Challenges in Digital Preservation – an ADS perspective

Dr Tim Evans 21-07-2016
The Archaeology Data Service:

• Established in 1996
• Based within the Department of Archaeology, University of York
• Digital archive for UK-based fieldwork and research in archaeology
  • Academic ‘research’ archives produced by Higher Education
  • The results of development-led fieldwork
What was the impetus in 1996?

- Is destructive!

- Resources are unique

- The preserved record becomes the main resource for future interpretation
Progress to-date

• Collections
  ○ 1,100,000 metadata records
  ○ 700+ rich archives
    ○ Raster + vector images
    ○ Databases
  ○ Geospatial data
Fig. 1. Major functions of the OAIS Reference Model from Consultative Committee for Space Data Systems (CCSDS), CCSDS 650.0-W-1, Producer-Archive Interface Methodology Abstract Standard. (OAIS). White Book, Issue 1, Draft Recommendation for Space Data System Standards.

http://archaeologydataservice.ac.uk/advice/preservation
Not just an off-site backup:

- Increasingly complicated systems stack
- Operating 25 Virtual machines supporting a variety of applications, databases etc
- File storage – including our own remote copies
- Validation checks (corruption of files)

DISASTER!
Fig. 1. Major functions of the OAIS Reference Model, as described in the Committee for Space Data Systems (CCSDS), CCSDS 650.0-W-4 RCM for the Methodology Abstract Standard, (OAIS), White Book, Issue 1, 2014. For details, refer to the Committee for Space Data System Standards.
Problem 1: People
Geographical Information Systems

**Preferred File Format**
- .SHP + .SHX
- .DBF
- ESRI Shapefile
- TIF + TFW
  - Geo-referenced
  - TIF Image
- GML
  - Geography Markup Language

**Accepted File Format**
- ESRI Grid
- .MIF + .MID
  - MapInfo Interchange Format
- .DDF
  - Spatial data transfer standard
- .EXP
  - MOSS
- .VPF
  - Vector product Format

**Documentation and Metadata**
- Date of capture/purchase
- Software, version & platform
- Purpose of GIS & what layers represents
- Details of coordinate systems or site grids & how data relates to them
- Method of data capture (e.g. total station survey)
- Any data source information (e.g. Purchased from OS)
- Scale/resolution of data at capture and at which it is stored
- Assessment of data quality (e.g. Root mean square error)

http://archaeologydataservice.ac.uk/advice
Museum of Archaeological Computing

The Museum was not intended as a static exhibition only but as a tool to facilitate data recovery where this is on now defunct media which is also operating system dependent in many cases.
<table>
<thead>
<tr>
<th>FID</th>
<th>Shape *</th>
<th>nyexcid</th>
<th>nmr_id</th>
<th>her_event_</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Point</td>
<td>ny0041</td>
<td>1023794</td>
<td>EYD6391</td>
</tr>
<tr>
<td>1</td>
<td>Point</td>
<td>ny0043</td>
<td>1023854</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Point</td>
<td>ny0044</td>
<td>1023984</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Point</td>
<td>ny0049</td>
<td>1028653</td>
<td>EYD6415</td>
</tr>
<tr>
<td>4</td>
<td>Point</td>
<td>ny0051</td>
<td>1030462</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Point</td>
<td>ny0052</td>
<td>1030840</td>
<td>ENY5461</td>
</tr>
<tr>
<td>6</td>
<td>Point</td>
<td>ny0053</td>
<td>1030859</td>
<td>ENY1032</td>
</tr>
<tr>
<td>7</td>
<td>Point</td>
<td>ny0058</td>
<td>1032430</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Point</td>
<td>ny0059</td>
<td>1032435</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Point</td>
<td>ny0060</td>
<td>1032437</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Point</td>
<td>ny0074</td>
<td>1033498</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Point</td>
<td>ny0078</td>
<td>1034566</td>
<td>ENY1793</td>
</tr>
<tr>
<td>12</td>
<td>Point</td>
<td>ny0083</td>
<td>1034965</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Point</td>
<td>ny0085</td>
<td>1034976</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Point</td>
<td>ny0086</td>
<td>1034980</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Point</td>
<td>ny0087</td>
<td>1034981</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Point</td>
<td>ny0089</td>
<td>1035020</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Point</td>
<td>ny0092</td>
<td>1035087</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Point</td>
<td>ny0098</td>
<td>1035182</td>
<td>EYD6521</td>
</tr>
<tr>
<td>19</td>
<td>Point</td>
<td>ny0099</td>
<td>1035195</td>
<td>ENYM587</td>
</tr>
<tr>
<td>20</td>
<td>Point</td>
<td>ny0104</td>
<td>1035291</td>
<td>EYD6527</td>
</tr>
<tr>
<td>21</td>
<td>Point</td>
<td>ny0116</td>
<td>1035938</td>
<td>EYD6549</td>
</tr>
</tbody>
</table>
Problem 2: Preservation via migration

• Every file accessioned preserved as a standardised format e.g. all raster images as uncompressed TIFF, databases as delimited text...

• Formats used based on a mixture of technical consideration (e.g. compression = data loss), but also judgements on longevity of format and ease of establishing future migrations.
Operational workflow
Example structure

- admin
  - licence.pdf
  - project_metadata
    - email_2016-04-30.txt
    - project_metadata.docx
  - {acc_no}
    - 2016-04-30
      - cd1
        - image_01.jpg
        - image_02.jpg
        - project_metadata.doc
      - cd2
        - database.mdb
        - db_documentation.doc
  - preservation
    - tif
      - image_01.tif
      - image_02.tif
    - csv
      - database_table1.csv
      - database_table2.csv
      - documentation
        - db_documentation.docx
  - dissemination
    - jpg
      - image_01.jpg
      - image_02.jpg
    - csv
      - database_table1.csv
      - database_table2.csv
      - documentation
        - db_documentation.pdf

http://archaeologydataservice.ac.uk
In more detail...

• Query of the ADS Object Management System returned:
  • 2,118,052 files
  • 860,190 of which are in SIP
    • 253 unique formats
  • By contrast only 28 unique formats used in the AIP
Most popular?

- 320,000+ .tif
- 180,000+ .jpg (1.01)
- 105,000+ .jpg (1.02)
- 20,000+ .jpg (raw)
- 17,000+ .pdf (1.4)
- 16,000+ .pdf (1.6)
- 13,000+ .doc (97-2003)
More data – less variation
"Cannot Open a database created with a previous version of your application" error in Access 2013 and Access 2016

This issue occurs when you try to use Access 2013 or later version of Microsoft Access to open an Access 97 database. To work around this issue, use a pre-Access 2013 version of Access to save the Access 97 database as an .accdb file:

2. On the Tools menu, click Database Utilities, click Convert Database, and then click to Access 2002-2003 file format.
3. Enter a name for the database, and then click Save.
5. Open the database in Access 2013 or later version of Access.
6. On the File tab, click Save As, select Access Database (*.accdb), and then click Save As.
7. In the Save As dialog box, click Save.

Cause

This issue occurs because recent versions of Access cannot convert Access 97 files.
Unusual formats

• 594 GSSI RADAN data file .dzt
  • A proprietary format created by a US-based geophysical company
  • At the moment *can only* be exported using RADAN software

• 21 Harris Matrix .hm
  • A stand-alone product created by a (now ended) research project
The document uses fonts on your system that are not available or are in a different location.

ZapfDingbats:
Font not found on the system.

Do you still wish to open the document?
User Reviews

Current Version

5 star 4
4 star 0
3 star 0
2 star 0
1 star 10
out of 14 votes

All Versions

5 star 4
4 star 0
3 star 0
2 star 0
1 star 10
out of 14 votes

Sort: [Date]

⭐⭐⭐⭐⭐ "If you want to waste 10 minutes..."
June 12, 2013 | By uk-guy567 | Version: Ai Viewer 3.2

Pros
None found

Cons
Crashes on opening under Win7
• CAD – defacto Autodesk’s AutoCAD software
• ADS had traditionally used DXF (R14) in AIP and DIP support for textual encoding (ASCII) and its primary purpose as an exchange format which could be used beyond Autodesk software
Due to the fast development of the AutoCAD software, the DXF format has seen almost as many version updates as the proprietary DWG format (which has seen eighteen new versions since 1982)

Effectively using an old format for new content – at risk of losing data!

As a result, the decision was made in early 2014 to change the ADS archiving policy and adopt DWG version 2010 (AC1024)
CAD migration
Lessons

- Did have a cost impact
- Mitigated by file management
- Perhaps most importantly: file format not as important as knowing what it was we were actually meant to be preserving = importance of metadata

Migrating 2 and 3D Datasets: Preserving AutoCAD at the Archaeology Data Service

*ISPRS Int. J. Geo-Inf.* **2016**, *5*(4), 44; doi:[10.3390/ijgi5040044](http://dx.doi.org/10.3390/ijgi5040044)
Will this ever change?
Future migrations

• **DOCX**
  
  • When ADS started using OOXML formats were using Office 2010, which only supported the ISO 29500 Transitional standard
  
  • Although it is anticipated that the transitional format will continue to be supported (including OpenOffice), is it better to move away now?
Problem 3: File management

- Query of the ADS Object Management System returned:
  - 2,118,052 files
  - 20+ Tb
- How do we know what we have?
- How do we manage ‘objects’ comprising different files/data streams?
Implementing **Fedora** (Flexible Extensible Digital Object Repository Architecture).
Ultimately: too complicated and with a high level of subsequent software development

Mundane tasks such as deletion of records were problematic

"Nothing that can't be achieved using a database and a programming language" (in our opinion!)

Fedora
<table>
<thead>
<tr>
<th>Object Id</th>
<th>Object Title</th>
<th>Data Type</th>
<th>Created Date</th>
<th>Object Representations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1328034</td>
<td>Map of traditional farmsteads in Greater Lincolnshire with house located with gable side on to yard</td>
<td>GIS</td>
<td>12-Sep-2015</td>
<td>side_on_point.dbf</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>side_on_point.prj</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>side_on_point.prj</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>side_on_point.shp</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>side_on_point.shx</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>side_on_point.gml</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>side_on_point.xsd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>side_on_point.zip</td>
</tr>
<tr>
<td>1328033</td>
<td>Map of traditional farmsteads in Greater Lincolnshire with detached farm houses</td>
<td>GIS</td>
<td>12-Sep-2015</td>
<td>detached_point.dbf</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>detached_point.prj</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>detached_point.shp</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>detached_point.shx</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>detached_point.gml</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>detached_point.xsd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>detached_point.zip</td>
</tr>
<tr>
<td>1328032</td>
<td>Map of traditional farmsteads in Greater Lincolnshire with additional detached elements to main plan</td>
<td>GIS</td>
<td>12-Sep-2015</td>
<td>d_point.dbf</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>d_point.prj</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>d_point.shp</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>d_point.shx</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>d_point.gml</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>d_point.prj</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>d_point.xsd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>d_point.zip</td>
</tr>
</tbody>
</table>

Related documents:
- JonathanParkhouse_Warwickshire.pdf
- LizPearson_Environmental.pdf
- MalcolmAtkin_Worcestershire.pdf
- MarkBowden_Earthworks.pdf
- MikeHodder_Birmingham.pdf
- PaulStamper_Countryside.pdf
- AnnWoodward_PotsPitsandMonuments_(2).pdf
- JamesCreig_Environmental.pdf
- JohnHalsted_BronzeAgeShropshire.pdf
- MikeD_British_Industrial.pdf
RUNS CHECKSUM
Error listing directory 'E:\ADS_preservation\Intarch\journal\issue38\7\urubici_5'.

Error retrieving file list for "E:\ADS_preservation\Intarch\journal\issue38\7\urubici_5\*.*".
The file or directory is corrupted and unreadable.
What’s an object?

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Type</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1373975</td>
<td>EAA_REPORT_9_VOL_2</td>
<td>Text</td>
<td>09-Nov-2015</td>
</tr>
<tr>
<td>1373974</td>
<td>EAA_REPORT_9_VOL_1</td>
<td>Text</td>
<td>30-Dec-2015</td>
</tr>
</tbody>
</table>

- EAA_REPORT_9_VOL_2.pdf
- EAA_REPORT_9_VOL_2.pdf
- EAA_REPORT_9_VOL_2.pdf
- EAA_REPORT_9_VOL_2.zip.001
- EAA_REPORT_9_VOL_2.zip.002

- EAA_REPORT_9_VOL_1.pdf
- EAA_REPORT_9_VOL_1.pdf
- EAA_REPORT_9_VOL_1.pdf
- EAA_REPORT_9_VOL_1.zip.001
- EAA_REPORT_9_VOL_1.zip.002
## Mis-identification

<table>
<thead>
<tr>
<th>File Name</th>
<th>MapInfo-GIS-files-for-the-Historic-Landscape-Character-Areas-and-Zones.TAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep Id</td>
<td>3356836</td>
</tr>
<tr>
<td>Purpose</td>
<td>Original</td>
</tr>
<tr>
<td>File Path</td>
<td>/ADS_presentation/arch-2248-1/original/5444/2016-02-04/MapInfo-GIS-files-for-the-Historic-Landscape-Character-Areas-and-Zones.TAB</td>
</tr>
<tr>
<td>File Type</td>
<td></td>
</tr>
<tr>
<td>File Type Ver</td>
<td></td>
</tr>
<tr>
<td>Pronom Id</td>
<td></td>
</tr>
<tr>
<td>Extension</td>
<td>tab</td>
</tr>
<tr>
<td>File Size</td>
<td>208</td>
</tr>
<tr>
<td>Last Modified Date</td>
<td>04-Feb-2016</td>
</tr>
<tr>
<td>Mime Type</td>
<td></td>
</tr>
<tr>
<td>Checksum</td>
<td>143d4b27e18936a0a93cdd348bb4ec61</td>
</tr>
<tr>
<td>Updated</td>
<td>04-Feb-2016</td>
</tr>
<tr>
<td>Updated By</td>
<td>kjn103</td>
</tr>
<tr>
<td>File Name</td>
<td>detached_point.gml</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Rep Id</td>
<td>3324042</td>
</tr>
<tr>
<td>Purpose</td>
<td>Preservation</td>
</tr>
<tr>
<td>File Path</td>
<td>/ADS_preservation/arch-2128-1/preservation/gml/GIS_Layers_and_Data/Farmstead_data Extractions/House Position/detached_point.gml</td>
</tr>
<tr>
<td>File Type</td>
<td>Extensible Markup Language</td>
</tr>
<tr>
<td>File Type Ver</td>
<td>1.0</td>
</tr>
<tr>
<td>Pronom Id</td>
<td>fmt/101</td>
</tr>
<tr>
<td>Extension</td>
<td>gml</td>
</tr>
<tr>
<td>File Size</td>
<td>11672363</td>
</tr>
<tr>
<td>Last Modified Date</td>
<td>17-Nov-2015</td>
</tr>
<tr>
<td>Mime Type</td>
<td>application/xml, text/xml</td>
</tr>
<tr>
<td>Checksum</td>
<td>3ae4f0ee91523b56b75b99b880ae29d</td>
</tr>
<tr>
<td>Updated</td>
<td>04-Dec-2015</td>
</tr>
<tr>
<td>Updated By</td>
<td>te1</td>
</tr>
</tbody>
</table>
Thanks!

tim.evans@york.ac.uk
http://archaeologydataservice.ac.uk/blog/
https://twitter.com/ADS_Chatter
https://twitter.com/ADS_Update