



Acquisition of Digital Records:

Lessons of the Ford Foundation International Fellowships Program Project

Jane Gorjevsky and Dina Sokolova
Columbia University

E-Archive Pilot Project

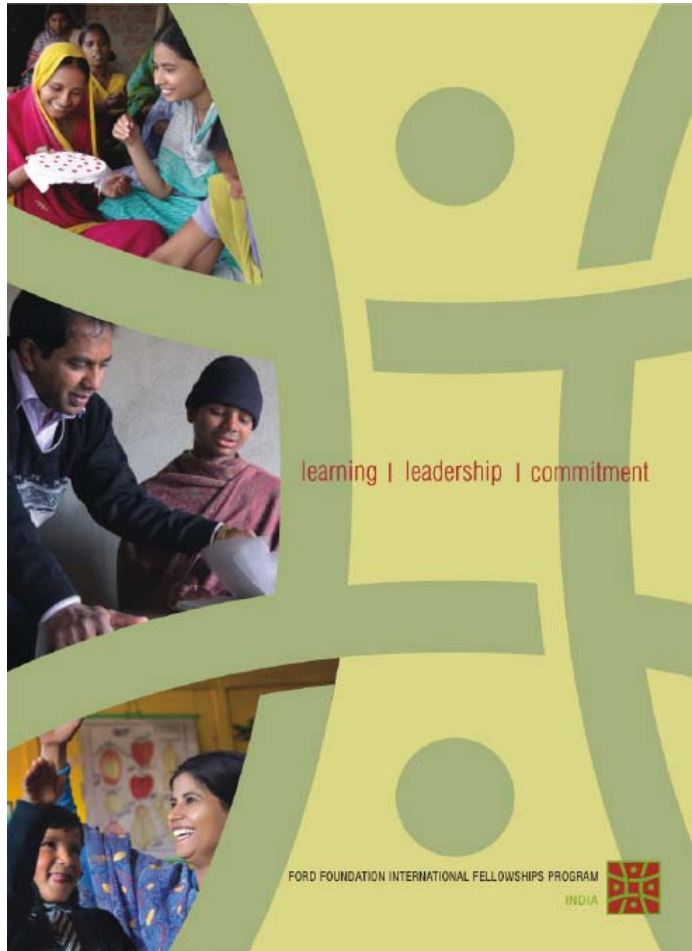
- ▶ Digital content acquisition procedures
- ▶ Hardware and software needs
- ▶ Sorting and weeding parameters and workflow
- ▶ Metadata creation or capture
- ▶ Preservation routines
- ▶ Access restrictions (tiered access)
- ▶ Finding aids and tools to view digital assets

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...**

Ford Foundation International Fellowships Program - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Ford Foundation International Fellowshi...

https://library.columbia.edu/bts/ford-ifp.html

COLUMBIA UNIVERSITY LIBRARIES / INFORMATION SERVICES

The IFP has since 2001 offered fellowships for post-graduate study to leaders from underserved communities in Asia, Africa, Latin America, and Russia, and will complete its work in 2014. Their archives include documentation and videos of the more than 3,300 IFP fellows who passed through the program as well as comprehensive planning and administrative files.

The completed IFP archive at Columbia will be of great interest to researchers and practitioners interested in the progress of social justice, community development, and access to higher education. Access to the paper and electronic archives will be integrated, offering researchers a fast and comprehensive way to study the content.

One of the key objectives of the grant is to enable Columbia Libraries to build out a full set of repository-based systems and services so that it can more easily acquire, ingest, process, preserve and make accessible both the paper and born-digital organizational records. The technological infrastructure built for this project will ultimately allow Columbia act as the central repository for the electronic records of other institutions whose archives are deposited at Columbia.

Information for International Partners

Electronic Manifest

The makelInventory program generates a complete listing of all the folders and files on the transportation drive, including file attributes. It will be compared to the listing generated from the same drive after it arrives at Columbia Libraries, helping us to ensure that no data loss or corruption occurs during the transfer.

makelInventory Instructions:

1. Click on "Download makelInventory.zip" below to download the program.
2. Unzip the folder.
3. Copy makelInventory folder into the main directory of the flash/hard drive.
4. Open makelInventory folder in Windows Explorer.
5. Drag the flash/hard drive icon into make_inventory.bat icon.
6. The cmd.exe window should open indicating that the process has started.
7. When the "Press any key to continue" prompt appears, press a key to close the window.
8. The fileInventory.xml file should appear in makelInventory folder (it might take several minutes to more than an hour for the process to finish, depending on the disk size).
9. Make sure that fileInventory.xml file is not empty.

[makelInventory: Step-by-step instructions](#)

[Download makelInventory.zip](#)

Archiving Email in Microsoft Outlook

1. Open Microsoft Outlook.
2. Navigate to the Archive window:
 - In Outlook 2003/2007: from the File menu, select Archive.
 - In Outlook 2010:

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_05_31_2013 [Read-Only] - Microsoft Word

IFP_records_access_rights(2012~2075)[¶]
(based-on-Transfer-Agreement-and-Deed-of-Gift)[¶]

UserGroup [¶]	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,IE) [¶]
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) [¶]
IIIE*staff-and-other-individuals-designated-by-IIIE [¶]	Yes [¶] (records-created-by-IIIE;records-created-by-IFP-in-case-of-legal-claim-(CUI)-should-provide-access-or-send-documents-to-IIIE) [¶]	Yes [¶] (same-as-researchers) [¶]	No [¶]	Yes [¶] (records-created-by-IIIE) [¶]

*IFP and IIIE also refer to the Oversight Bodies if IIIE 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:

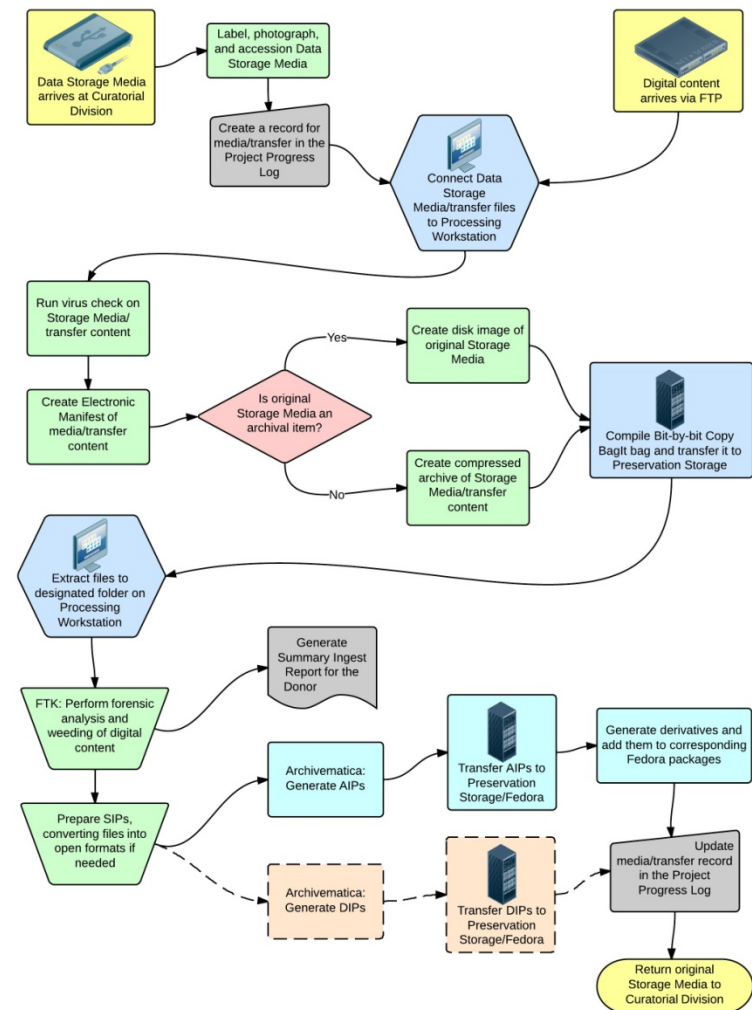
IFP\Newsletter\Alumni Meeting\... \...\...\Fifth meeting
October 23-28, 2008\Agenda\IFP Assembly\Other\07.jpg

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



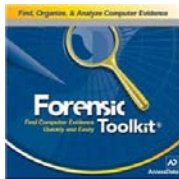
Technological Tools



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

Hashdeep

- makeInventory program



- Forensic Toolkit (FTK)

archivematica®

- Archivematica

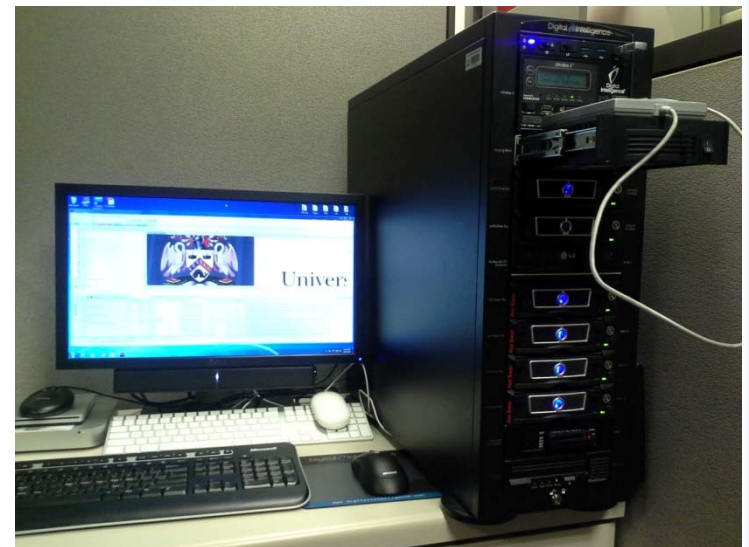
Processing Workstation

➤ **FRED:**

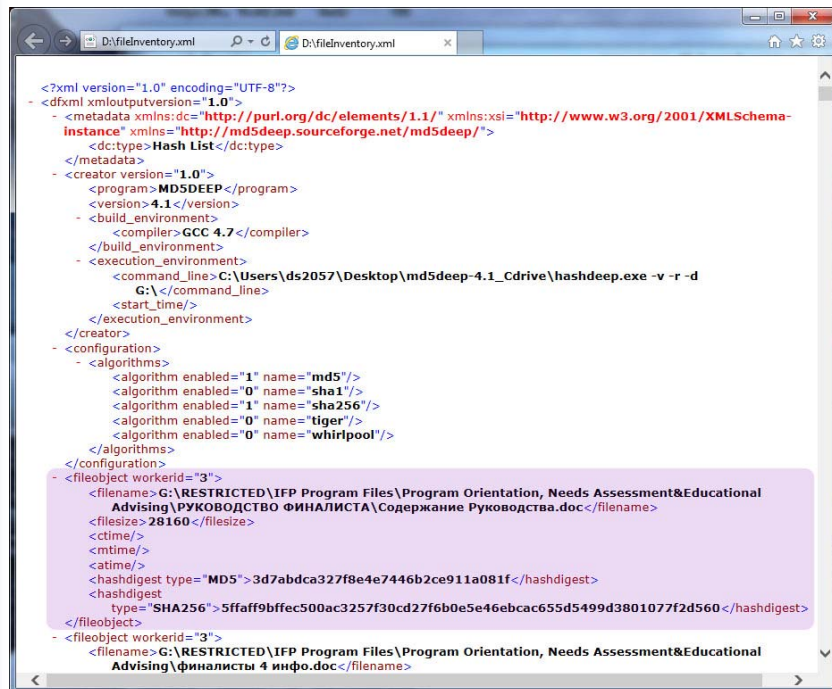
- Create bit-by-bit copy of the original transfer and metadata using write-blocking device and external disk drives (PC-formatted storage media)
- Perform content analysis and selection using Forensic Toolkit

➤ **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

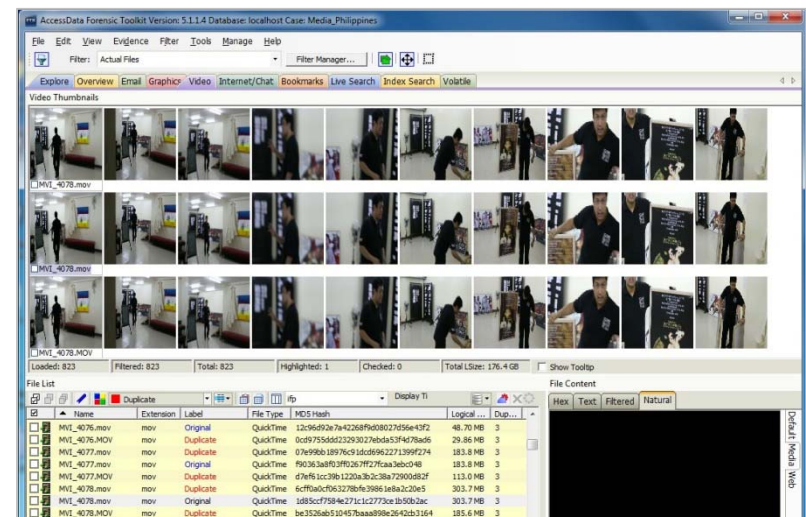
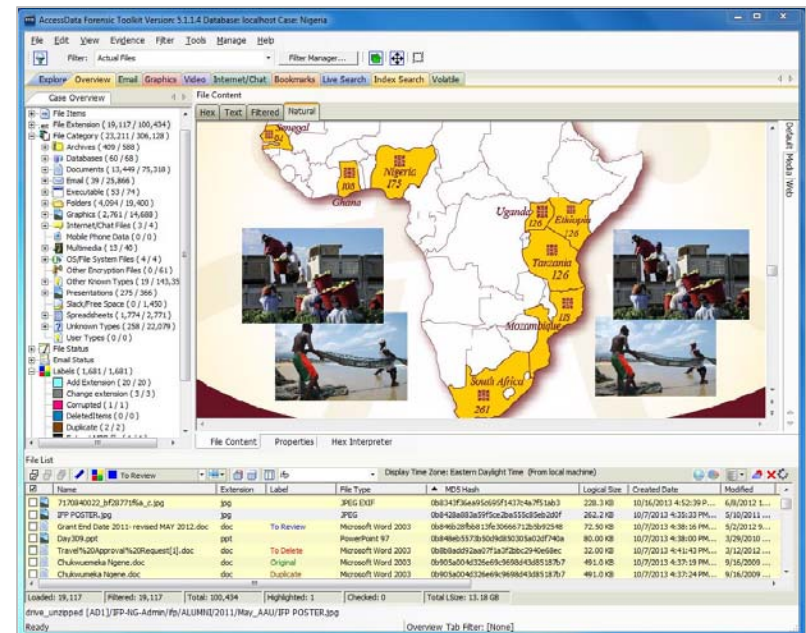


```
<?xml version="1.0" encoding="UTF-8"?>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://md5deep.sourceforge.net/md5deep/">
  <dc:type>Hash List</dc:type>
</rdf:RDF>
- <creator version="1.0">
  <program>MD5DEEP</program>
  <version>4.1</version>
  <build_environment>
    <compiler>GCC 4.7</compiler>
  </build_environment>
  <execution_environment>
    <command_line>C:\Users\ds2057\Desktop\md5deep-4.1_Cdrive\hashdeep.exe -v -r -d
    G:\</command_line>
    <start_time/>
  </execution_environment>
</creator>
- <configuration>
  <algorithms>
    <algorithm enabled="1" name="md5"/>
    <algorithm enabled="0" name="sha1"/>
    <algorithm enabled="1" name="sha256"/>
    <algorithm enabled="0" name="tiger"/>
    <algorithm enabled="0" name="whirlpool"/>
  </algorithms>
</configuration>
- <fileobject workerid="3">
  <filename>G:\RESTRICTED\IFP Program Files\Program Orientation, Needs Assessment&Educational
  Advising\РУКОВОДСТВО ФИНАЛИСТА\Содержание Руководства.doc</filename>
  <filesize>28160</filesize>
  <ctime/>
  <mtime/>
  <atime/>
  <hashdigest type="MD5">3d7abdca327f8e4e7446b2ce911a081f</hashdigest>
  <hashdigest
  type="SHA256">5ffaff9bffc500ac3257f30cd27f6b0e5e46ebcac655d5499d3801077f2d560</hashdigest>
</fileobject>
- <fileobject workerid="3">
  <filename>G:\RESTRICTED\IFP Program Files\Program Orientation, Needs Assessment&Educational
  Advising\финалисты 4 инфо.doc</filename>
```

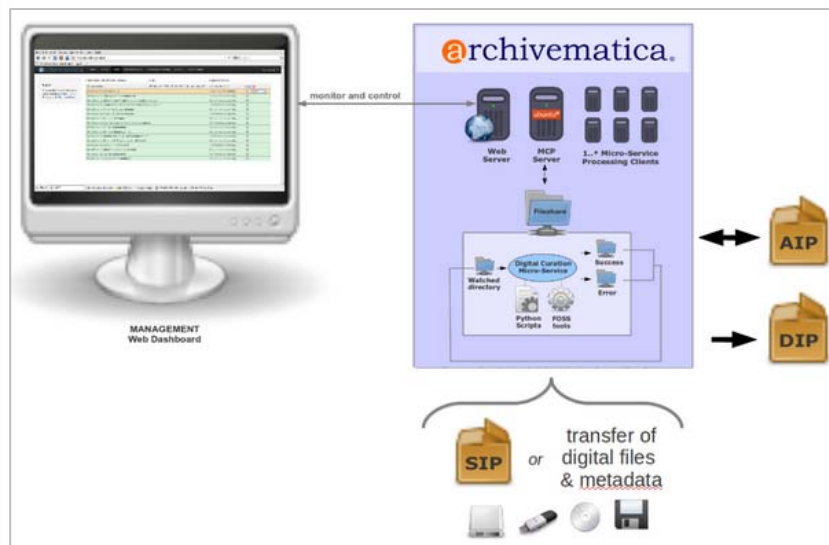
- Windows program based on Hashdeep
- Records filenames/paths, file sizes, checksums in MD5 and SHA formats
- Retains filenames in their original languages
- Run on transfer media by both content donors and Columbia Libraries
- Inventories are compared to ensure content integrity

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



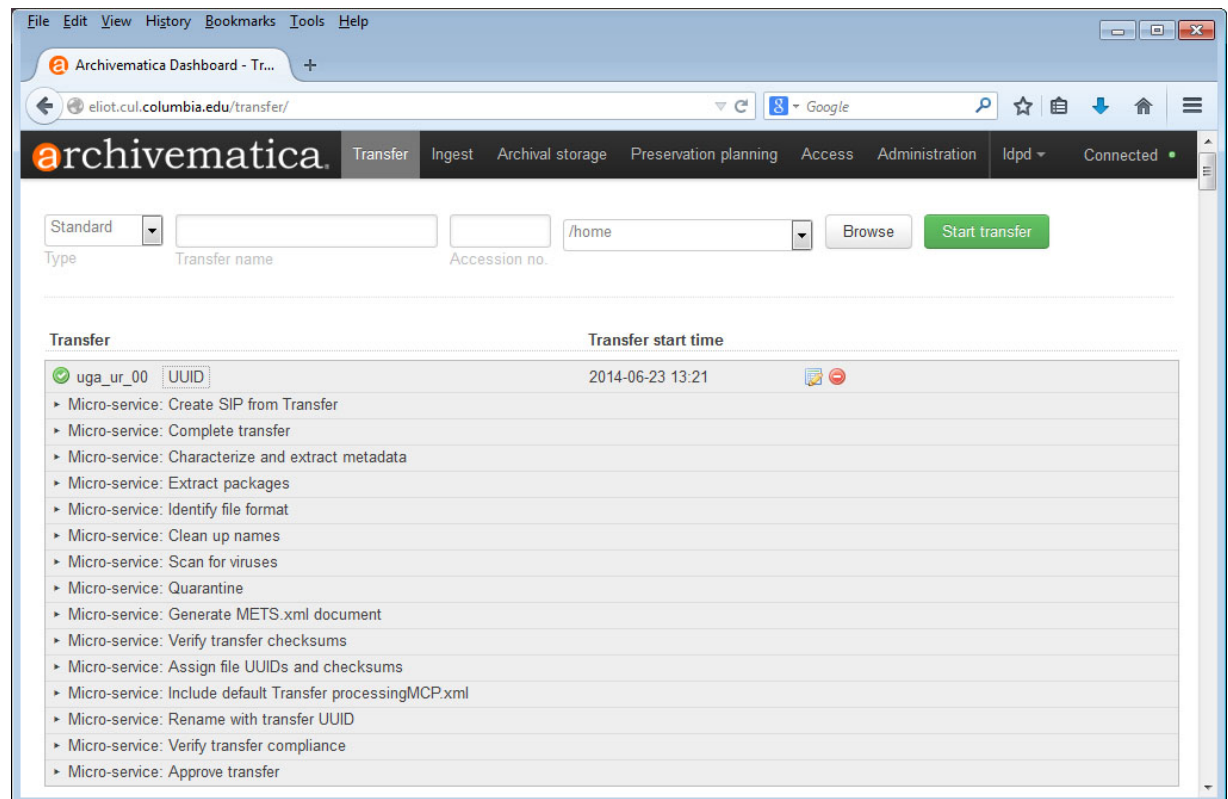
- Open-source OAIS-compliant digital preservation system
- Compiles SIPs and produces AIPs/DIPs
- Preserves files in original formats and normalizes them to preservation/access formats
- Generates METS files containing technical, structural, descriptive, rights, and PREMIS preservation metadata
- Access: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

The screenshot shows the Archivematica web interface. The browser address bar displays the URL: `eliot.cul.columbia.edu/ingest/6ec9ce8b-b967-4b1a-82bc-2e9eacd3843f/rights/`. The Archivematica navigation bar includes links for Transfer, Ingest, Archival storage, Preservation planning, Access, Administration, and Idpd. The breadcrumb trail indicates the current location: `Ingest / uga_ur_00 / Rights`. The main heading is "Rights" for the object `uga_ur_00`. Below this is a table with the following data:

Act	Basis	Restriction(s)	Start	End	
Access	Donor	Allow	2012/07/25	(open)	Edit Delete
Disseminate	Donor	Conditional	2012/07/25	(open)	Edit Delete

Below the table is an "Add" button.

Archivematica: Descriptive Metadata

- Dublin Core metadata at the SIP level

The screenshot shows the Archivematica web interface in a browser window. The address bar shows the URL: `eliot.cul.columbia.edu/ingest/6ec9ce8b-b967-4b1a-82bc-2e9eacd3843f/metadata/25/`. The browser's address bar also shows "Google". The Archivematica logo is in the top left, and navigation tabs for "Transfer", "Ingest", "Archival storage", "Preservation planning", "Access", "Administration", and "Idpd" are in the top right. Below the navigation bar, a breadcrumb trail reads "Ingest / uga_ur_00 / Metadata / Edit".

The main content area is titled "Metadata" and shows the identifier "uga_ur_00". Below this, a dropdown menu labeled "Applies to" is set to "uga_ur_00". A note states: "Metadata can be added at the SIP/AIP level only".

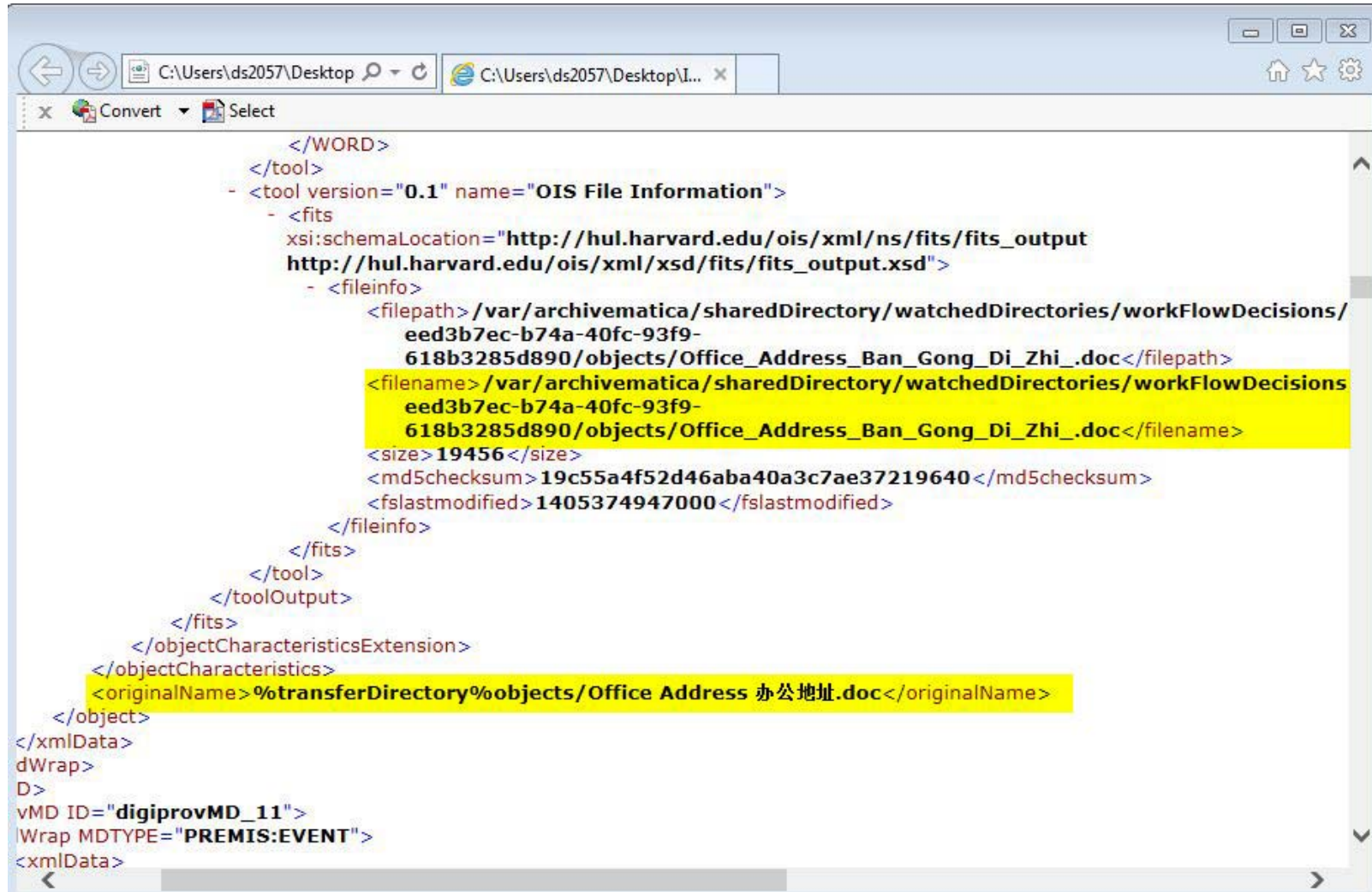
The "Title" field is highlighted with a blue border and contains the text: "Ford Foundation International Fellowships Program Records, Series IV: International Partners, Subseries 14:". To the right of this field is a grey box with the text: "A name given to the resource. (ISO15836)".

Other fields include:

- Creator:** Association of the Advancement of Higher Education and Development
- Subject:** (empty field)
- Description:** Unrestricted Born-Digital Records of IFP International Partner in Uganda
- Publisher:** (empty field)
- Contributor:** Uganda
- Date:** (empty field) with a note: "Use ISO 8061 (YYYY-MM-DD or YYYY-MM-DD/YYYY-MM-DD)"
- Type:** Collection
- Format:** (empty field)
- Identifier:** 9489034_uga_ur_00
- Source:** (empty field)
- Relation:** (empty field)
- Language:** (empty field)

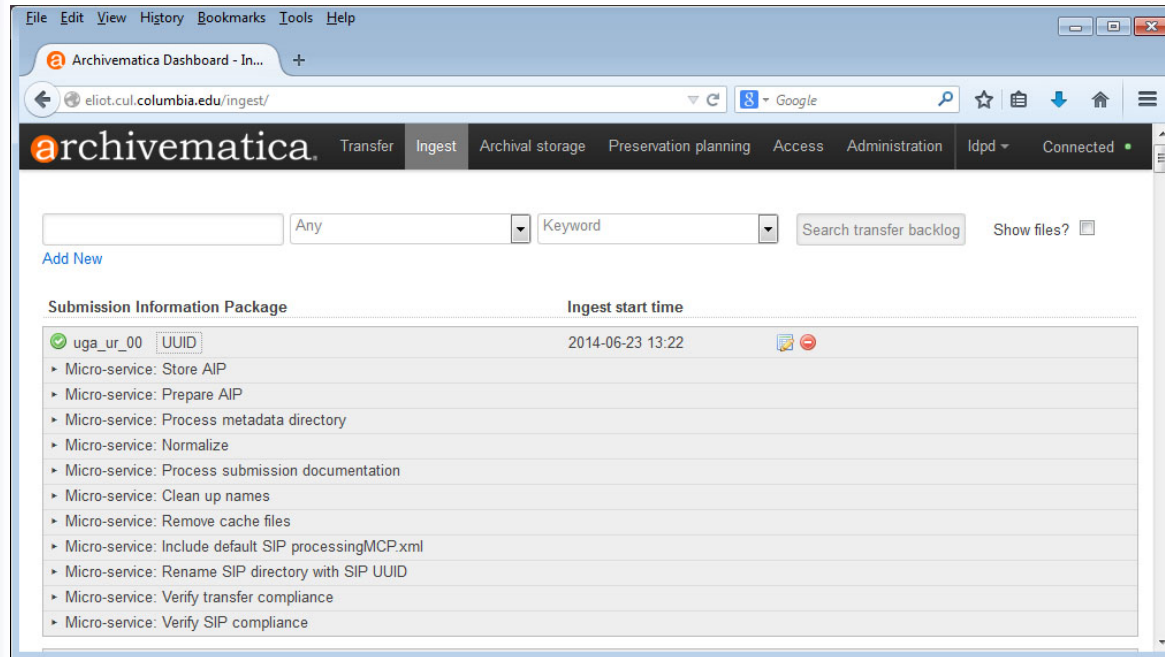
Archivematica: Filenames

- Original and normalized filenames in METS file:



```
</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>
```

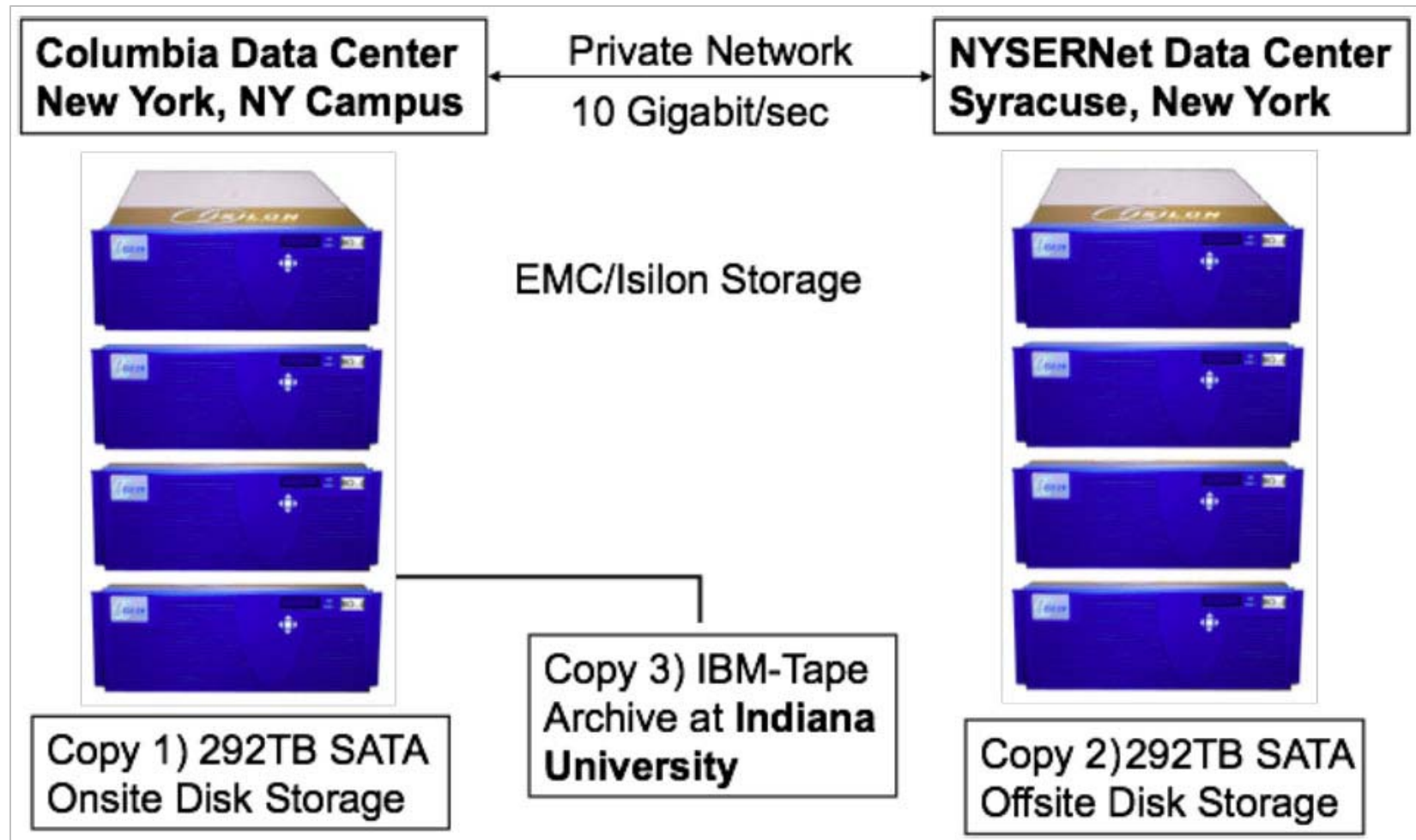
Archivematica: AIPs



- 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