A Moving Story of Data

In Out, In Out, And Shake It All About

A Moving Story of Data
320 contributors

Any kind of archive

Contribute as and when they wish

Collection level and multi-level

Use our cataloguing tool or export from systems
What we set out to achieve

» Migrate descriptions from old system --- normalise
» Workflow for new system --- repeatable processing
» Rigorous treatment of data --- consistency, structure
» Sustainable model
  › Entities: Archives, Repositories, Names
» Take data in automated way & at scale
» Deal effectively with complex hierarchy & revisions
Check for:

- Valid EAD
- Hub requirements
- Content

Revisions --- version control issues

- More than one version of a description
- Which is the master copy?
- Can change title, reference, filename

Solution needs to work for wide variety of descriptions
Knowledge Integration: CIIM
Is Supplier Known

If No, Determine Supplier

If Yes, Read file from "in progress" repository

Global Transformations

Supplier Transformations

Validate

Has Error been Resolved?

Yes

No
What are we looking to achieve?

What are the required data conditions?

What is the rationale for them?

How can they be implemented consistently?
’Hub Required’ Conditions

» References must be unique
» Repository name must not include the archon code
» Level values from a controlled list
» Language must be provided at the top level
» Access conditions must be provided at the top level
» Normalised dates must use the correct ISO format
» URLs must use http://
» [Title below 500 characters]
Well-formed XML? Known XML format? Known supplier?

Valid EAD/EAG? Good Hub EAD?

XML (original) XML (current) XML (processed) JSON

Hub Admin

Cons

ole

Elastic

Search

CIIM
### Upload Data File

**Organisation:**
- choose an organisation

**Select File:**
- Browse...

**Upload**

### Refine Results

**Order by:**
- Date
- Filename
- Title
- ↓ desc

**The number of files to display:**
- 10

**Workflow Status:**
- Errors

### Showing 1 - 10 of 34 Files

<table>
<thead>
<tr>
<th>Select All</th>
<th>Archives Hub to Resolve</th>
<th>Contributor to Resolve</th>
<th>Publish</th>
<th>Repeat All Transforms</th>
<th>Repeat Global Transforms</th>
<th>Reprocess Original</th>
</tr>
</thead>
</table>

- **Select All**
- **Archives Hub to Resolve**
- **Contributor to Resolve**
- **Publish**
- **Repeat All Transforms**
- **Repeat Global Transforms**
- **Reprocess Original**

**EAD**

- **Wynn Hall Mss**
  - Wynn
  - bangor Wynn.xml
  - 6 Errors
  - 08-Jun-2018 09:04
  - View Audits

- **Huddersfield Branch of the British Federation of Women Graduates**
  - WGH
  - huddersfield WGH.xml
  - 4 Errors
  - 18-Jan-2018 15:57
  - View Audits

- **Ted Hughes Network Archive**
  - THN
  - huddersfield THN.xml
  - 1 Errors
  - 18-Jan-2018 15:57
  - View Audits

- **huddersfield tli.xml**
  - 1 Errors
  - 14-May-2018 15:37
  - View Audits

- **St. Andrew's Society Archive**
  - STA
  - huddersfield STA.xml
  - 1 Errors
  - 18-Jun-2018 15:57
  - View Audits

- **Gnoll Estate Collection**
  - RBSW Gn
  - westglos RBSW Gn.xml
  - 1 Errors
  - 11-Sep-2017 12:18
  - View Audits

**Actions**
XML Validation Error, Location: /ead[1]/archdesc[1]/dsc[1]/c[3]/c[2]/d[1]/unitid[1]
For the persistent or current unitid, element unitid attribute identifier value MUST be unique. [Reference(s): WYNN/736] [identifier: WYNN/736]

XML Validation Error, Location: /ead[1]/archdesc[1]/dsc[1]/c[3]/c[2]/c[1]/d[1]/unitid[1]
For the persistent or current unitid, element unitid attribute identifier value MUST be unique. [Reference(s): WYNN/736] [identifier: WYNN/736]

XML Validation Error, Location: /ead[1]/archdesc[1]/dsc[1]/c[3]/c[5]/d[1]/unitid[1]
For the persistent or current unitid, element unitid attribute identifier value MUST be unique. [Reference(s): WYNN/808] [identifier: WYNN/808]
<table>
<thead>
<tr>
<th></th>
<th>EAD</th>
<th>Transformation Name</th>
<th>Organisation</th>
<th>Date/Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EAD</td>
<td>TidyThemed</td>
<td>Global Organisation for Workflow</td>
<td>18-Jul-2017 15:23</td>
</tr>
<tr>
<td>2</td>
<td>EAD</td>
<td>FixNormalisedDates</td>
<td>Global Organisation for Workflow</td>
<td>30-Nov-2016 18:04</td>
</tr>
<tr>
<td>3</td>
<td>EAD</td>
<td>RepairExtref</td>
<td>Global Organisation for Workflow</td>
<td>07-Dec-2017 18:11</td>
</tr>
<tr>
<td>4</td>
<td>EAD</td>
<td>FixLangcodes</td>
<td>Global Organisation for Workflow</td>
<td>26-Nov-2016 15:26</td>
</tr>
<tr>
<td>5</td>
<td>EAD</td>
<td>DecupAll</td>
<td>Global Organisation for Workflow</td>
<td>29-Jul-2016 10:40</td>
</tr>
<tr>
<td>6</td>
<td>EAD</td>
<td>TidySourceAndRules</td>
<td>Global Organisation for Workflow</td>
<td>17-Apr-2018 15:58</td>
</tr>
<tr>
<td>7</td>
<td>EAD</td>
<td>TidyEmphAlt</td>
<td>Global Organisation for Workflow</td>
<td>30-Nov-2016 18:21</td>
</tr>
<tr>
<td>8</td>
<td>EAD</td>
<td>TidyGeogName</td>
<td>Global Organisation for Workflow</td>
<td>26-Apr-2017 17:00</td>
</tr>
<tr>
<td>9</td>
<td>EAD</td>
<td>NormaliseLevels</td>
<td>Global Organisation for Workflow</td>
<td>16-May-2016 15:11</td>
</tr>
<tr>
<td>10</td>
<td>EAD</td>
<td>RepairRepoName</td>
<td>Global Organisation for Workflow</td>
<td>11-Dec-2017 12:59</td>
</tr>
</tbody>
</table>
Contributor
“Local”, specific to Contributor

Group
Common to multiple Contributors

Global
Common to all Contributors
The pipeline model

» Same model applies to all data formats:
  › EAD (archive descriptions)
  › EAG (repository descriptions)
  › EAC (people and organisations)
» Descriptions often created to work within one system
» Designed for one interface
» Constraints of what is achievable within one system
» Data may not be suitable for an aggregator, for interoperability, for different systems and environments
» ...but it is hard to modify systems or change cataloguing practices
» Need a compromise
The Enemies of Blend (1)

» Missing references
» Duplicate unique references
» Missing mandatory fields
» Titles that are just a name...or they are a page long
» Cutting and pasting resulting in Jean HÃƒÆ’Ã†â€™™
» Dates in all manner of formats: [ ( c ca c. ? - / 1890-1831
» Data that isn’t structured
» Customised fields

» Printouts first

» Levels values in welsh

» Unesco subject terms that aren’t from the unesco thesaurus

» Place as name, name as subject, subject as genre, subject as place...

» Ghost levels – physical versus intellectual
An archivist typically thinks about describing an archive in a way that a human user can understand, accessed from their reading room or website; we think about it as structured data for humans and machines, within a global context.
Blend: Archives Hub

Archives Hub

Search Results

Filter

Applied Filters

Level

collection

Filter by

Holding Repository

Level

collection (65)

section (246)

item (116)

Creator

Subject

Date

Digital Content

Relevance

Show

20

Expanded mode

64

Archive Collections

1

Themed Collections

0

Repositories

< Previous

1

2

3

...

4

Next >

Notes on Magic

19th Century

1

Show on map

Treatise on Magic

1854-1857

2

Show on map

Batak Pustaha [Magic Book]

n.d.

3

Show on map

'Religion, myth and magic in Tangala'
» OAI-PMH
  › Archives Portal Europe
  › Cendari

» Elastic Search API
  › Archives Wales
  › Micro sites
  › European Holocaust Research Institute
Enter the URL of your repository: https://oai.archiveshub.jisc.ac.uk/ArchivesHub

Select a set: Institution: GB-103 (University College London Archives)

Select a metadata prefix: ead

Select harvesting method: Harvest by verb ListIdentifiers/GetRecord (fail-safe)

Select an interval for your harvesting: 1 months

Harvest only on weekends? No

Set last harvesting date (if existing) - dd/MM/YYYY: 

Select a user profile: Standard

Activation status: Enabled