Infrastructure Development:

Multiple Digital Content Types in a Single Collection

Dina Sokolova and Jane Gorjevsky,
Columbia University

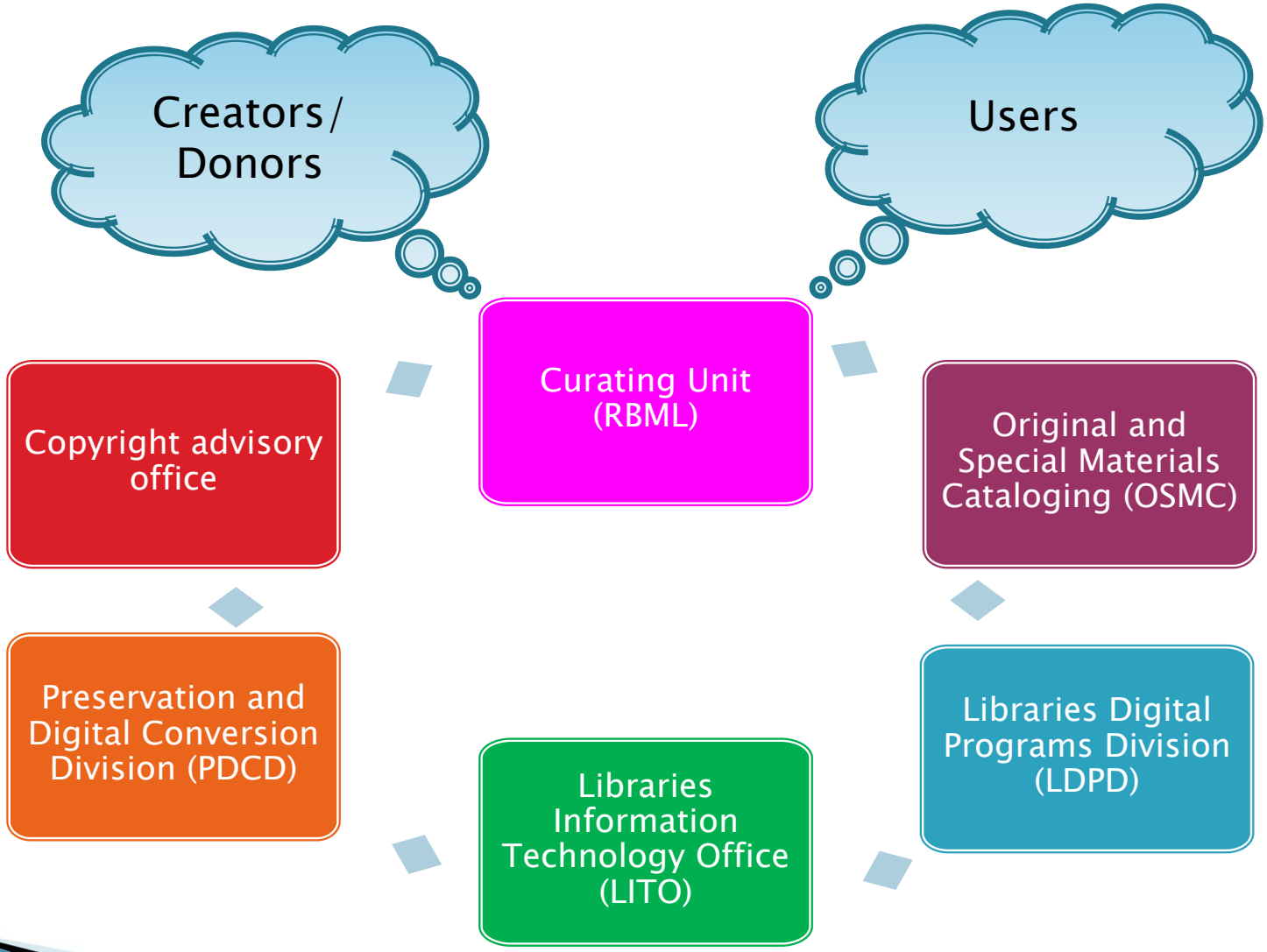
Digital Content in Columbia Special Collections

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

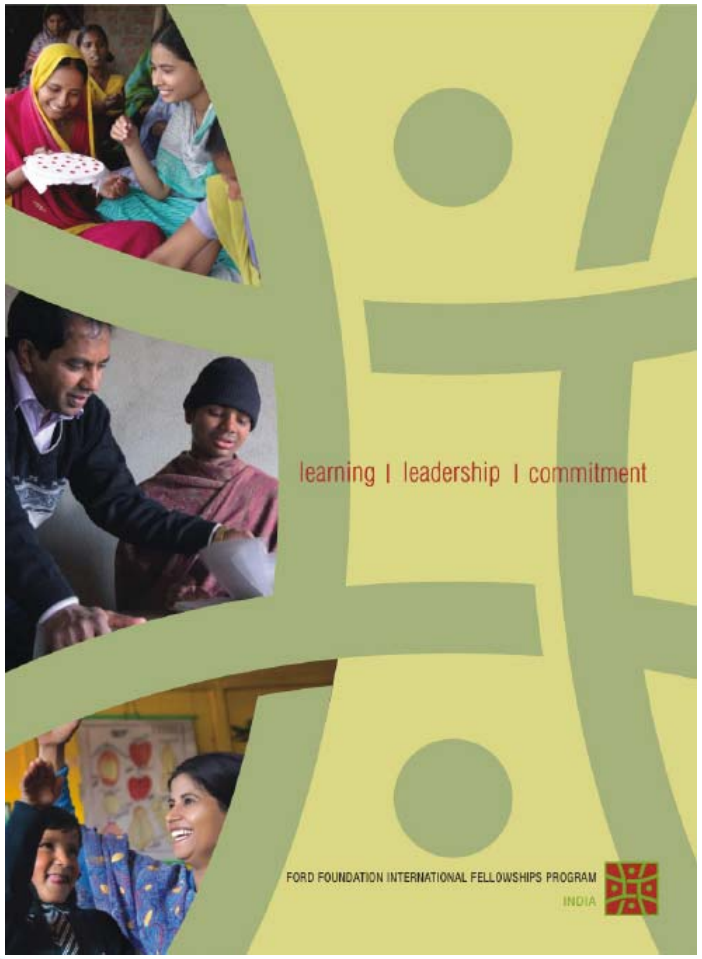
Digital Content Pilot Project

- 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

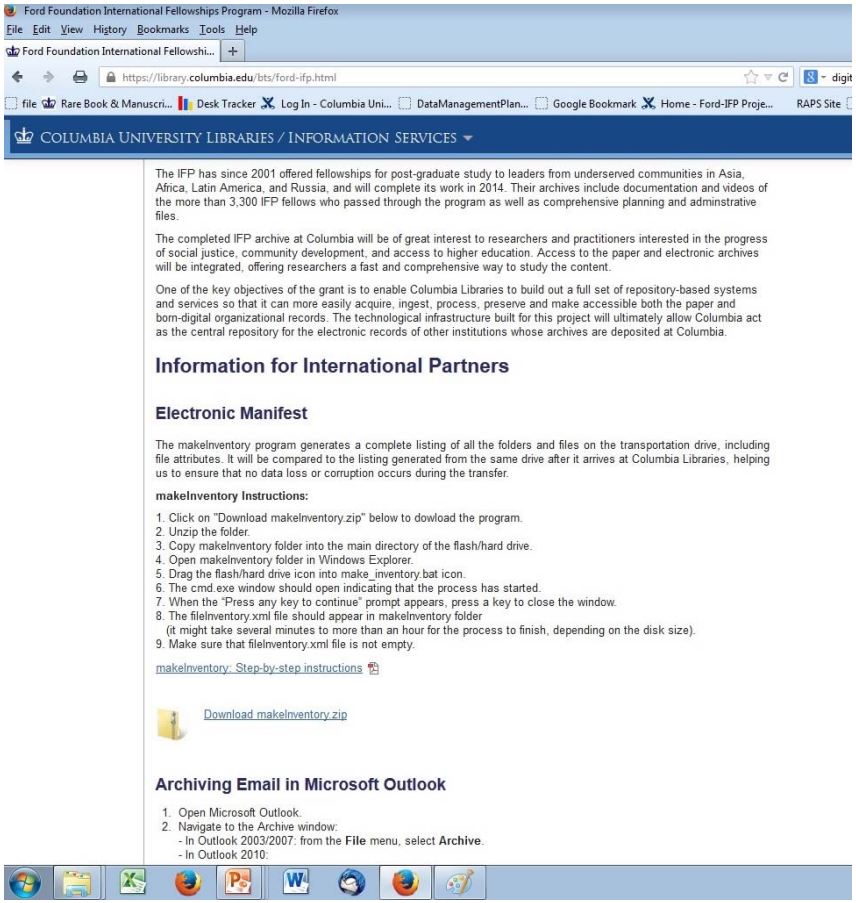
- Offered fellowships for post-graduate study to more than 4,300 people via offices in 22 countries with an overall program management by Secretariat in New York 2001 – 2013

Records Scope and Content

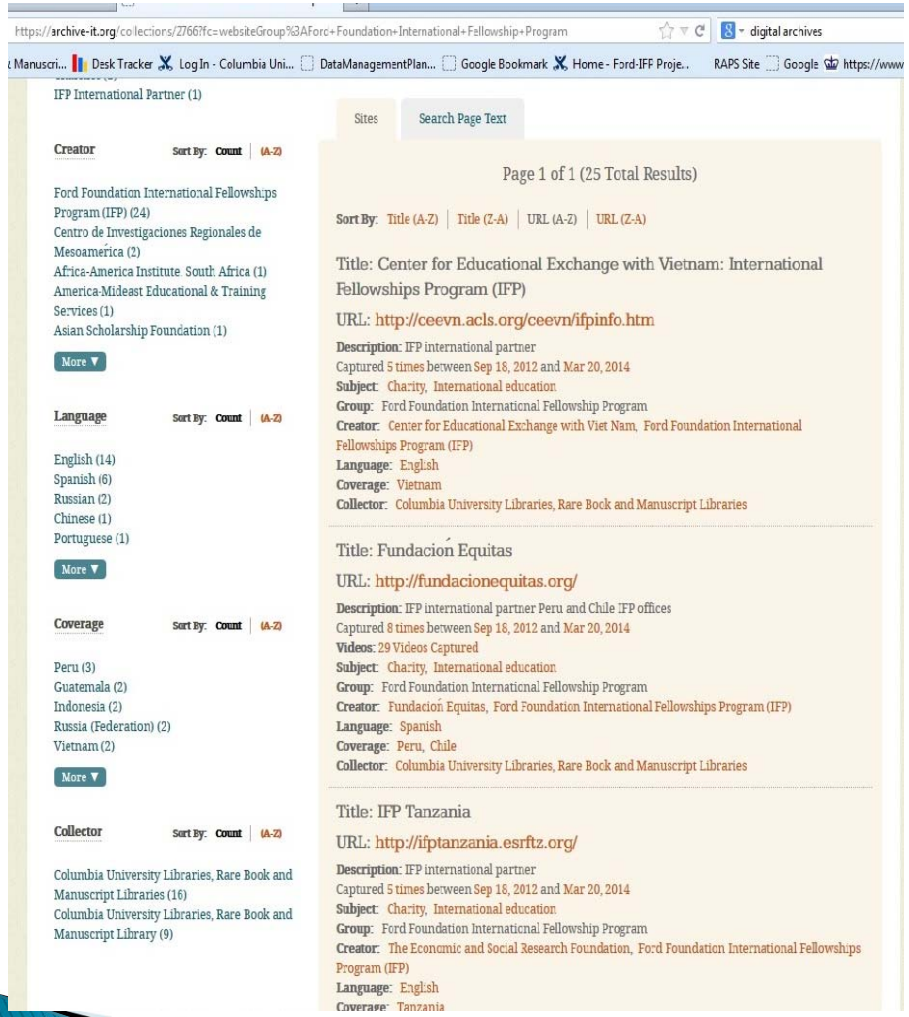
- Paper and digital records from 22 International partner organizations, New York Secretariat and CHEPS (Center for Higher Education Policy Studies)
- Materials include:
  - Office documents
  - Time-based (audio and video) materials
  - Databases
  - Email correspondence
  - Websites
  - Academic and 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



Archiving Web Resources



- 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:

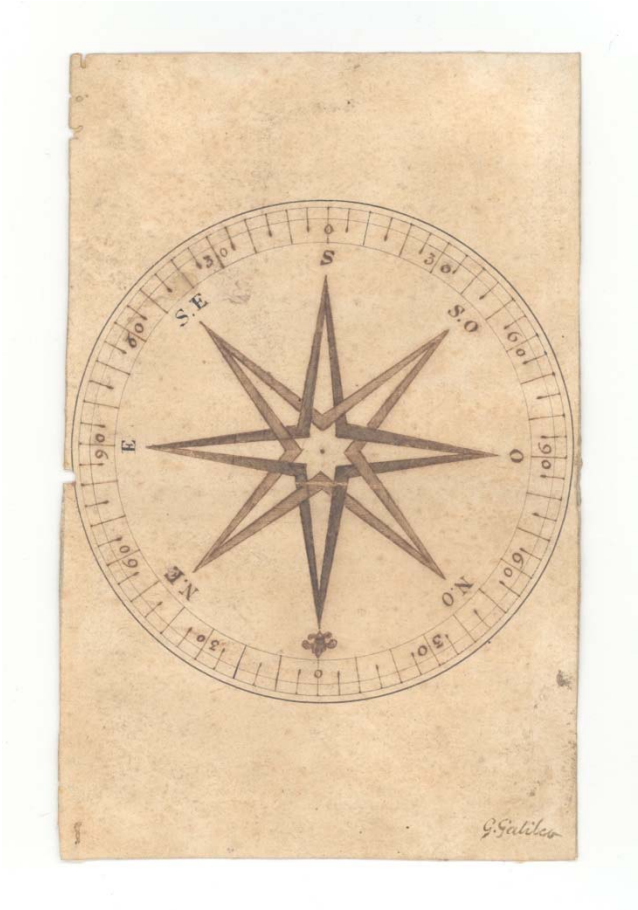
  IFP\Newsletter\Alumni Meeting\... \...\...\Fifth meeting October 23-28, 2008\Agenda\IFP Assembly\Other\07.jpg

- Foreign languages:

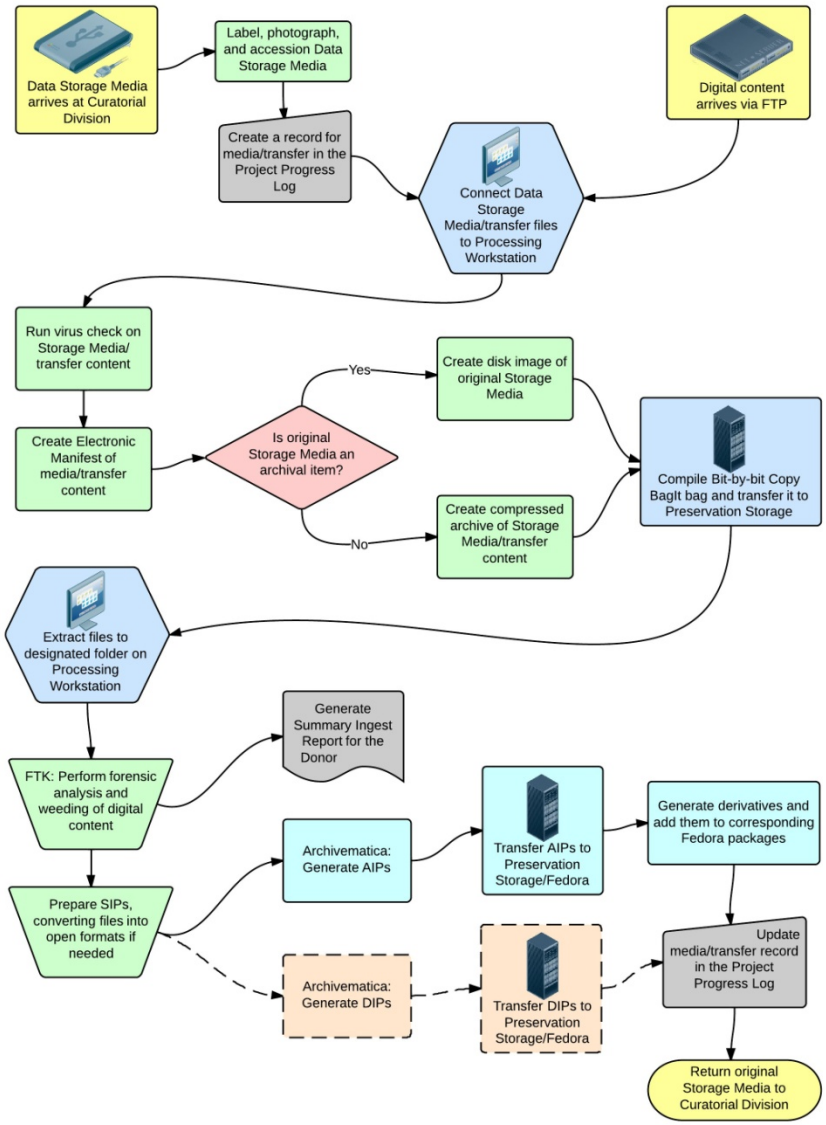
  IFP\...\...\Foto bersama usai sidang kongres Perhimpunan Pelajar Indonesia Australia di Balai Kartini Gedung KBRI Canberra, 2012.jpg (A group photograph of Indonesian students taken after the congress in front of the Indonesian Embassy in Canberra, Australia, 2012)

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

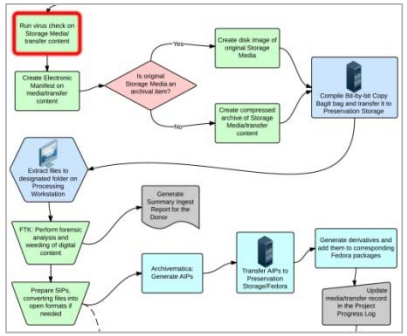


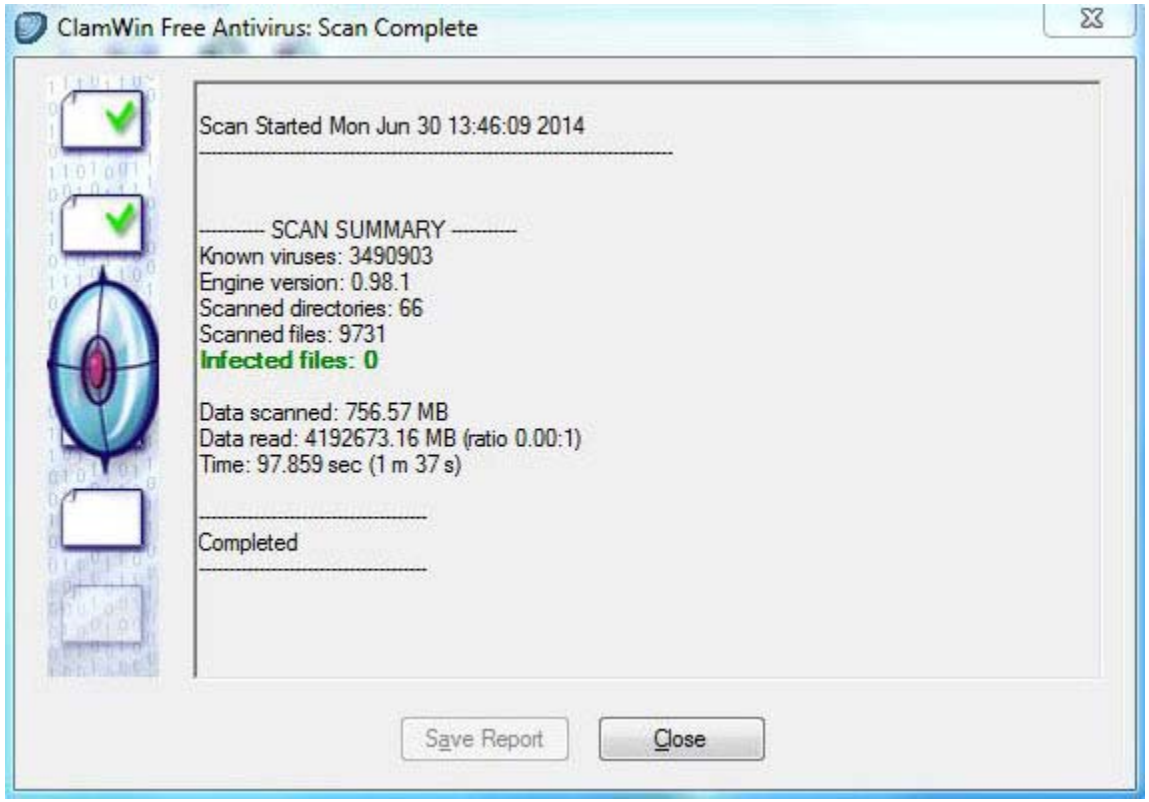
Processing Workstation

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

Virus Check

ClamWin (ClamXav): initial virus check

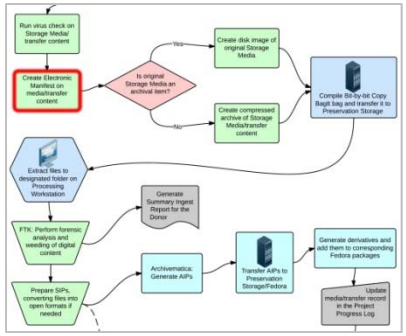




Electronic Manifest

Hashdeep

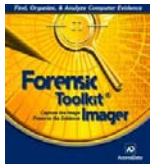
makeInventory program: verifying content integrity



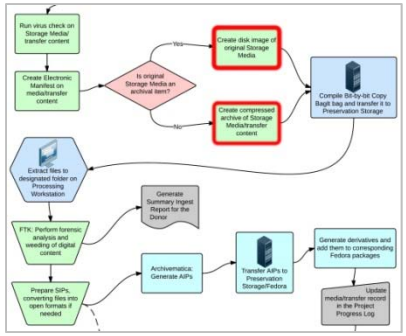
D:\fileInventory.xml


```
<?xml version="1.0" encoding="UTF-8"?>
- <dfxml xmloutputversion="1.0">
  - <metadata 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">5ffaff9bffec500ac3257f30cd27f6b0e5e46ebcac655d5499d3801077f2d560</hashdigest>
    </fileobject>
  - <fileobject workerid="3">
      <filename>G:\RESTRICTED\IFP Program Files\Program Orientation, Needs Assessment&Educational Advising\финалисты 4 инфо.doc</filename>
```

Disk Imaging

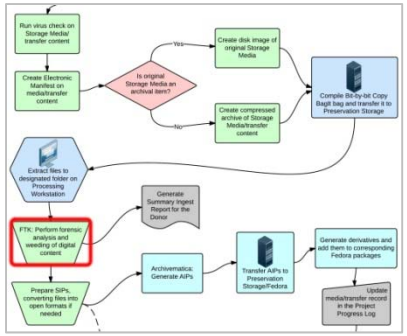


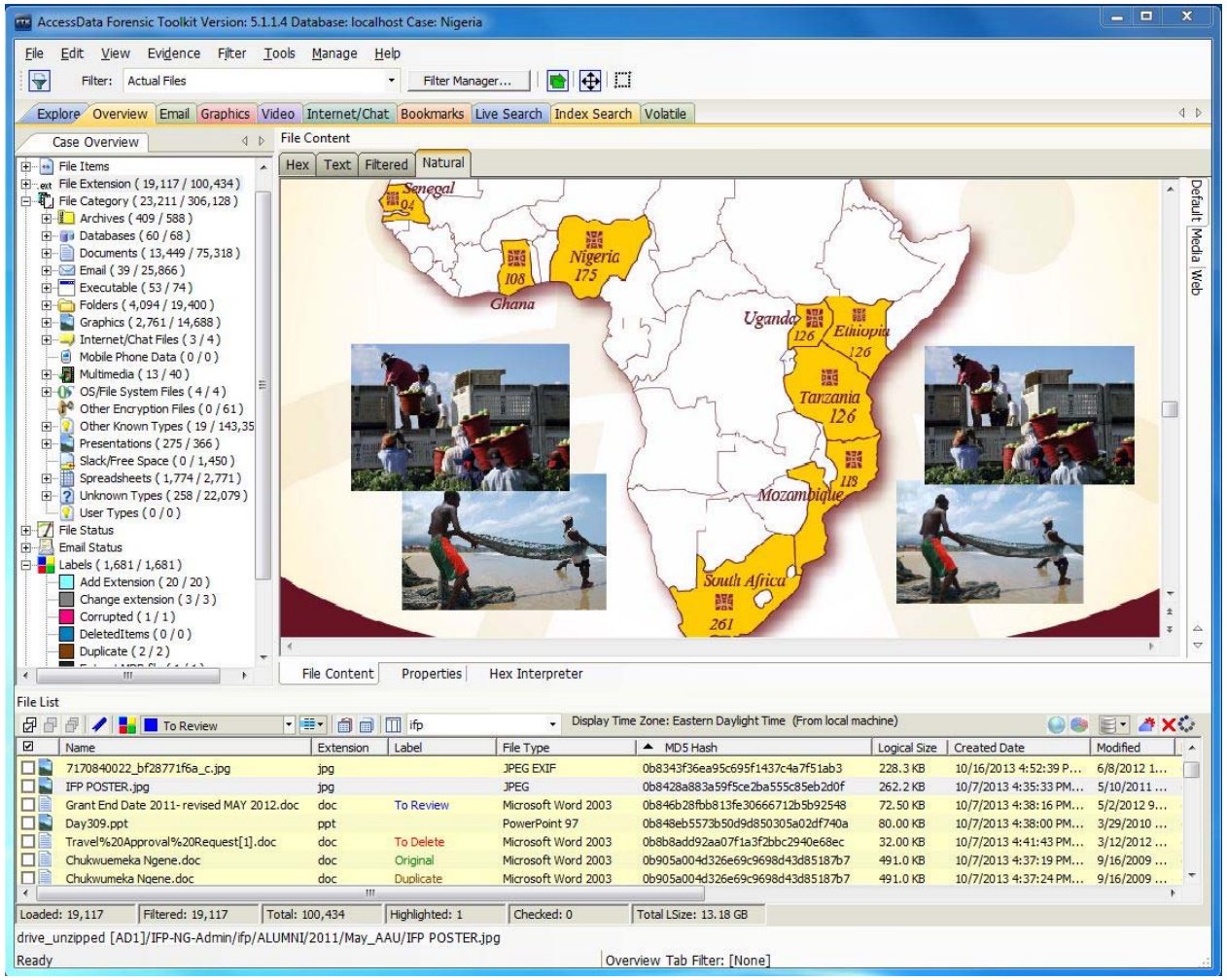
AccessData FTK Imager: creating disk images (CDs/DVDs, ZIP and Floppy Disks)



Appraisal and Selection

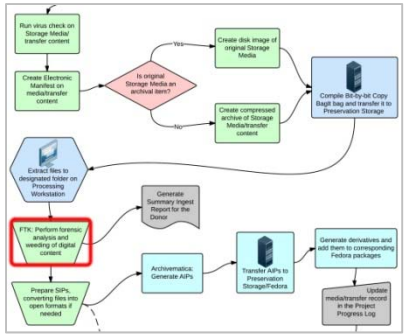
Forensic Toolkit (FTK): content review and weeding

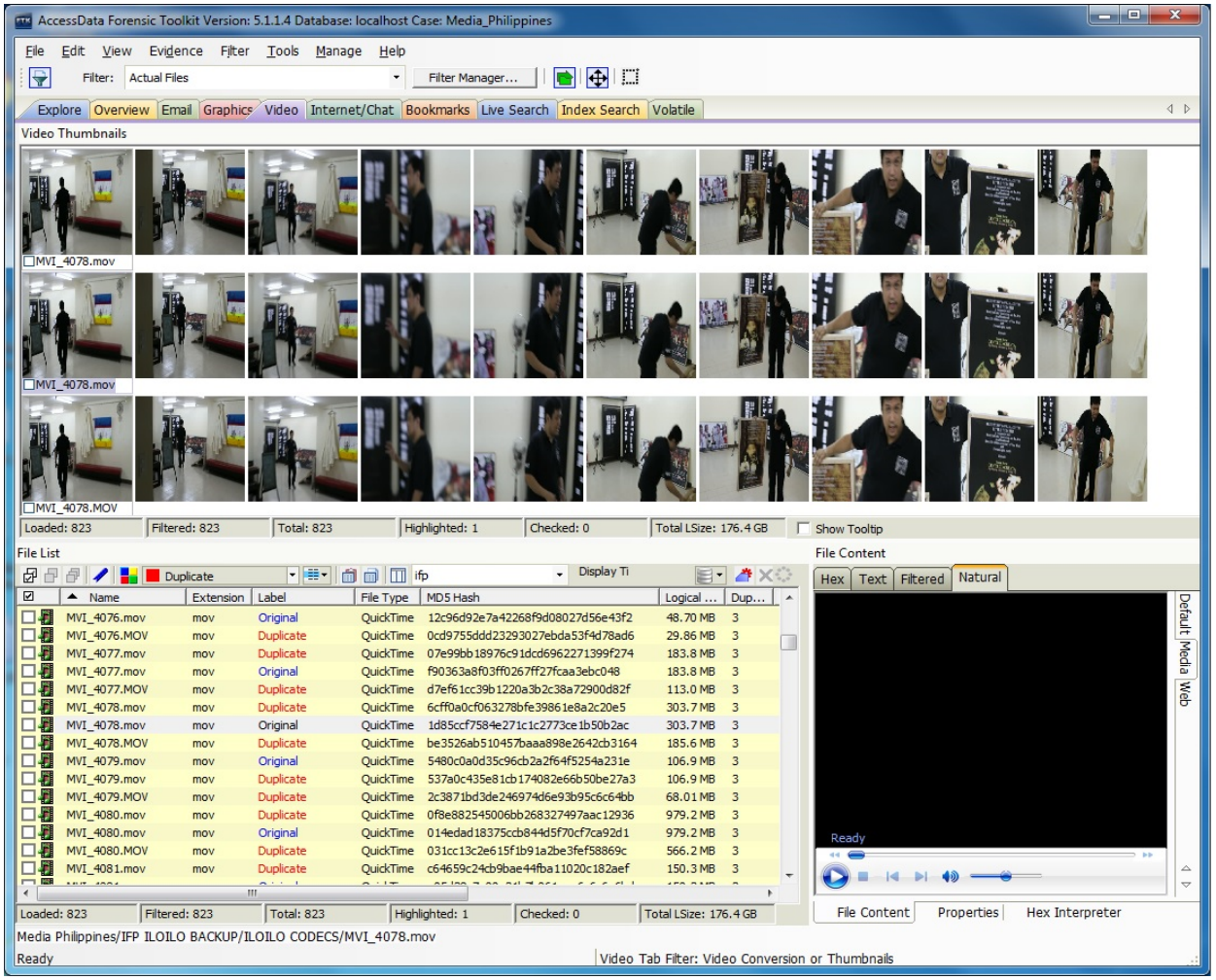




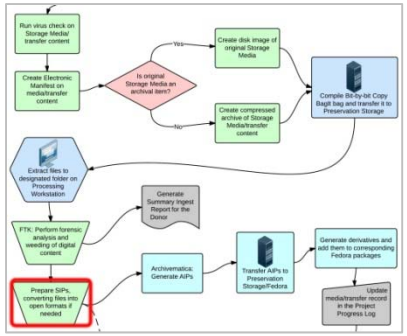
Appraisal and Selection (cont.)

Forensic Toolkit (FTK): finding duplicate content





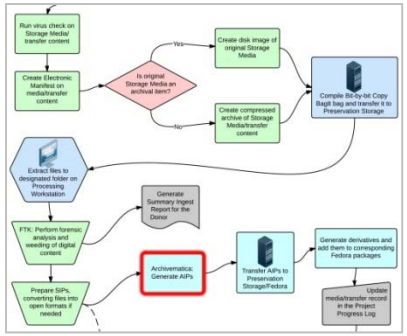
Preparing Content for SIPs

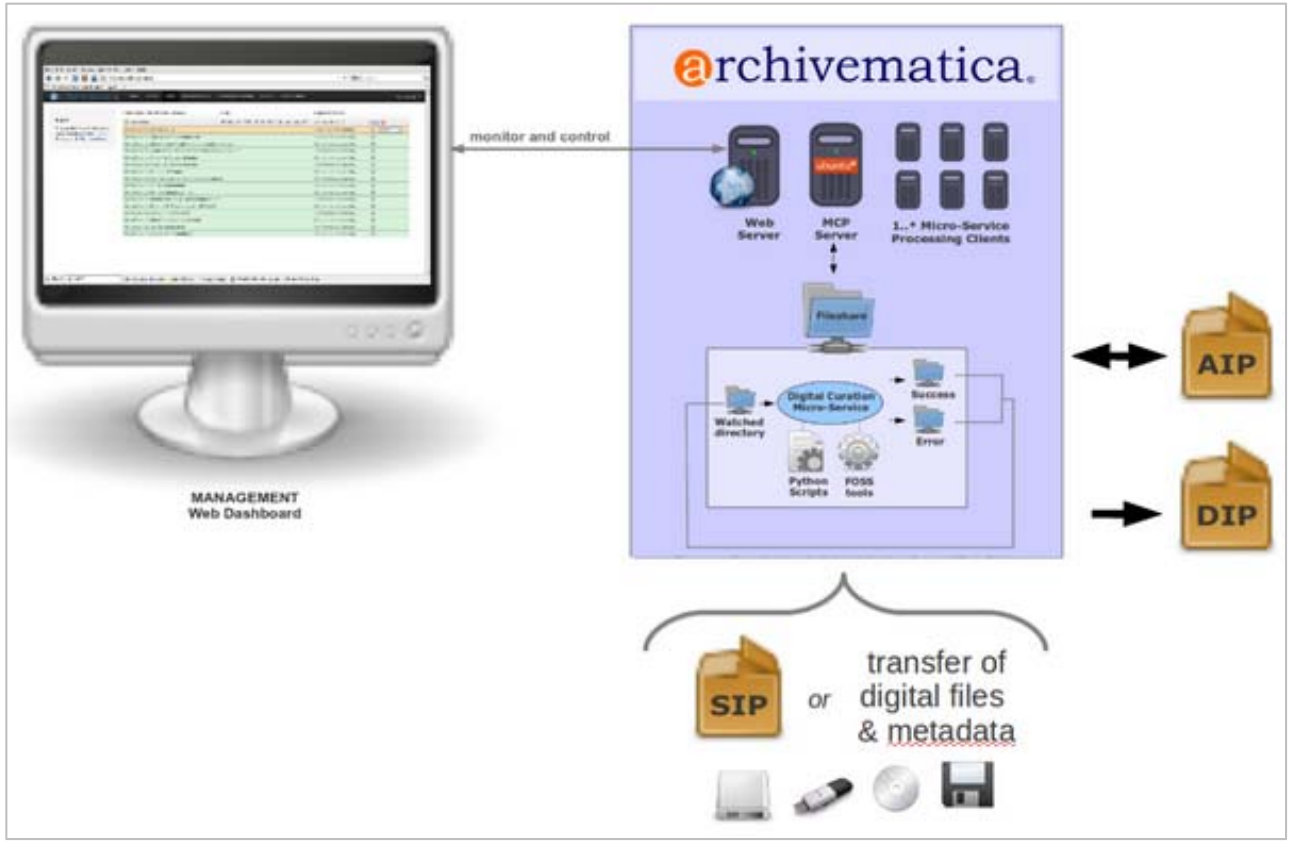


- SIPs for each office are based on access restrictions (Unrestricted, Onsite, Restricted)
- Aid4Mail e-Discovery Archivist: converting email from multiple formats (eml, mbx, msg, pst, sbd, Pegasus mail) to MBOX
- Db-preservation-toolkit
  Database Preservation Toolkit
  Database Preservation Toolkit: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

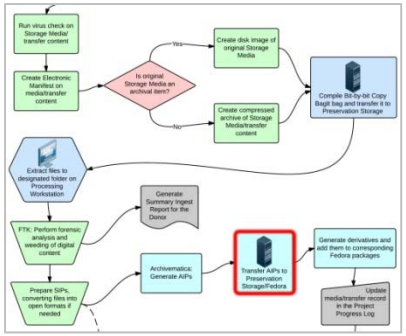
Generating AIPs

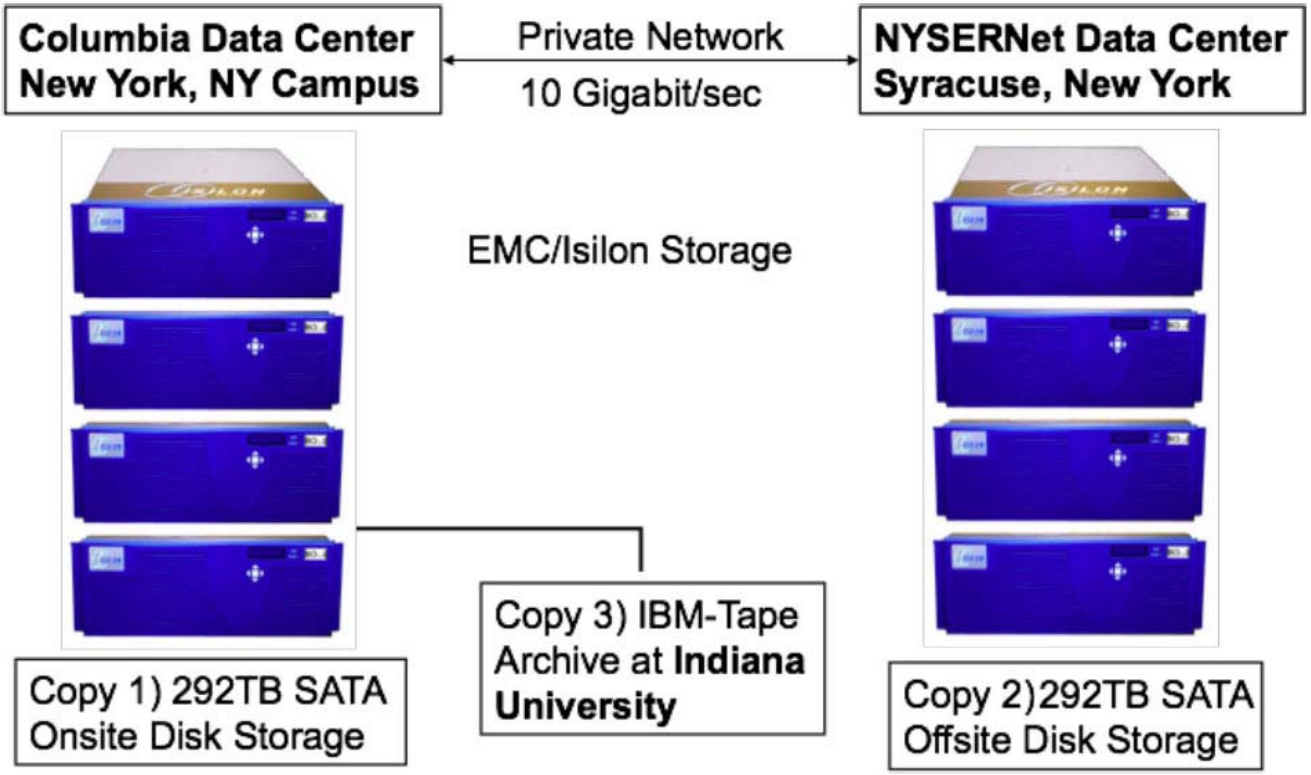
Archivematica: digital preservation





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
jg2138@columbia.edu