and produces AIPs/DIPs
- Preserves files in original formats and normalizes them to preservation/access formats
- Generates METS files containing technical, structural, descriptive, rights, and PREMIS preservation metadata
- Access: ICA-AtoM, DSpace, CONTENTdm

Archivematica: Content Preparation

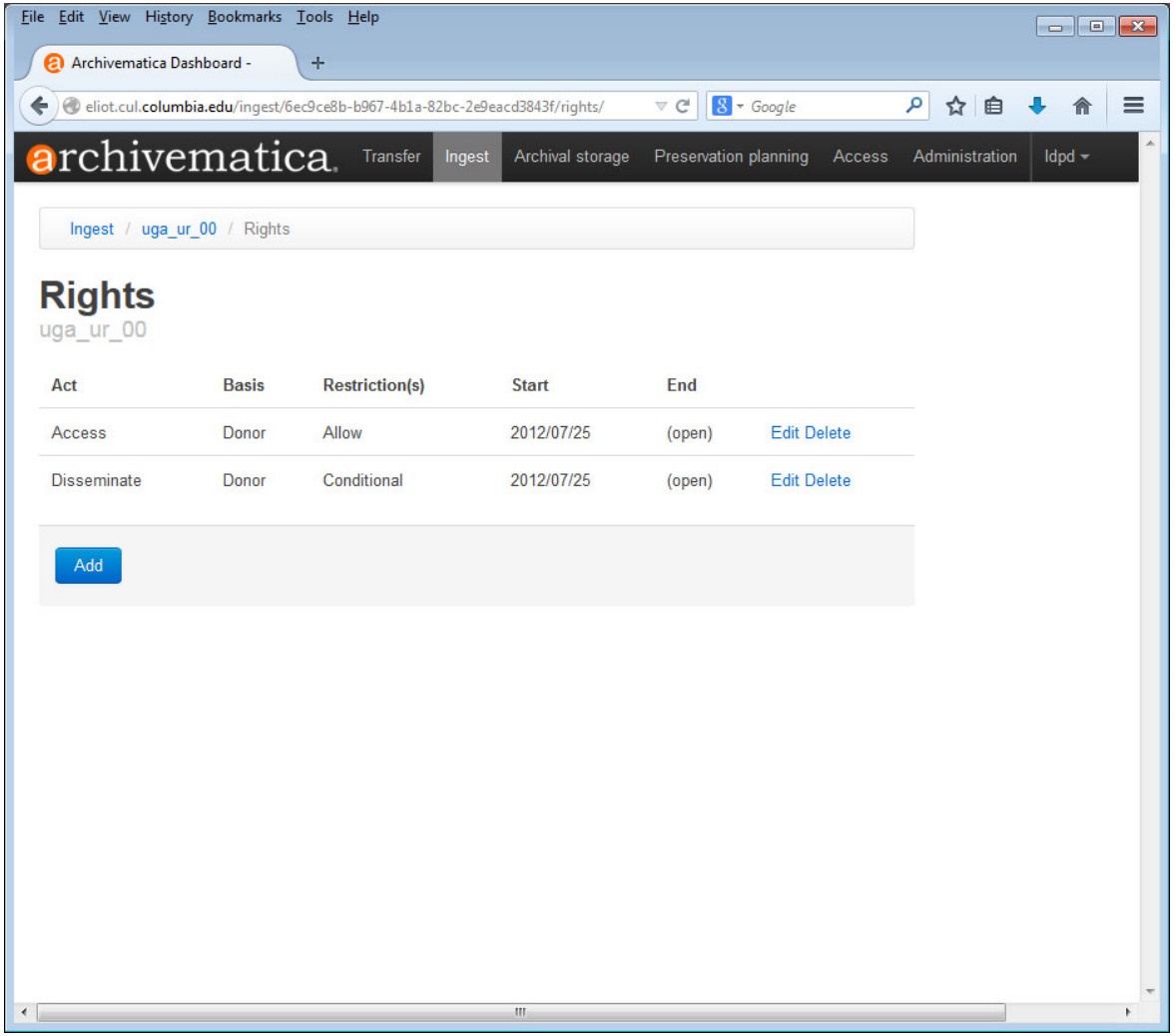
- Content pre-processing:
  - Convert email from multiple formats (eml, mbx, msg, pst, sbd, Pegasus mail) to MBOX
  - Convert Microsoft Access databases to XML format
  - Outsource conversion of content of commercially produced video DVDs, audio CDs, and mini DV-tapes to preservation formats
  - Extract data from ZIP and RAR archives
- Compiling SIPs:
  - Unrestricted, Onsite, Restricted for each office
  - SIP size can be limited
  - Number of files in AIP < 1100

Archivematica: SIPs

- Assign unique IDs
- Verify content integrity
- Perform virus check
- Clean up filenames
- Perform file format identification
- Extract metadata
- Generate METS.xml file



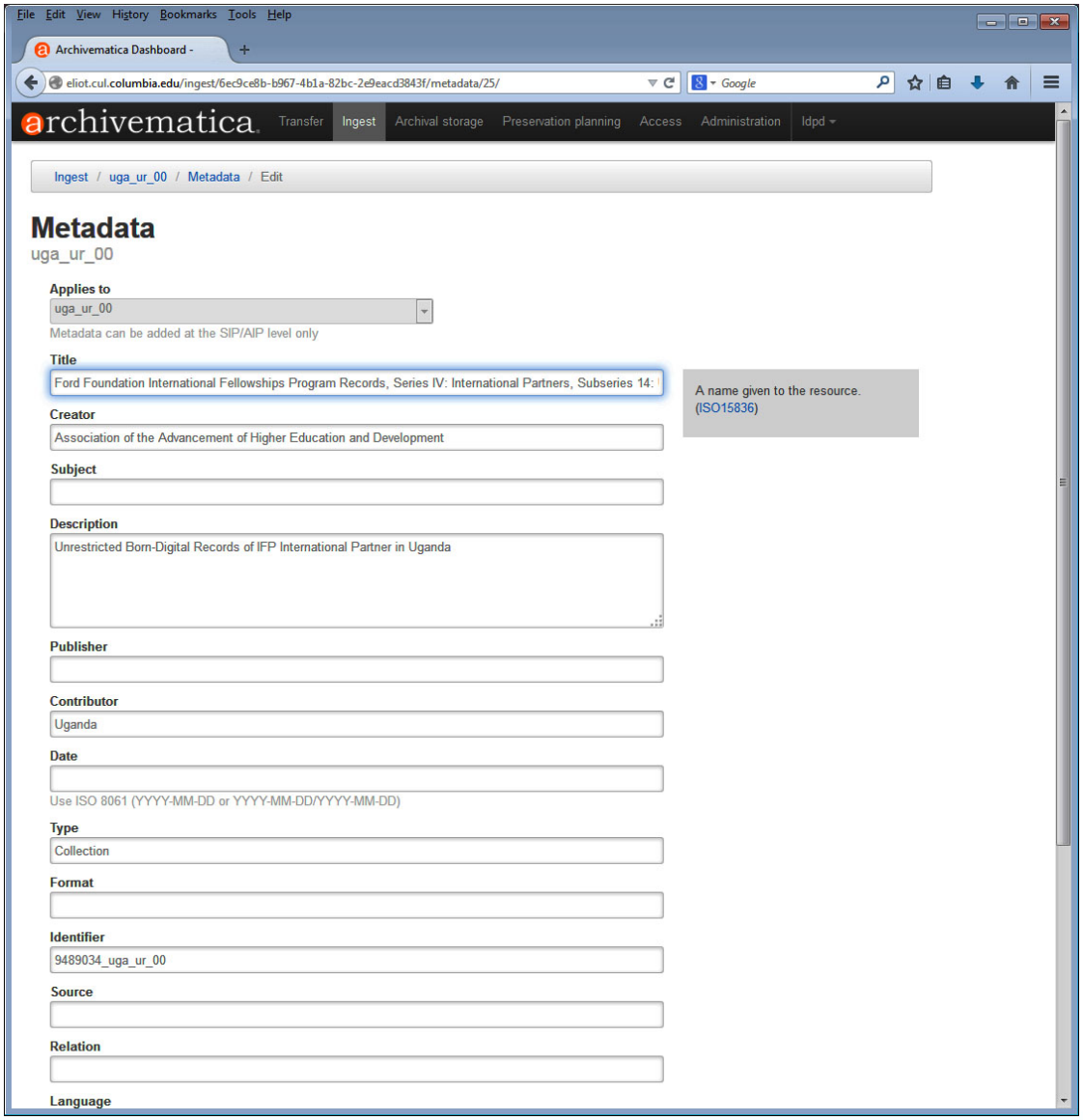
Archivematica: Rights Metadata



- PREMIS rights at the SIP level

Archivematica: Descriptive Metadata

- Dublin Core metadata at the SIP level

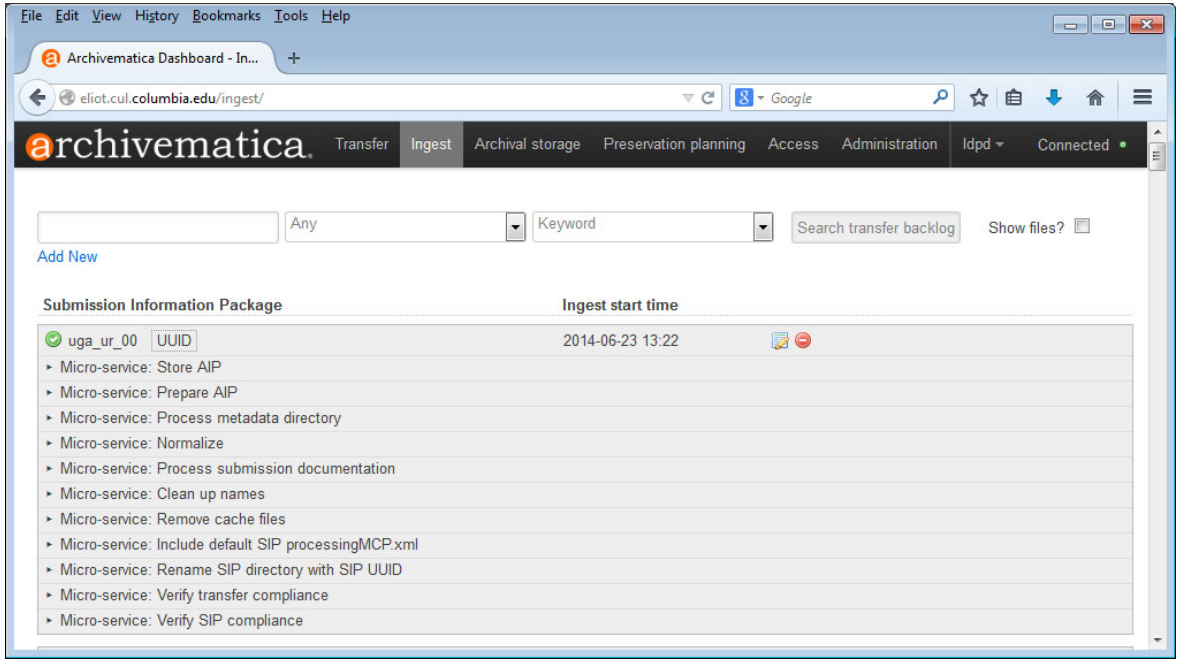


Archivematica: Filenames

- Original and normalized filenames in METS file:


```
C:\Users\ds2057\Desktop   C:\Users\ds2057\Desktop\I...
Convert  Select
                    </WORD>
                  </tool>
                - <tool version="0.1" name="OIS File Information">
                  - <fits
                    xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits_output
                    http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd">
                      - <fileinfo>
                          <filepath>/var/archivematica/sharedDirectory/watchedDirectories/workFlowDecisions/
                            eed3b7ec-b74a-40fc-93f9-
                            618b3285d890/objects/Office_Address_Ban_Gong_Di_Zhi_.doc</filepath>
                          <filename>/var/archivematica/sharedDirectory/watchedDirectories/workFlowDecisions
                            eed3b7ec-b74a-40fc-93f9-
                            618b3285d890/objects/Office_Address_Ban_Gong_Di_Zhi_.doc</filename>
                          <size>19456</size>
                          <md5checksum>19c55a4f52d46aba40a3c7ae37219640</md5checksum>
                          <fslastmodified>1405374947000</fslastmodified>
                        </fileinfo>
                    </fits>
                  </tool>
                </toolOutput>
              </fits>
            </objectCharacteristicsExtension>
          </objectCharacteristics>
          <originalName>%transferDirectory%objects/Office Address 办公地址.doc</originalName>
        </object>
</xmlData>
dWrap>
D>
vMD ID="digiprovMD_11">
Wrap MDTYPE="PREMIS:EVENT">
<xmlData>
```

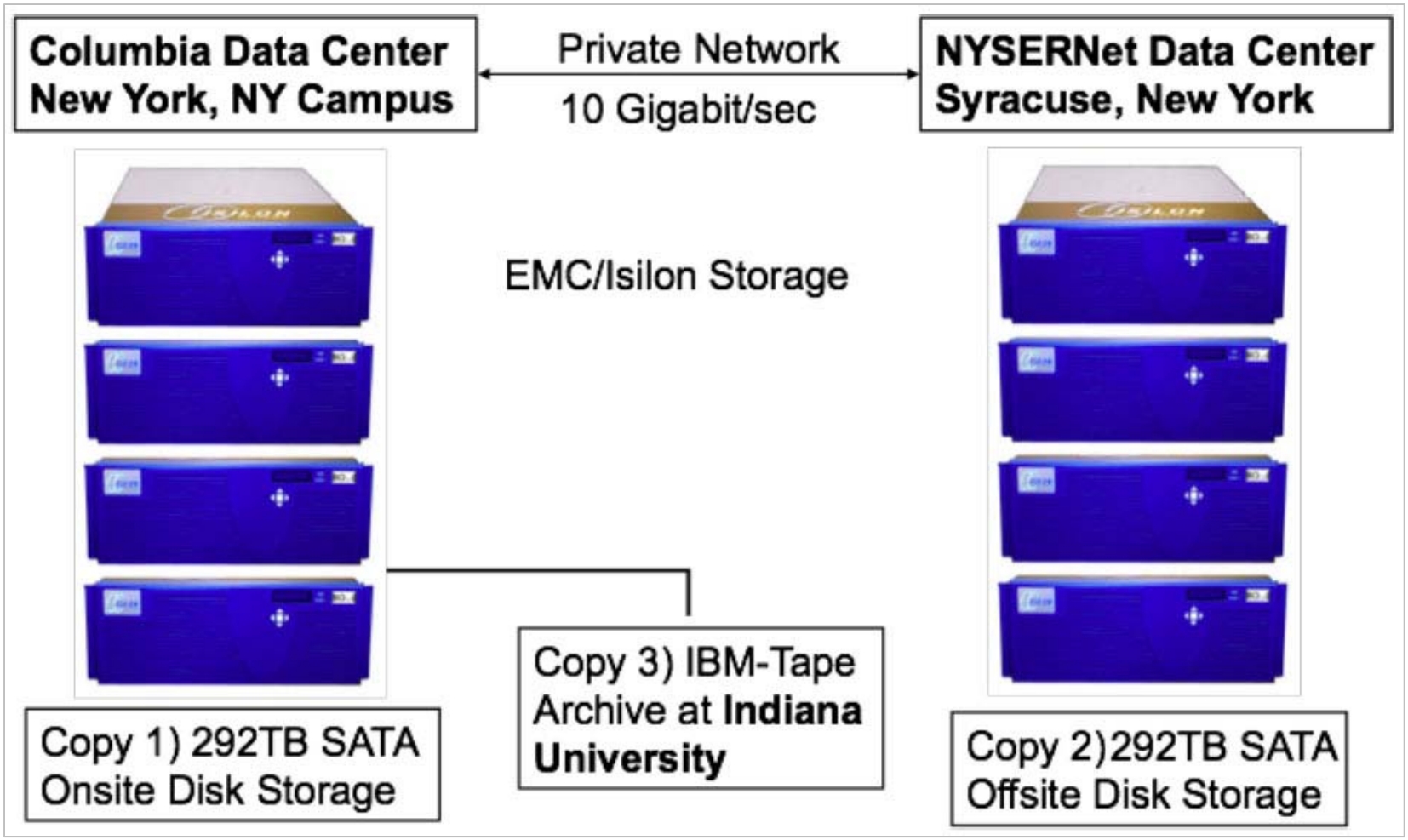
Archivematica: AIPs



- Normalize objects for preservation
- Populate METS.xml
- Generate and store AIP

Preservation Storage

- AIPs in Bagit format are ingested into Preservation Repository



Thank you!

Questions? Contact us:
jg2138@columbia.edu
ds2057@columbia.edu