

# Infrastructure Development: Multiple Digital Content Types in a Single Collection

Dina Sokolova and Jane Gorjevsky,  
*Columbia University*

# Digital Content in Columbia Special Collections

- ▶ Digitized
- ▶ Digital-born records (modern and legacy)
- ▶ Delivered-digital
- ▶ Harvested online materials

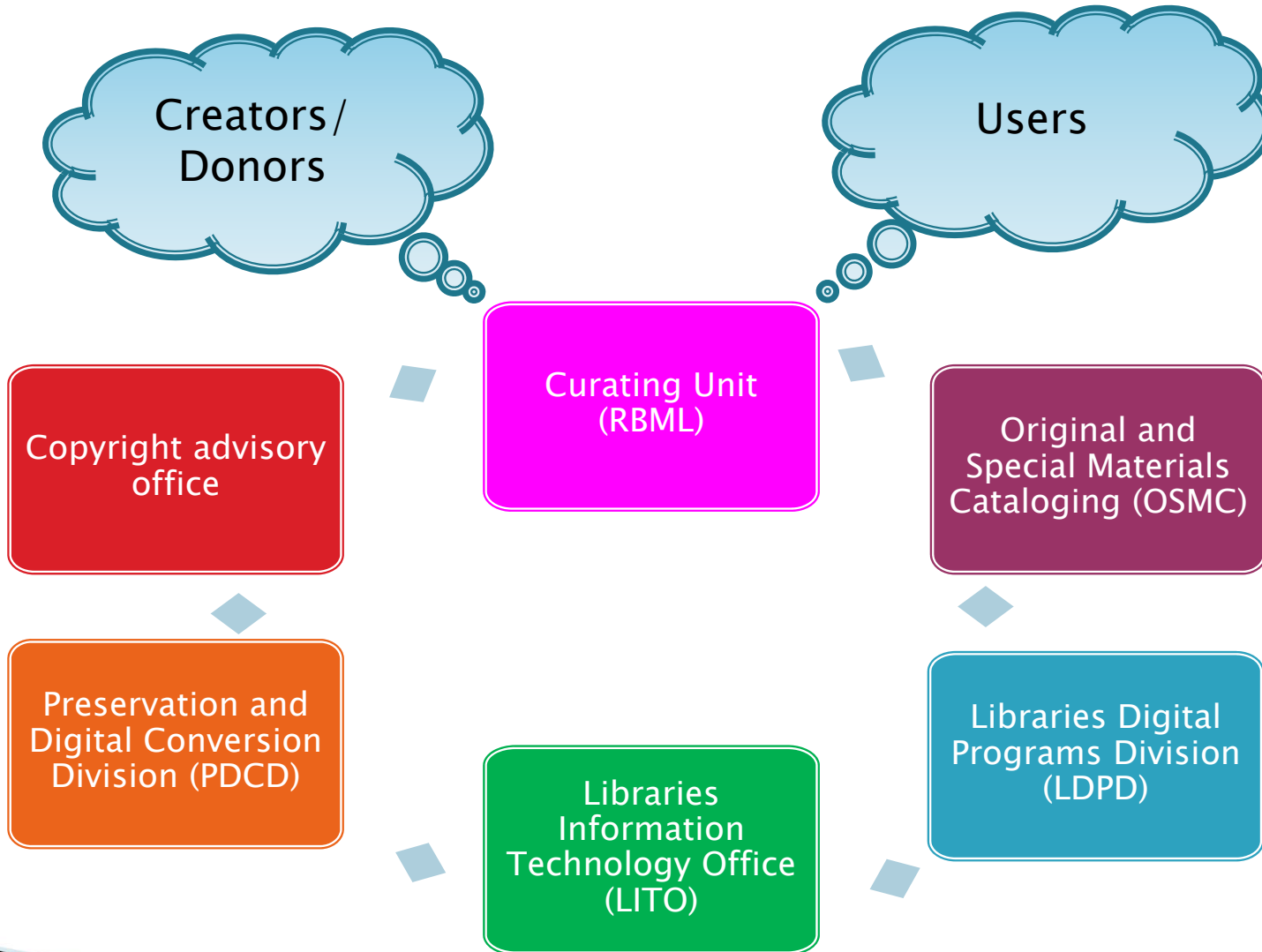


# Digital Content Pilot Project

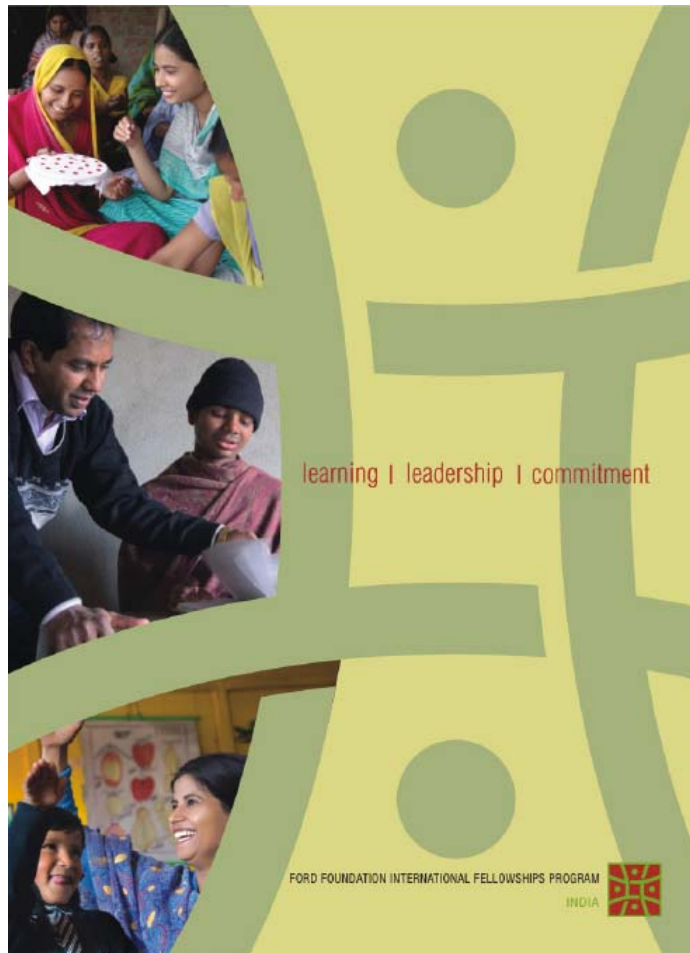
- ▶ Acquisition procedures
- ▶ Hardware and software
- ▶ Sorting and weeding workflow
- ▶ Metadata capture and enhancement
- ▶ Preservation routines for various content types and file formats
- ▶ Finding aids for hybrid digital/analog collections
- ▶ Tiered access system



# Organizational Roles



# Ford Foundation International Fellowships Program



- ▶ 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 and a web interface

*Funded by Ford Foundation grant,  
October 2011*



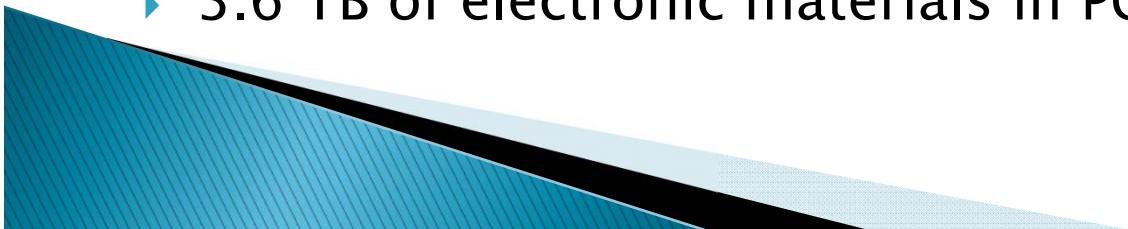
# International Fellowships Program

- ▶ 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 2001 - 2013



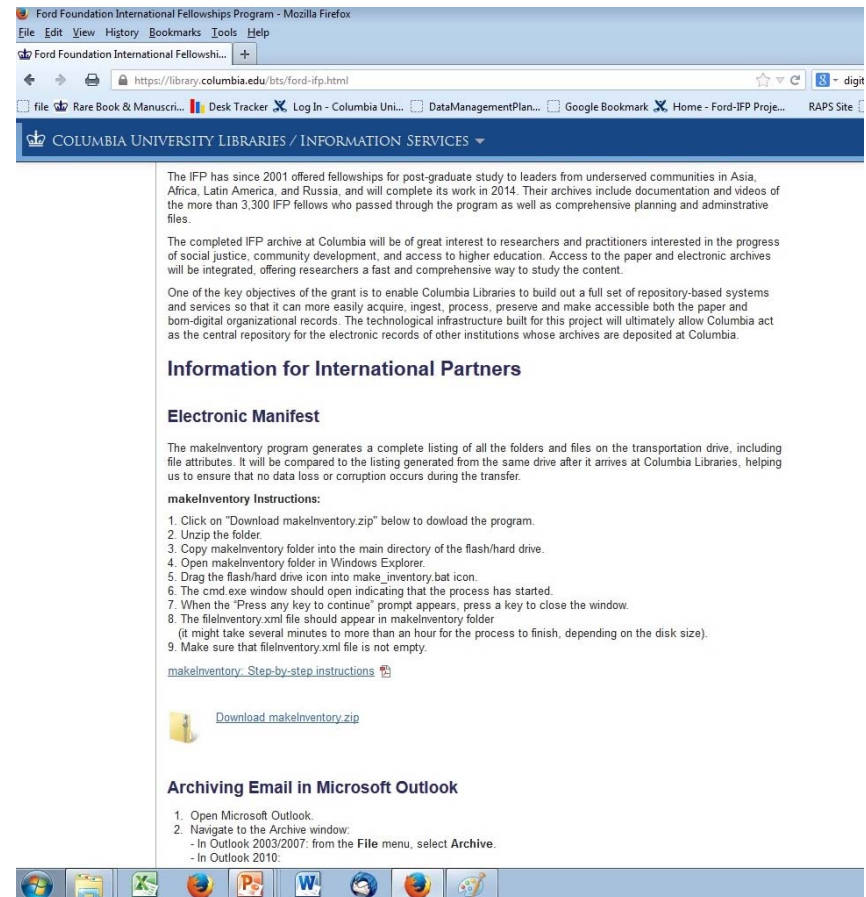
# Records Scope and Content

- ▶ 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



# Work with Donors

- ▶ Record surveys (2010, 2012)
- ▶ Selection and sorting guidelines
- ▶ Transfer instructions and tools
- ▶ Record Samples
- ▶ Additional file/folder naming, selection and sorting recommendations



The screenshot shows a Mozilla Firefox browser window displaying the Ford Foundation International Fellowships Program page. The address bar shows the URL: <https://library.columbia.edu/bts/ford-ifp.html>. The page content includes:

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 makeInventory 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.

**makeInventory Instructions:**

1. Click on "Download makeInventory.zip" below to download the program.
2. Unzip the folder.
3. Copy makeInventory folder into the main directory of the flash/hard drive.
4. Open makeInventory 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 makeInventory 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.

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

[Download makeInventory.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:



# Archiving Web Resources

The screenshot shows a search results page on archive.org. The browser address bar displays the URL: <https://archive-it.org/collectors/Z766ffc=websiteGroup%3AForc+Foundation+International+Fellowship+Program>. The page title is 'IFP International Partner (1)'. The search results are sorted by 'Count' and show 'Page 1 of 1 (25 Total Results)'. The results are categorized by 'Creator', 'Language', 'Coverage', and 'Collector'. The first result is for the 'Center for Educational Exchange with Vietnam: International Fellowships Program (IFP)' with URL <http://ceevn.acls.org/ceevn/ifpinfo.htm>. The second result is for 'Fundación Equitas' with URL <http://fundacionequitas.org/>. The third result is for 'IFP Tanzania' with URL <http://ifptanzania.esrftz.org/>. Each result includes a description, subject, group, creator, language, coverage, and collector information.

- ▶ CUL program administered by OSMC using archive.org toolset
- ▶ RBML:
  - Verifies URLs
  - Provides metadata
  - Specifies capturing frequency
  - Monitors captures
  - Adds descriptive metadata

# Transferring Digital Files

## Hard drives or flash drives

RBML:  
Received  
Accessioned  
Housed  
Labeled  
Forwarded to LDPD for  
transfer

LDPD:  
Data transferred to server  
Inventory and Initial ingest  
report generated

## Obsolete media

RBML:  
Found  
Inventoried using template  
Separated  
Forwarded to LDPD for  
transfer

LDPD:  
Data transferred to server  
Initial ingest report  
generated

Original Media  
returned to  
RBML,  
shipped  
offsite

Selection and  
weeding using  
digital  
forensics  
tools (LDPD,  
RBML)

# Initial Assumptions

- ▶ Most materials in English
- ▶ Pre-selected and sorted into 3 access categories
- ▶ No access to “embargoed files” 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



# 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 Digital Conversion
  - Standards
  - Vendor communications
  - Quality control



# Content Challenges

- ▶ Selection and sorting by creators 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



# 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

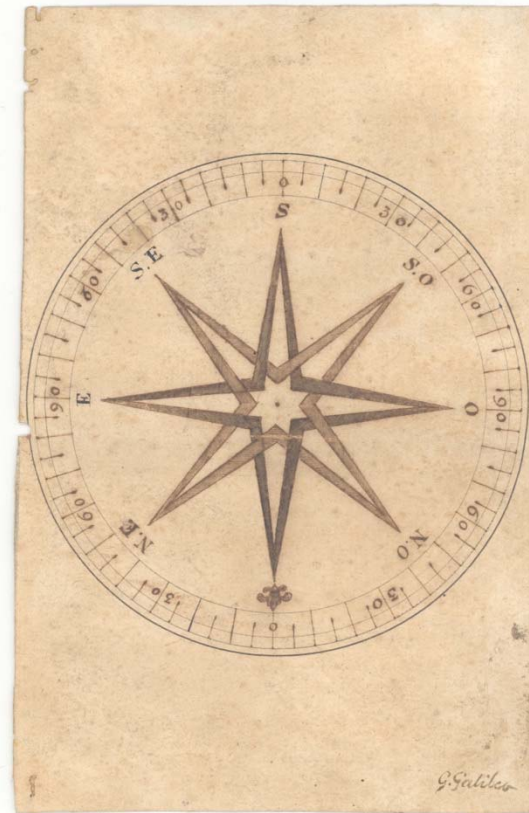
- ▶ Foreign languages:

- 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)

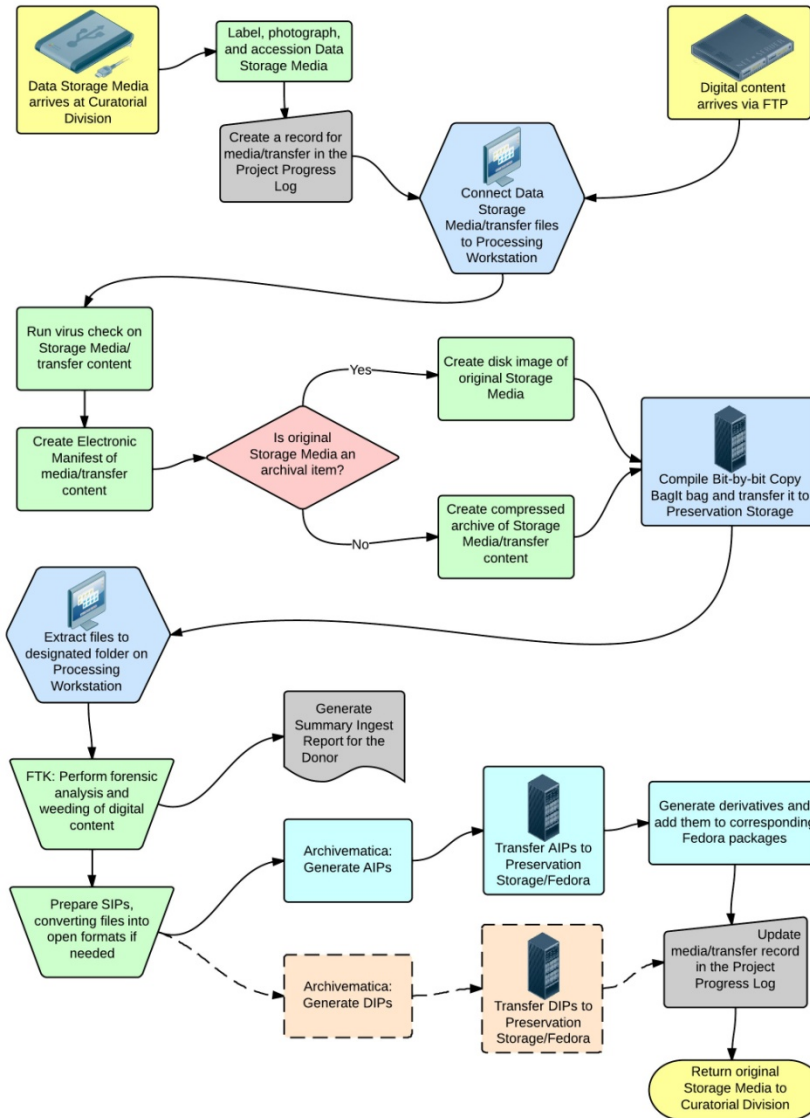


# Internal Guidelines and Documentation

- ▶ Curatorial surveys (pre-acquisition)
- ▶ Record Transfer Documentation
- ▶ Accessioning workflow
- ▶ Weeding routines
- ▶ File format action plan
- ▶ Pre-processing and ingest workflows



# Digital Preservation Workflow



# Processing Workstation



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

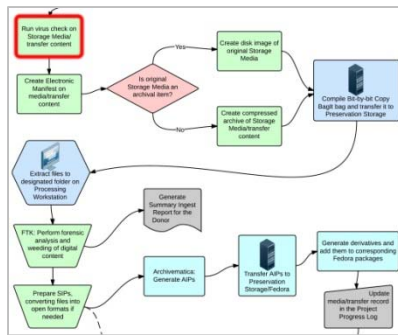




# Virus Check



ClamWin (ClamXav): initial virus check



ClamWin Free Antivirus: Scan Complete

Scan Started Mon Jun 30 13:46:09 2014

----- SCAN SUMMARY -----  
Known viruses: 3490903  
Engine version: 0.98.1  
Scanned directories: 66  
Scanned files: 9731  
**Infected files: 0**

Data scanned: 756.57 MB  
Data read: 4192673.16 MB (ratio 0.00:1)  
Time: 97.859 sec (1 m 37 s)

Completed

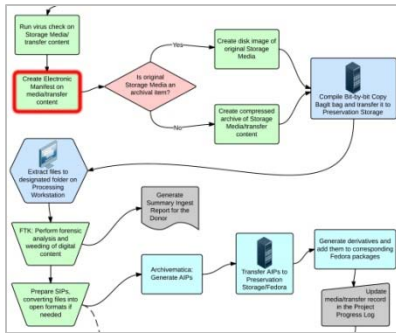
Save Report Close



# Electronic Manifest

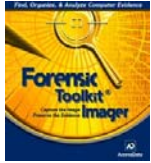
Hashdeep

makeInventory program: verifying content integrity

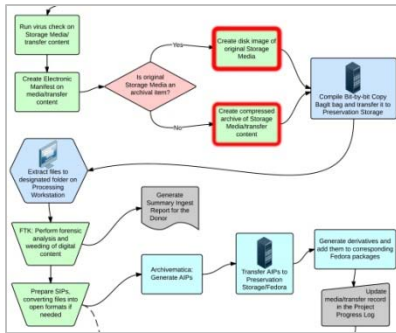


```
<?xml version="1.0" encoding="UTF-8"?>
- <dfxml xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://md5deep.sourceforge.net/md5deep/">
  <dc:type>Hash List</dc:type>
</metadata>
- <creator version="1.0">
  <program>MD5DEEP</program>
  <version>4.1</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>
```

# Disk Imaging



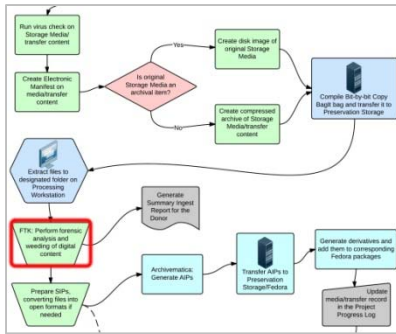
AccessData FTK Imager: creating disk images (CDs/DVDs, ZIP and Floppy Disks)





# Appraisal and Selection

## Forensic Toolkit (FTK): content review and weeding



AccessData Forensic Toolkit Version: 5.1.1.4 Database: localhost Case: Nigeria

File Edit View Evidence Filter Tools Manage Help

Filter: Actual Files Filter Manager...

Explore Overview Email Graphics Video Internet/Chat Bookmarks Live Search Index Search Volatile

Case Overview File Content

File Items

- File Extension (19,117 / 100,434)
- File Category (23,211 / 306,128)
  - Archives (409 / 588)
  - Databases (60 / 68)
  - Documents (13,449 / 75,318)
  - Email (39 / 25,866)
  - Executable (53 / 74)
  - Folders (4,094 / 19,400)
  - Graphics (2,761 / 14,688)
  - Internet/Chat Files (3 / 4)
  - Mobile Phone Data (0 / 0)
  - Multimedia (13 / 40)
  - OS/File System Files (4 / 4)
  - Other Encryption Files (0 / 61)
  - Other Known Types (19 / 143,35)
  - Presentations (275 / 366)
  - Slack/Free Space (0 / 1,450)
  - Spreadsheets (1,774 / 2,771)
  - Unknown Types (258 / 22,079)
  - User Types (0 / 0)
- File Status
- Email Status
- Labels (1,681 / 1,681)
  - Add Extension (20 / 20)
  - Change extension (3 / 3)
  - Corrupted (1 / 1)
  - DeletedItems (0 / 0)
  - Duplicate (2 / 2)

File Content

Hex Text Filtered Natural

Senegal 104  
Ghana 108  
Nigeria 175  
Uganda 126  
Ethiopia 126  
Tanzania 126  
Mozambique 118  
South Africa 261

File List

Name	Extension	Label	File Type	MD5 Hash	Logical Size	Created Date	Modified
7170840022_bf28771f6a_c.jpg	jpg		JPEG EXIF	0b8343f36ea95c695f1437c4a7f51ab3	228.3 KB	10/16/2013 4:52:39 P...	6/8/2012 1...
IFP POSTER.jpg	jpg		JPEG	0b8428a883a59f5ce2ba555c85eb2d0f	262.2 KB	10/7/2013 4:35:33 PM...	5/10/2011 ...
Grant End Date 2011- revised MAY 2012.doc	doc	To Review	Microsoft Word 2003	0b846b28fbb813fe30666712b5b92548	72.50 KB	10/7/2013 4:38:16 PM...	5/2/2012 9...
Day309.ppt	ppt		PowerPoint 97	0b848eb5573b50d9d850305a02df740a	80.00 KB	10/7/2013 4:38:00 PM...	3/29/2010 ...
Travel%20Approval%20Request[1].doc	doc	To Delete	Microsoft Word 2003	0b8b8add92aa07f1a3f2bbc2940e68ec	32.00 KB	10/7/2013 4:41:43 PM...	3/12/2012 ...
Chukwumeka Ngene.doc	doc	Original	Microsoft Word 2003	0b905a004d326e69c9698d43d85187b7	491.0 KB	10/7/2013 4:37:19 PM...	9/16/2009 ...
Chukwumeka Ngene.doc	doc	Duplicate	Microsoft Word 2003	0b905a004d326e69c9698d43d85187b7	491.0 KB	10/7/2013 4:37:24 PM...	9/16/2009 ...

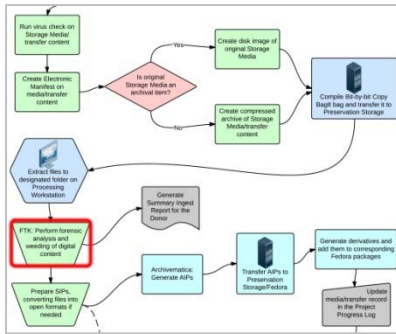
Loaded: 19,117 Filtered: 19,117 Total: 100,434 Highlighted: 1 Checked: 0 Total LSize: 13.18 GB

drive\_unzipped [AD1]/IFP-NG-Admin/IFP/ALUMNI/2011/May\_AAU/IFP POSTER.jpg

Ready Overview Tab Filter: [None]

# Appraisal and Selection (cont.)

## Forensic Toolkit (FTK): finding duplicate content




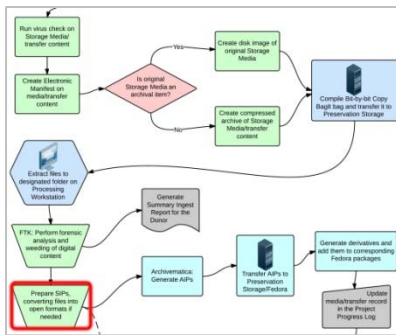
The screenshot displays the AccessData Forensic Toolkit (FTK) interface. The top menu includes File, Edit, View, Evidence, Filter, Tools, Manage, and Help. The main window shows a grid of video thumbnails for a file named 'MVI\_4078.MOV'. Below the thumbnails, a 'File List' table identifies duplicate files. The table has columns for Name, Extension, Label, File Type, MD5 Hash, Logical Size, and Duplication Count. The 'Label' column uses color coding: green for 'Original' and red for 'Duplicate'.


Name	Extension	Label	File Type	MD5 Hash	Logical Size	Duplication Count
MVI_4076.mov	mov	Original	QuickTime	12c96d92e7a42268f9d08027d56e43f2	48.70 MB	3
MVI_4076.MOV	mov	Duplicate	QuickTime	0cd9755ddd23293027ebda53f4d78ad6	29.86 MB	3
MVI_4077.mov	mov	Duplicate	QuickTime	07e99bb18976c91dcd6962271399f274	183.8 MB	3
MVI_4077.MOV	mov	Duplicate	QuickTime	f90363a8f03ff0267f27fcaa3ebc048	183.8 MB	3
MVI_4077.MOV	mov	Duplicate	QuickTime	d7ef61cc39b1220a3b2c38a72900d82f	113.0 MB	3
MVI_4078.mov	mov	Duplicate	QuickTime	6cffa0cf63278bfe39861e8a2c20e5	303.7 MB	3
MVI_4078.MOV	mov	Duplicate	QuickTime	1d85ccf7584e271c1c2773ce1b50b2ac	303.7 MB	3
MVI_4078.MOV	mov	Duplicate	QuickTime	be3526ab510457baaa898e2642cb3164	185.6 MB	3
MVI_4079.mov	mov	Original	QuickTime	5480c0ad0d35c96cb2a2f64f5254a231e	106.9 MB	3
MVI_4079.MOV	mov	Duplicate	QuickTime	537a0c435e81cb174082e66b50be27a3	106.9 MB	3
MVI_4079.MOV	mov	Duplicate	QuickTime	2c3871bd3de246974d6e93b5c6c64bb	68.01 MB	3
MVI_4080.mov	mov	Duplicate	QuickTime	0f8e882545006bb268327497aac12936	979.2 MB	3
MVI_4080.MOV	mov	Duplicate	QuickTime	014edad18375ccb844d5f70cf7ca92d1	979.2 MB	3
MVI_4080.MOV	mov	Duplicate	QuickTime	031cc13c2e615f1b91a2be3fe58869c	566.2 MB	3
MVI_4081.MOV	mov	Duplicate	QuickTime	c64659c24cb9bae44fba11020c182aef	150.3 MB	3



# Preparing Content for SIPs

- ▶ SIPs for each office are based on access restrictions (Unrestricted, Onsite, Restricted)
- ▶  **Aid4Mail e-Discovery Archivist:** converting email from multiple formats (eml, mbx, msg, pst, sbd, Pegasus mail) to MBOX



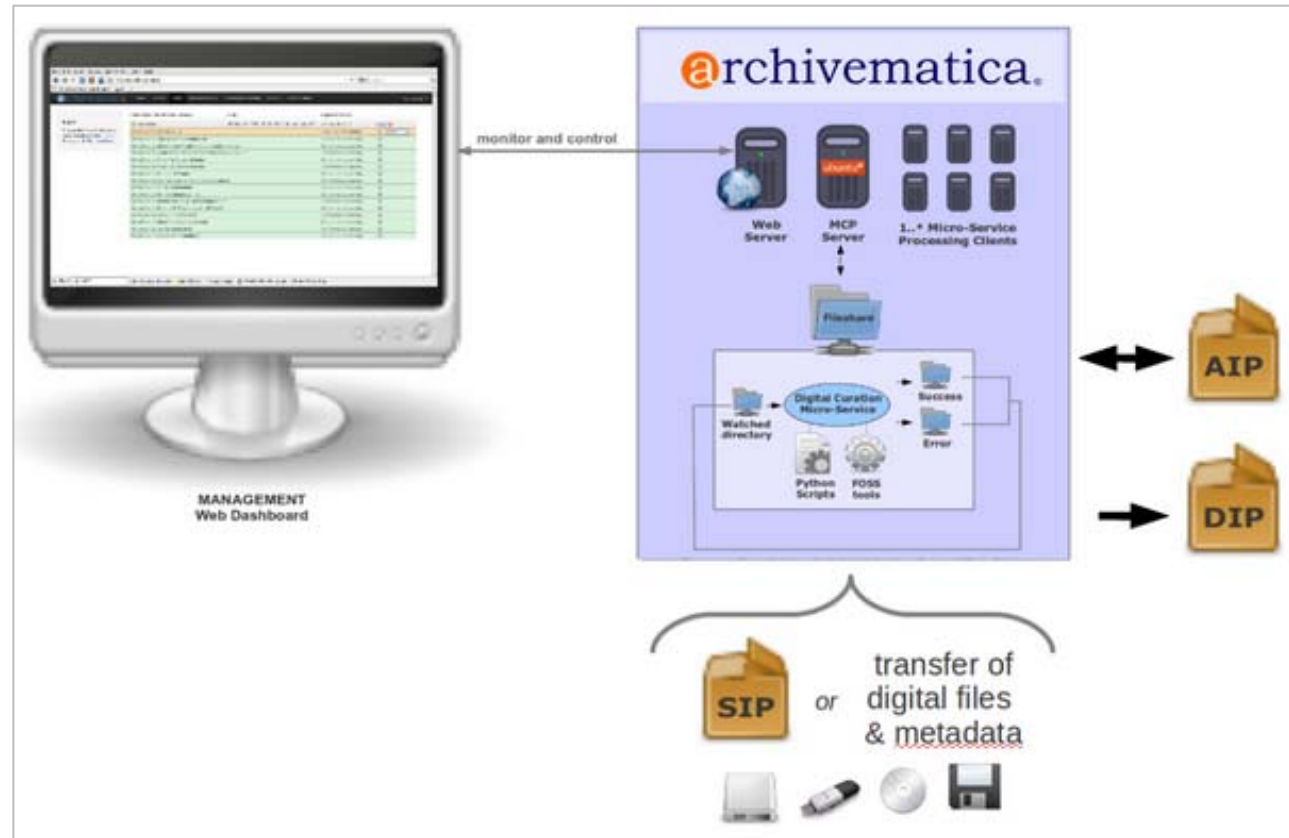
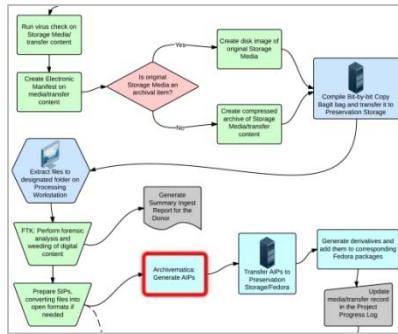
- ▶  **Database Preservation Toolkit:** converting Microsoft Access databases to XML format
- ▶ **Original formats:** SQL-based databases, statistical datasets (sav, spss)
- ▶ **External Vendor:** converting content of commercially produced video DVDs, audio CDs, and mini DV-tapes to preservation formats



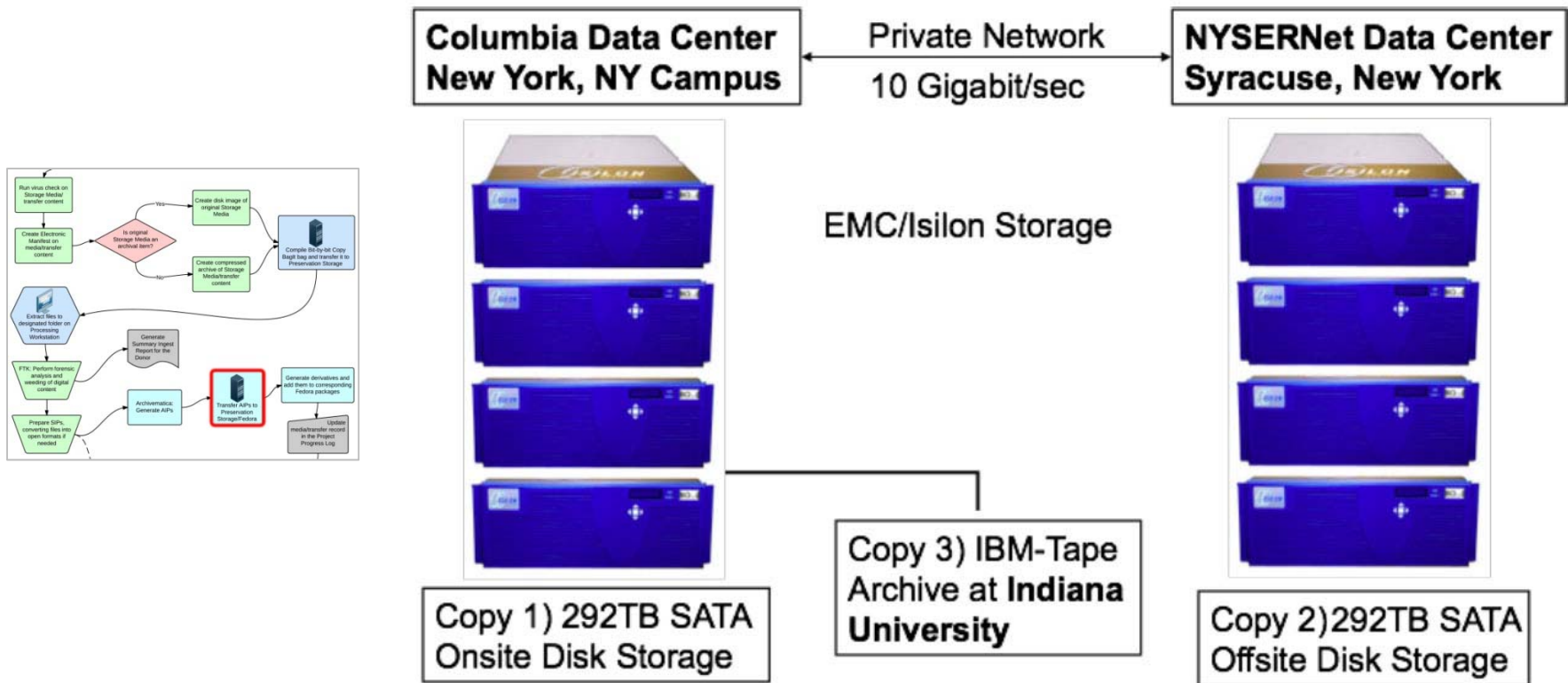
# Generating AIPs



Archivematica: digital preservation

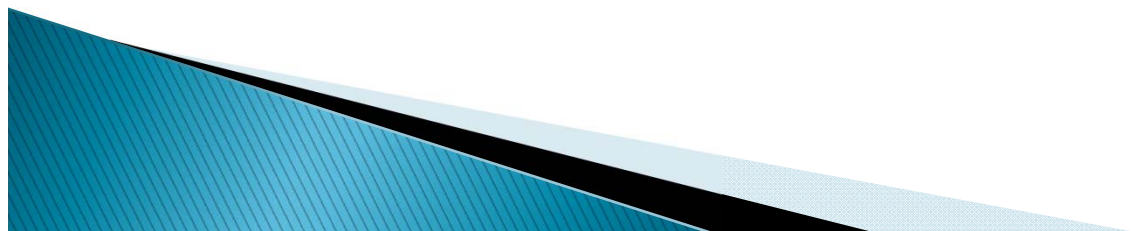


# Preservation Storage



# Metadata

- ▶ Descriptive metadata: makeInventory program, Archivematica
- ▶ Technical metadata: makeInventory program, FTK, Archivematica
- ▶ Preservation metadata: Archivematica
- ▶ Rights metadata: Archivematica



# Relevance to National Agenda

- ▶ **Digital Content Areas:**
  - Electronic records
  - Research data
  - Websites
  - Audiovisual materials
- ▶ **Organizational Roles, Policies, and Practices:**
  - Education of content creators
  - Large amount of complex data
  - Rights management
  - Security and compliance policies
- ▶ **Technical Infrastructure Development:**
  - Integration of digital forensics tools
  - Ensuring content integrity
  - File format action plan development





# Thank you!

Questions? Contact us:  
[ds2057@columbia.edu](mailto:ds2057@columbia.edu)  
[jg2138@columbia.edu](mailto:jg2138@columbia.edu)

