Making sense of digital collections

What I wish I knew before I started...

Adrian Brown, 24 January 2012
Things I wish I’d known #1

Ingest accounts for up to 90% of digital repository activity
Digital information is hugely (and increasingly) diverse
Hello Adrian Brown. We have recommendations for you. (Not Adrian?)

Adrian's Amazon.co.uk  Deals of the Week  Gift Certificates  Gifts & Wish Lists

Your Account  Help

Adrian's Amazon.co.uk  Recommended for you
(If you're not Adrian Brown, click here.)

These recommendations are based on items you own and more.

View: All  |  New Releases  |  Coming Soon

1. Transition
   by Iain Banks (Sep 3, 2009)
   Average Customer Review: ⭐⭐⭐⭐⭐ (21)
   In stock
   RRP: £18.99
   Price: £9.49
   30 used & new from £5.22
   I own it  |  Not interested  |  Rate this item
   Recommended because you said you owned Complicity (Abacus Paperback) and more (Fix this)

2. Zima Blue
   by Alastair Reynolds (April 30, 2009)
   Average Customer Review: ⭐⭐⭐⭐⭐ (8)
   In stock
   RRP: £18.99
   Price: £12.28
   24 used & new from £6.96
Using SPARQL With Geographic Data

The SPARQL endpoint for geo data is available at http://os.services.tso.co.uk/geo/sparql.

```sparql
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT * WHERE {
  ?s ?p ?o
} LIMIT 10
```

Sample Queries

This query returns a list and names of Westminster Constituencies, and their containing county electoral divisions.

```sparql
PREFIX rdf: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX admingeo: <http://data.ordnancesurvey.co.uk/ontology/admingeo/>
PREFIX spatialrelations: <http://data.ordnancesurvey.co.uk/ontology/spatialrelations/>

SELECT ?constituency ?constituencyname ?cdvision ?name WHERE {
  {?constituency rdfs:label ?constituencyname .}
  {?cdvision rdfs:label ?name .} 
  {?constituency admingeo:WestminsterConstituency .}
  {?constituency foaf:name ?constituencyname .}
  {?cdvision admingeo:CountyElectoralDivision .}
  {?cdvision foaf:name ?name .}
  {?constituency spatialrelations:contains ?cdvision .}
} LIMIT 10
```
Things I wish I’d known #3

You rarely get nice, standard, complete SIPs
DIGITAL PRESERVATION PROJECT BOARD

2nd Meeting

30 Sept 2008, 11am

HL Director of Information Services’ Office

AGENDA

1. Apologies
2. Minutes of Last Meeting
3. Highlight Report
4. Communications Plan – for approval
5. AOB
6. Date of Next Meeting – 15th December 2008, 2.30pm
Digital workflows are highly sensitive to variation
QED: There are as many ingest workflows as there are archivists to design them
Workflows...

Start

- Test of Archival Value Met?
  - No: Destruction
  - Yes: Collections Policy Met?
    - No: Transfer
    - Yes: Selection Criteria Met?
      - No: Transfer
      - Yes: Ingest
        - End

Or

End

No
Management of Semi-Current Records → Appraisal & Selection → Transfer → Preservation & Maintenance → Technology Watch → Resource Discovery → Delivery & Presentation
...etc...

Ingest Process for Web Archive Data Type

1. Archivist manually adds records to CALM series
2. Disk plugged in at agreed location and structure copied to network
3. Information copied into directory structure which matches CALM reference
4. Workflow manually kicked off
5. Workflow copies data to quarantine area and virus checks
6. 30 day delay
7. Virus check and fix
8. SDB gets CALM Reference from directory structure

LEGEND
- **SDB**: Applications or manual interface involved in the step
- **ARCHIVES**: Link to CALM
- **CALM**: Import data removed from working and quarantine areas
- **CALM**: Relevant information added to the CALM interface
- **SDB**: SDB identifies link to presentation area from CALM and substitutes URI to landing page
- **SDB**: SDB creates DU and manifestations. Imports CALM descriptive data as appropriate using API
- **SDB**: SDB builds structure duplicating that of CALM Reference using API

Houses of Parliament
Parliamentary Archives
These are all GOOD things!
Ingest workflow

Pre-ingest activity

Create, Acquire or Transform SIP

Quarantine SIP

Validate SIP

Characterise SIP

Identify formats

Validate objects

Extract metadata

Enhance SIP metadata

Generate AIP

Store AIP

End
We can automate the boring stuff
Workflow Details

<table>
<thead>
<tr>
<th>Workflow Context</th>
<th>Full ingest workflow with...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflow Definition</td>
<td>Full Ingest Workflow (Manual Selection, Store Metadata File, Thumbnails)</td>
</tr>
<tr>
<td>Workflow ID</td>
<td>19</td>
</tr>
<tr>
<td>Workflow State</td>
<td>Completed</td>
</tr>
<tr>
<td>Date Started</td>
<td>04.10.11 11:19:00</td>
</tr>
<tr>
<td>Date Finished</td>
<td>04.10.11 11:21:37</td>
</tr>
<tr>
<td>Number of Files</td>
<td>26</td>
</tr>
<tr>
<td>Total Size</td>
<td>2 MB</td>
</tr>
<tr>
<td>Collection Code</td>
<td>CALM/GH1</td>
</tr>
<tr>
<td>Submission name</td>
<td>CALM Export Test</td>
</tr>
<tr>
<td>Top Level Record</td>
<td>Sup Ingest 2</td>
</tr>
</tbody>
</table>

Step Progress

<table>
<thead>
<tr>
<th>State</th>
<th>Name</th>
<th>Progress</th>
<th>Started</th>
<th>Finished</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Select</td>
<td></td>
<td>04.10.11 11:19:00</td>
<td>04.10.11 11:19:07</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>Copy XIP Package</td>
<td></td>
<td>04.10.11 11:19:07</td>
<td>04.10.11 11:19:11</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>Virus Check</td>
<td></td>
<td>04.10.11 11:19:11</td>
<td>04.10.11 11:19:33</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>Finitest Check</td>
<td></td>
<td>04.10.11 11:19:35</td>
<td>04.10.11 11:19:35</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>Metadata Integrity</td>
<td></td>
<td>04.10.11 11:19:35</td>
<td>04.10.11 11:19:35</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>Content Integrity</td>
<td></td>
<td>04.10.11 11:19:57</td>
<td>04.10.11 11:20:19</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>Characterise</td>
<td></td>
<td>04.10.11 11:20:19</td>
<td>04.10.11 11:20:51</td>
<td>View</td>
</tr>
<tr>
<td>✔️</td>
<td>Store Files</td>
<td></td>
<td>04.10.11 11:20:51</td>
<td>04.10.11 11:20:55</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>Store Metadata</td>
<td></td>
<td>04.10.11 11:20:55</td>
<td>04.10.11 11:21:18</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>Store Metadata File</td>
<td></td>
<td>04.10.11 11:21:18</td>
<td>04.10.11 11:21:25</td>
<td></td>
</tr>
</tbody>
</table>
8.4 The Bainesse pottery.

The Bainesse pottery provides a useful start to the quantified sequences as it has a long series of phases spanning the second to earlier third century. There is comparatively little evidence of Flavian-Trajanic on the site, and, the material of this date, can be seen both in the low mortaria dated to first century in comparison second (Fig 000.00), especially in the low proportion...
Things I wish I’d known #5b

A digital archivist’s role is to manage the exceptions
Characterisation is one of the most important tasks
Characterisation

- Identification
  - What is it?

- Validation
  - Is it really it?

- Property extraction
  - What can we measure about it?
  - Technical and significant properties
Preservation planning
Collection profiling

**Formats**

- Acrobat PDF 1.3 - Portable Document Format (Version 1.4): 2%
- Adobe PDF Format (Version 1.3): 1%
- Adobe Portable Document Format (Version 1.0): 1%
- GIF Image File Format (Version 1.2): 1%
- Microsoft PowerPoint Presentation (Version 97-2002): 2%
- Portable Network Graphics (Version 1.0): 3%
- Extensible Markup Language (Version 1.0): 3%
- Other: 8%

**Age**

- 1990-1999: 23%
- 2000-2009: 22%
- 2010-2019: 18%
- 2020-2029: 16%
- Unknown: 4%

**Volume**

- JPEG File Interchange Format (Version 1.0): 18%
- Acrobat PDF 1.3 - Portable Document Format (Version 1.4): 12%
- Portable Network Graphics (Version 1.0): 8%
- Other: 8%

**Usage**

- Houses of Parliament Parliamentary Archives

---

![Graph showing file formats, age, volume, and usage](image_url)
Things I wish I’d known #7

Characterisation is very amenable to automation
Format identification

DROID (Digital Record Object Identification)

File List

<table>
<thead>
<tr>
<th>Status</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C:\Documents and Settings\Adrian\My Documents\Digital Archiving\File Formats\Test Files\Final Test Files\04 small.jpg</td>
</tr>
<tr>
<td></td>
<td>C:\Documents and Settings\Adrian\My Documents\Digital Archiving\File Formats\Test Files\Final Test Files\10 ASCII...</td>
</tr>
<tr>
<td></td>
<td>C:\Documents and Settings\Adrian\My Documents\Digital Archiving\File Formats\Test Files\Final Test Files\110 binary...</td>
</tr>
<tr>
<td></td>
<td>C:\Documents and Settings\Adrian\My Documents\Digital Archiving\File Formats\Test Files\Final Test Files\115...</td>
</tr>
</tbody>
</table>

Identification results

| File | C:\Documents and Settings\Adrian\My Documents\Digital Archiving\File Formats\Test Files\Final Test Files\04 small.jpg |
| PUID | fmt44 |
| Format | JPEG File Interchange Format |
| Version | 1.02 |
| Status | Positive (Specific Format) |
| Warning |  |

Add Files  Remove Files  Remove All

Identify  Cancel

File 84 of 84 analysed
Validation and property extraction
We can automatically extract vast amounts of metadata about digital objects
We probably shouldn’t
Things I wish I’d known #9

Don’t try to model the world
Models...

- Character encodings
- Compression methods
- File formats
- Software
- Storage media
- Operating systems
- Hardware platforms
- Protocols
...etc
Things I wish I’d known #10

You don’t have to solve every problem yourself - collaborate
Collaboration
Collaboration
Collaboration
Questions?

Web: http://www.parliament.uk/archives
Twitter: @UKParlArchives, @realAdrianBrown