

# Adding Metadata and Ingesting Large Born-Digital Archives with Archivematica

Dina Sokolova and Jane Gorjevsky  
*Columbia University*

# Archives of the Ford Foundation International Fellowships Program

- Large-scale project funded by the Ford Foundation grant
- Key goals:
  - Permanently preserve IFP paper and electronic records
  - Provide access to IFP digital archives based on three types of user access:
    - ◆ publicly accessible
    - ◆ viewable onsite only
    - ◆ embargoed until 2075

# International Fellowships Program Overview

The screenshot shows the homepage of the Ford Foundation International Fellowships Program. At the top left is the logo, a stylized globe with dots, and the text "FORD FOUNDATION INTERNATIONAL FELLOWSHIPS PROGRAM". To the right are navigation links: "LOGIN >>", "SELECT LANGUAGE" with a dropdown arrow, a "Google Custom Search" box, and "Choose Country" with a dropdown arrow. Below these are social media icons for Facebook, Twitter, YouTube, and LinkedIn, and a "USER GUIDES >>" link. A banner below the navigation reads "A Decade of Advanced Study Opportunities for Social Change Leaders Worldwide". On the left is a blue sidebar menu with links: Home, About IFP, IFP Legacy Online, News & Publications, Multimedia, Thesis Library, Alumni Opportunities, Alumni Search, Alumni Login, and Contact Us. The main content area features a large photo of women in traditional attire with the text "Linking Higher Education To Social Change". Below the photo are three sections: "Community Forums" (Share resources, pose questions, and dialogue with other IFP alumni), "IFP Alumni Tracking Study" (As IFP closes, new doors open. Learn about the 10-year IFP alumni tracking study and how you can participate.), and "Manage Your Profile" (Tell us about yourself! Share personal and professional interests, add a photo, send messages, and interact with other features of the site.). Each section has a "GO >>" link.

- Program was active in 2001 – 2013
- Program offered fellowships for post-graduate study to social justice leaders from underserved communities in Asia, Africa, Latin America, Russia, and the Middle East

## Scope of Materials

- 3.6 TB of electronic materials, received from 22 International partner organizations, New York Secretariat and CHEPS (Center for Higher Education Policy Studies):
  - Planning and administrative documents
  - Audiovisual materials
  - Databases
  - Email correspondence
  - Website content
  - Academic and personal records of fellows
  - Surveys, interviews and statistical reports
  - Datasets

# Challenges

- About 350,000 files in 245 formats, 10 languages, 7 non-roman character sets
- Filenames and directory paths as the only source of descriptive metadata
- Long filenames/file paths (> 260 characters)
- Multiple languages and non-Roman character sets:

Original:

Ð“Ð¾¼Ñ€Ð±Ð°Ñ‡Ð¼Ð²-Ð¸µ Ñ…Ð¾¼Ñ‡ÑƒÑŒÑ’Ð°Ð²Ð°Ñ,ÑŒÑŒÑŒ.doc

Normalized:

\_\_\_\_\_ - \_\_\_\_\_ .doc

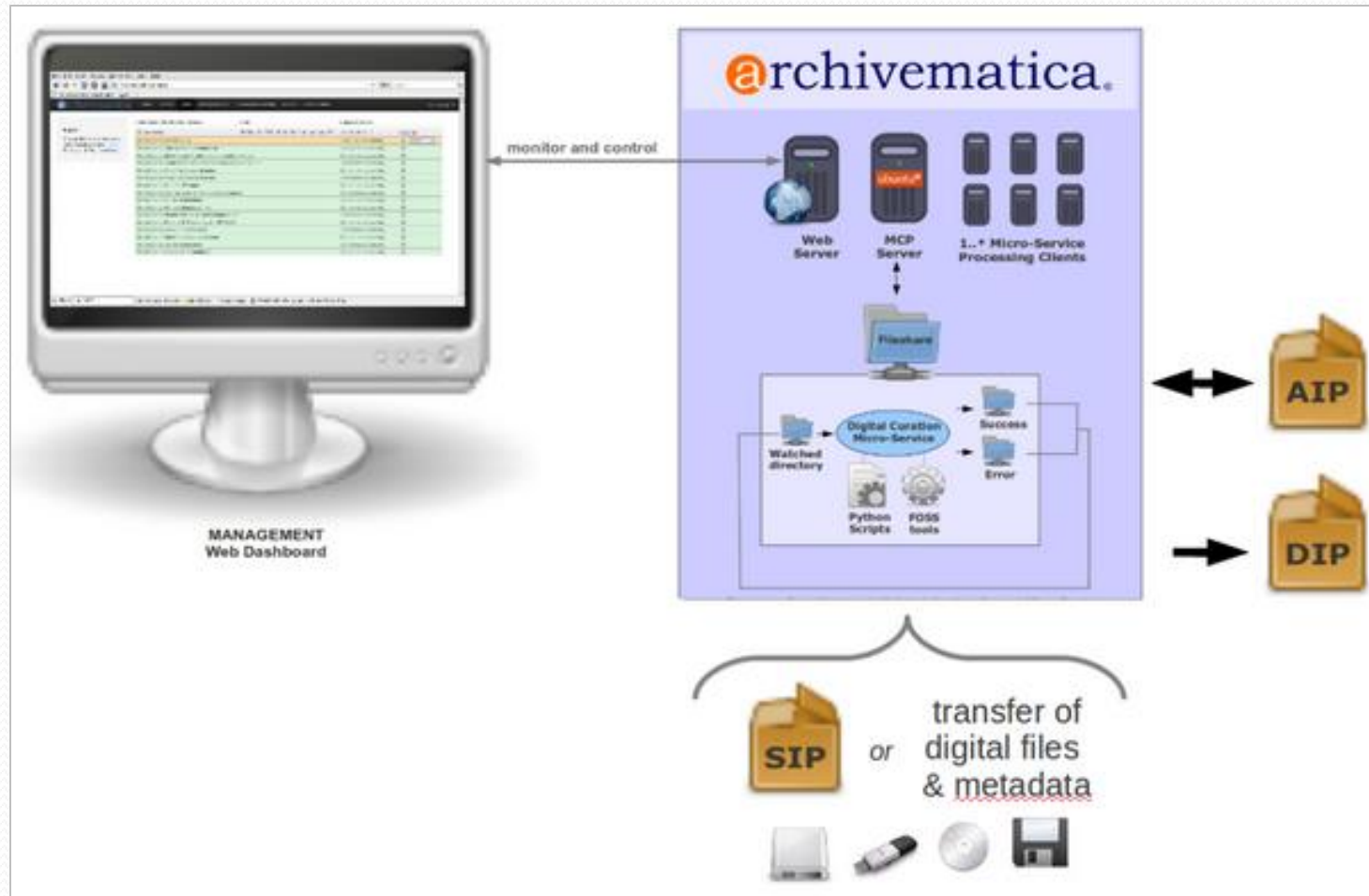
- Appraisal and Selection
- Privacy and confidentiality concerns

# Preparing Content for SIPs

- Submission Information Packages (SIPs) for each office are based on access restrictions (Unrestricted, Onsite, Restricted)
- Content preparation:
  - Converting email from multiple formats (eml, mbx, msg, pst, sbd, Pegasus mail) to MBOX
  - Converting Microsoft Access databases to XML format
  - Outsourcing conversion of content of commercially produced video DVDs, audio CDs, and mini DV-tapes to preservation formats
  - Extracting data from ZIP and RAR archives
  - Establishing SIP size

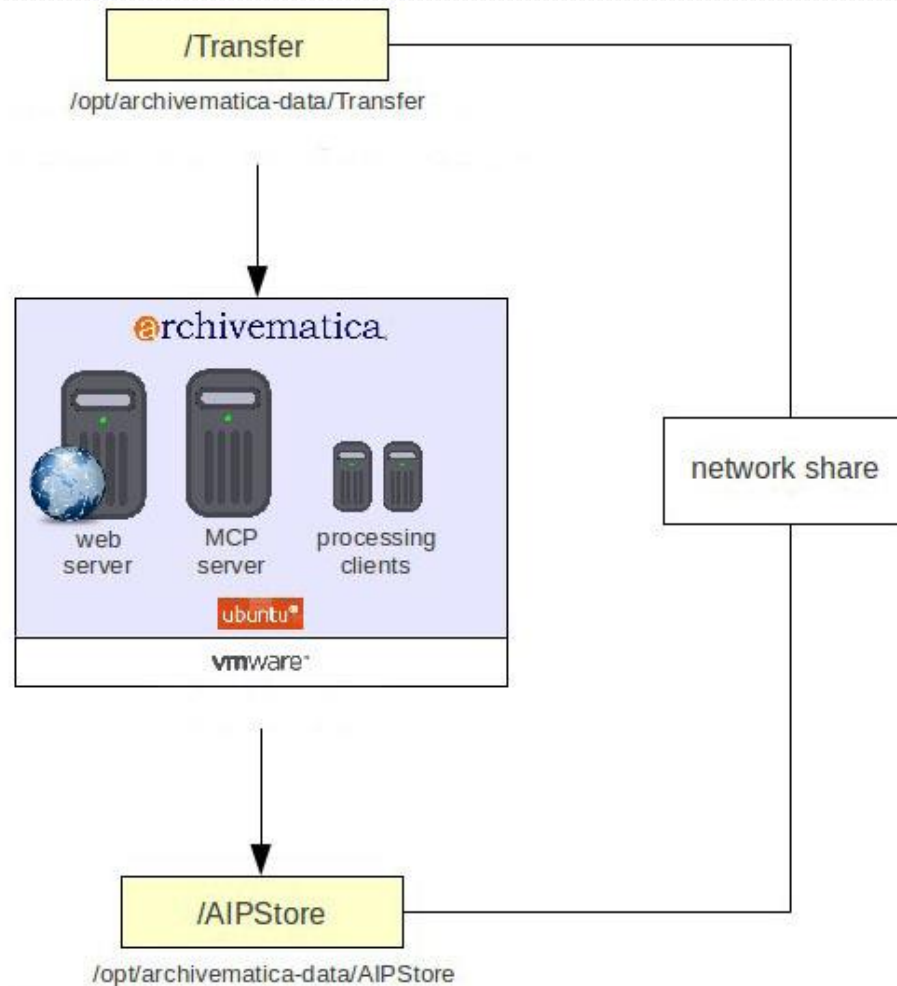
# Archivematica

- OAIS-compliant digital preservation system



# Archivematica at CUL

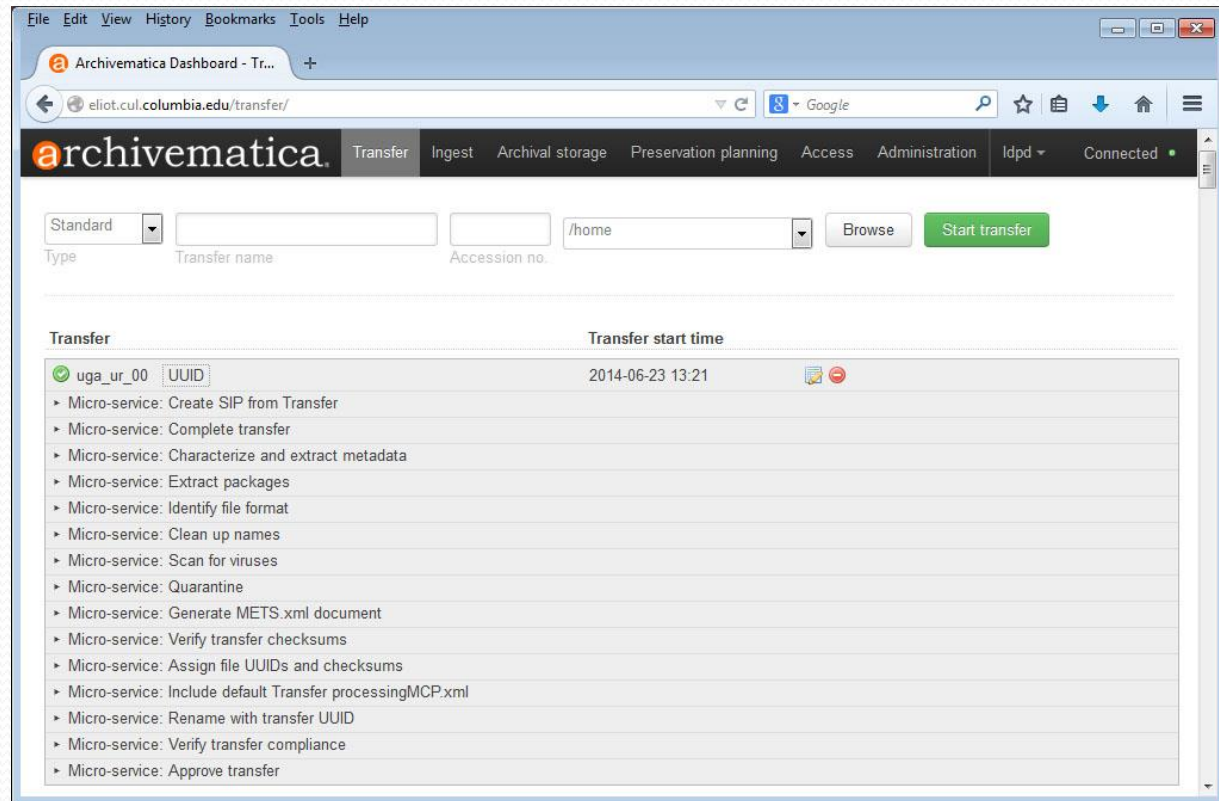
- Dedicated Ubuntu virtual machine on CUL server with mounted network storage





# Submission Information Packages

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



The screenshot displays the Archivematica web interface. At the top, there is a navigation menu with options: File, Edit, View, History, Bookmarks, Tools, and Help. Below this is a browser window showing the URL 'eliot.cul.columbia.edu/transfer/'. The main header features the 'archivematica' logo and a navigation bar with links: Transfer, Ingest, Archival storage, Preservation planning, Access, Administration, Idpd, and Connected.

The main content area contains a form for configuring a transfer. It includes a dropdown menu for 'Type' (set to 'Standard'), input fields for 'Transfer name' and 'Accession no.', a dropdown for the destination path (set to '/home'), a 'Browse' button, and a green 'Start transfer' button.

Below the form is a table titled 'Transfer' with columns for 'Transfer' and 'Transfer start time'. The table shows a single transfer entry with a green checkmark, the name 'uga\_ur\_00', and a UUID. The start time is '2014-06-23 13:21'. A list of micro-services is shown below the entry, including: Create SIP from Transfer, Complete transfer, Characterize and extract metadata, Extract packages, Identify file format, Clean up names, Scan for viruses, Quarantine, Generate METS.xml document, Verify transfer checksums, Assign file UUIDs and checksums, Include default Transfer processingMCP.xml, Rename with transfer UUID, Verify transfer compliance, and Approve transfer.

# Rights Metadata

- PREMIS rights at the SIP level

Archivematica Dashboard - eliot.cul.columbia.edu/ingest/6ec9ce8b-b967-4b1a-82bc-2e9eacd3843f/rights/25

archivematica Transfer Ingest Archival storage Preservation planning Access Administration Idpd

Ingest / uga\_ur\_00 / Rights / Edit

### Rights

uga\_ur\_00

**Basis**  
Donor

**Donor agreement start date**  
2012/07/25  
Use ISO 8061 (YYYY-MM-DD)

**Donor agreement end date**  
2075/01/01  
Use ISO 8061 (YYYY-MM-DD)

Open End Date

**Donor documentation identifier:**

**Type**  
Donor Transfer Agreement and Deed of Gift

**Value**  
5268074v.2

**Role**

**Donor agreement note**  
Restrictions on access by donor request

Save Next Cancel

Archivematica Dashboard - eliot.cul.columbia.edu/ingest/6ec9ce8b-b967-4b1a-82bc-2e9eacd3843f/rights/grar

archivematica Transfer Ingest Archival storage Preservation planning Access Administration Idpd

Ingest / uga\_ur\_00 / Rights / Edit

### Rights

uga\_ur\_00

**Act**  
Access

**Grant/restriction**  
Allow

**Start**  
2012/07/25  
Use ISO 8061 (YYYY-MM-DD)

**End**  
Use ISO 8061 (YYYY-MM-DD)

Open End Date

**Grant/restriction note**  
No restrictions

**Act**  
Disseminate

**Grant/restriction**  
Conditional

**Start**  
2012/07/25  
Use ISO 8061 (YYYY-MM-DD)

**End**  
Use ISO 8061 (YYYY-MM-DD)

Open End Date

**Grant/restriction note**  
Permission to publish material from the collection must be requested from the Rare Book and Manuscript Library (RBML). The RBML approves permission to publish that which it physically owns; the responsibility to secure copyright permission rests with the patron.

# Descriptive Metadata

- Dublin Core metadata at the SIP level

The screenshot shows the Archivematica Dashboard interface. The browser address bar displays the URL: `eliot.cul.columbia.edu/ingest/6ec9ce8b-b967-4b1a-82bc-2e9eacd3843f/metadata/25/`. The page title is "Archivematica Dashboard". The main navigation menu includes: Transfer, Ingest, Archival storage, Preservation planning, Access, Administration, and Idpd. The breadcrumb trail is: Ingest / uga\_ur\_00 / Metadata / Edit.

## Metadata

uga\_ur\_00

**Applies to**  
uga\_ur\_00  
Metadata can be added at the SIP/AIP level only

**Title**  
Ford Foundation International Fellowships Program Records, Series IV: International Partners, Subseries 14: I

**Creator**  
Association of the Advancement of Higher Education and Development

**Subject**

**Description**  
Unrestricted Born-Digital Records of IFP International Partner in Uganda

**Publisher**

**Contributor**  
Uganda

**Date**  
Use ISO 8061 (YYYY-MM-DD or YYYY-MM-DD/YYYY-MM-DD)

**Type**  
Collection

**Format**

**Identifier**  
9489034\_uga\_ur\_00

**Source**

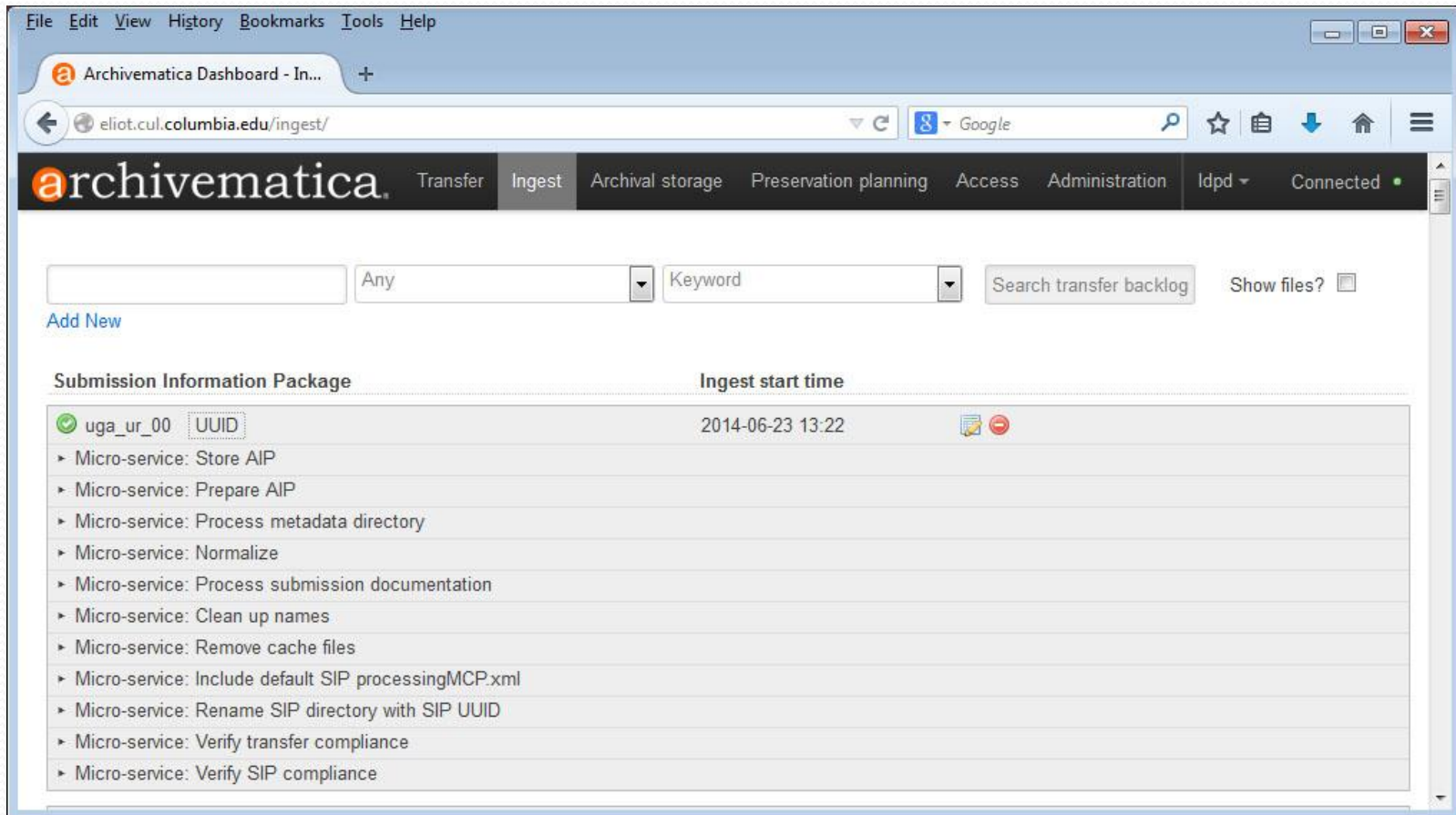
**Relation**

**Language**

A name given to the resource. (ISO15836)

# Archival Information Packages

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



The screenshot shows the Archivematica Dashboard in a web browser. The browser address bar displays "eliot.cul.columbia.edu/ingest/". The dashboard navigation menu includes "Transfer", "Ingest", "Archival storage", "Preservation planning", "Access", "Administration", "Idpd", and "Connected".

Search filters are set to "Any" for the first dropdown and "Keyword" for the second. A "Search transfer backlog" button and a "Show files?" checkbox are also visible.

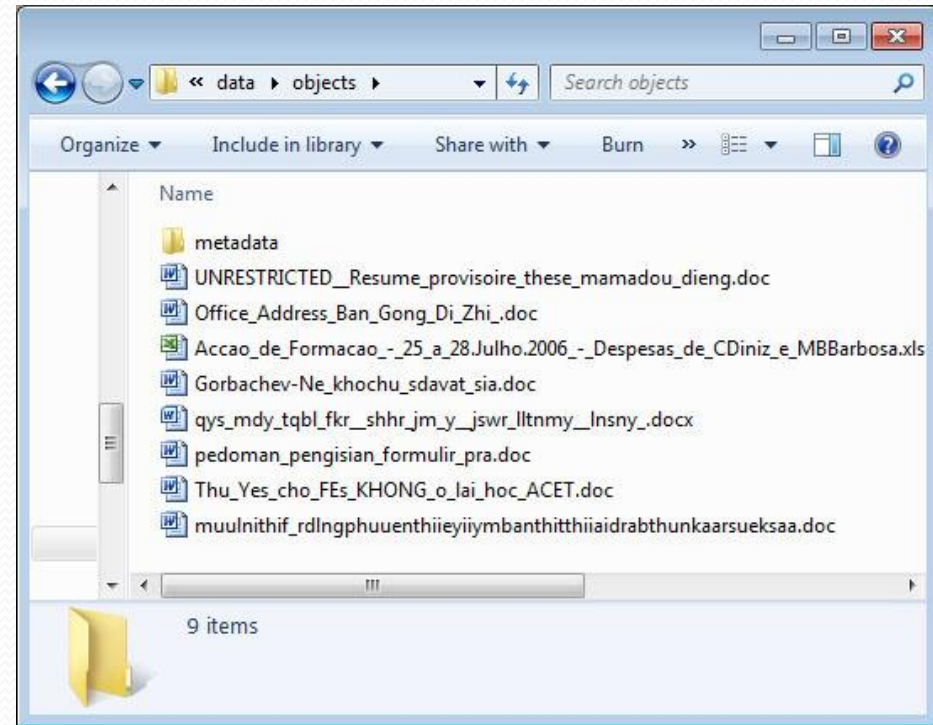
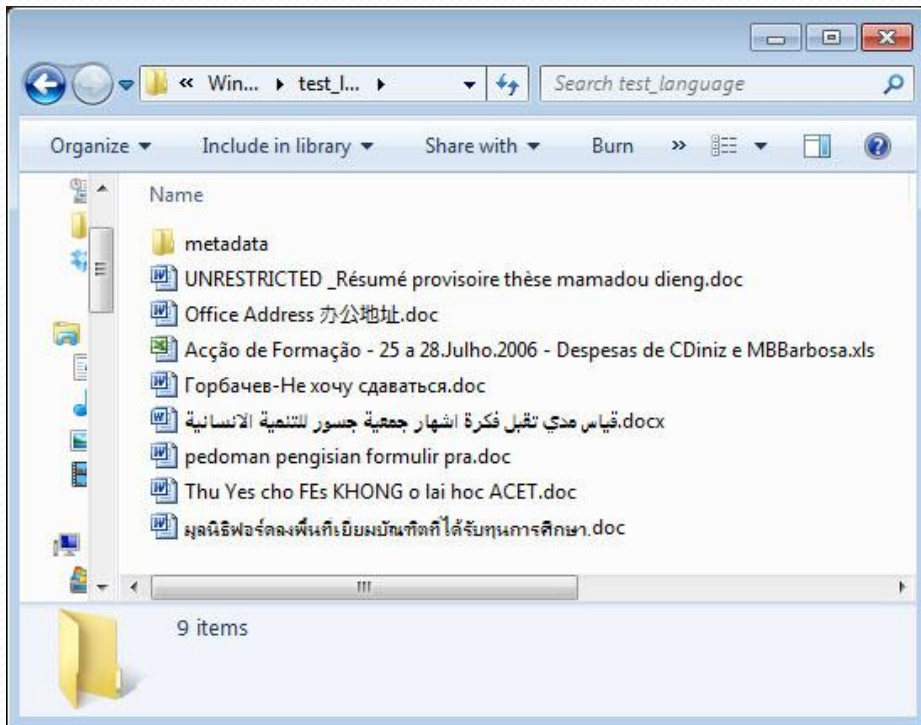
The main content area displays a table with the following data:

| Submission Information Package  | Ingest start time |
|---|-------------------|
| <input checked="" type="checkbox"/> uga_ur_00 <input type="text" value="UUID"/> | 2014-06-23 13:22  |
| ▶ Micro-service: Store AIP  |                   |
| ▶ Micro-service: Prepare AIP  |                   |
| ▶ Micro-service: Process metadata directory                                     |                   |
| ▶ Micro-service: Normalize  |                   |
| ▶ Micro-service: Process submission documentation                               |                   |
| ▶ Micro-service: Clean up names   |                   |
| ▶ Micro-service: Remove cache files   |                   |
| ▶ Micro-service: Include default SIP processingMCP.xml                          |                   |
| ▶ Micro-service: Rename SIP directory with SIP UUID                             |                   |
| ▶ Micro-service: Verify transfer compliance                                     |                   |
| ▶ Micro-service: Verify SIP compliance  |                   |

# Filename Normalization

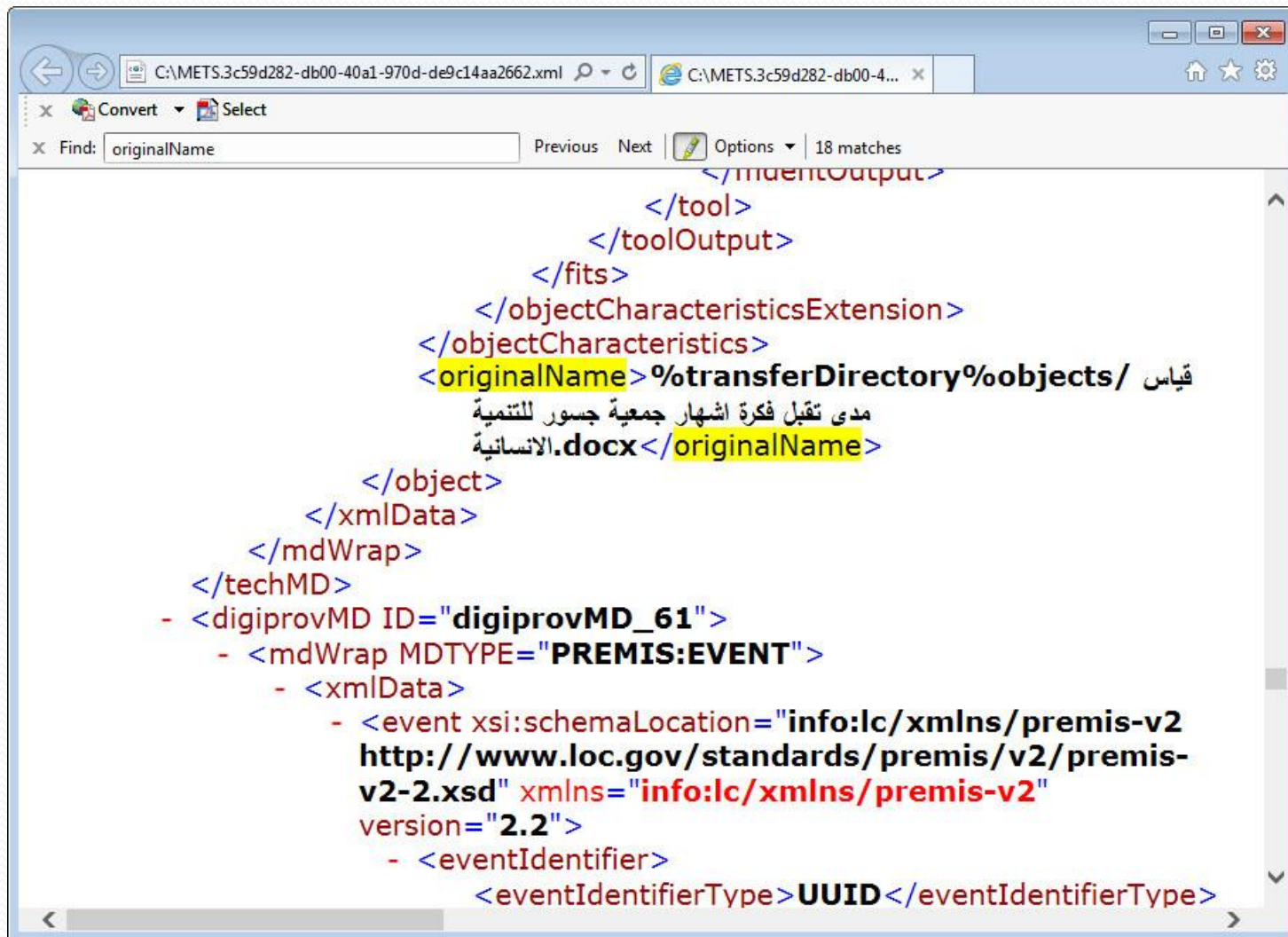
Original

Normalized



# Descriptive Metadata in METS

- Original filenames are retained in METS file

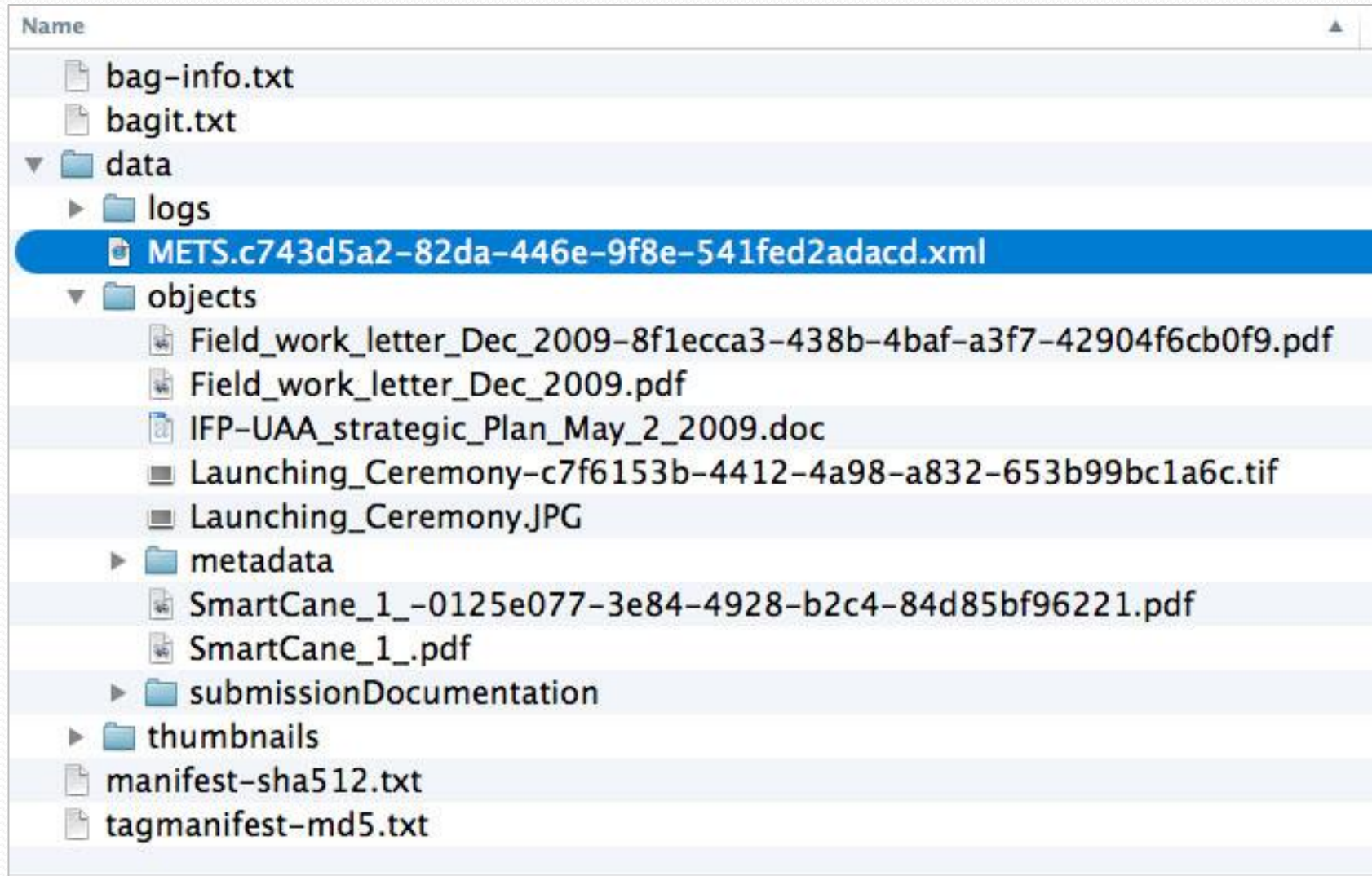


The screenshot shows a text editor window with the following content:

```
C:\METS.3c59d282-db00-40a1-970d-de9c14aa2662.xml
C:\METS.3c59d282-db00-4...
Convert Select
Find: originalName Previous Next Options 18 matches
</indentOutput>
  </tool>
</toolOutput>
</fits>
</objectCharacteristicsExtension>
</objectCharacteristics>
<originalName>%transferDirectory%objects/ قياس
مدى تقبل فكرة اشهار جمعية جسر للتنمية
الانسانية.docx</originalName>
</object>
</xmlData>
</mdWrap>
</techMD>
- <digiprovMD ID="digiprovMD_61">
  - <mdWrap MDTYPE="PREMIS:EVENT">
    - <xmlData>
      - <event xsi:schemaLocation="info:lc/xmlns/premis-v2
http://www.loc.gov/standards/premis/v2/premis-
v2-2.xsd" xmlns="info:lc/xmlns/premis-v2"
version="2.2">
        - <eventIdentifier>
          <eventIdentifierType>UUID</eventIdentifierType>
```

# Storing AIPs

- AIPs in Bagit format are ingested into Preservation Repository



# Thank you!

Contact us:

[ds2057@columbia.edu](mailto:ds2057@columbia.edu)

[jg2138@columbia.edu](mailto:jg2138@columbia.edu)