Using DROID to create file level metadata:

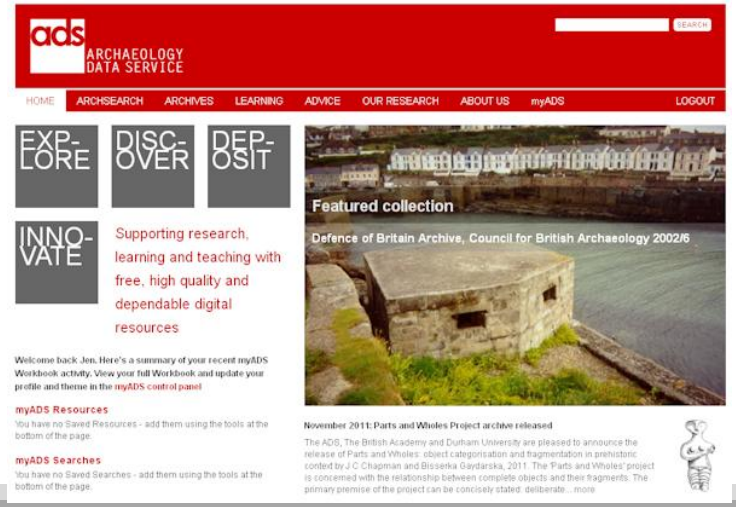
a case study from the Archaeology Data Service

Jenny Mitcham
Jenny.mitcham@york.ac.uk

A bit of background...

- Digital archive set up in 1996
- Originally part of the Arts and Humanities Data Service (AHDS)
- We archive data produced by archaeologists
- We make that data freely available on-line

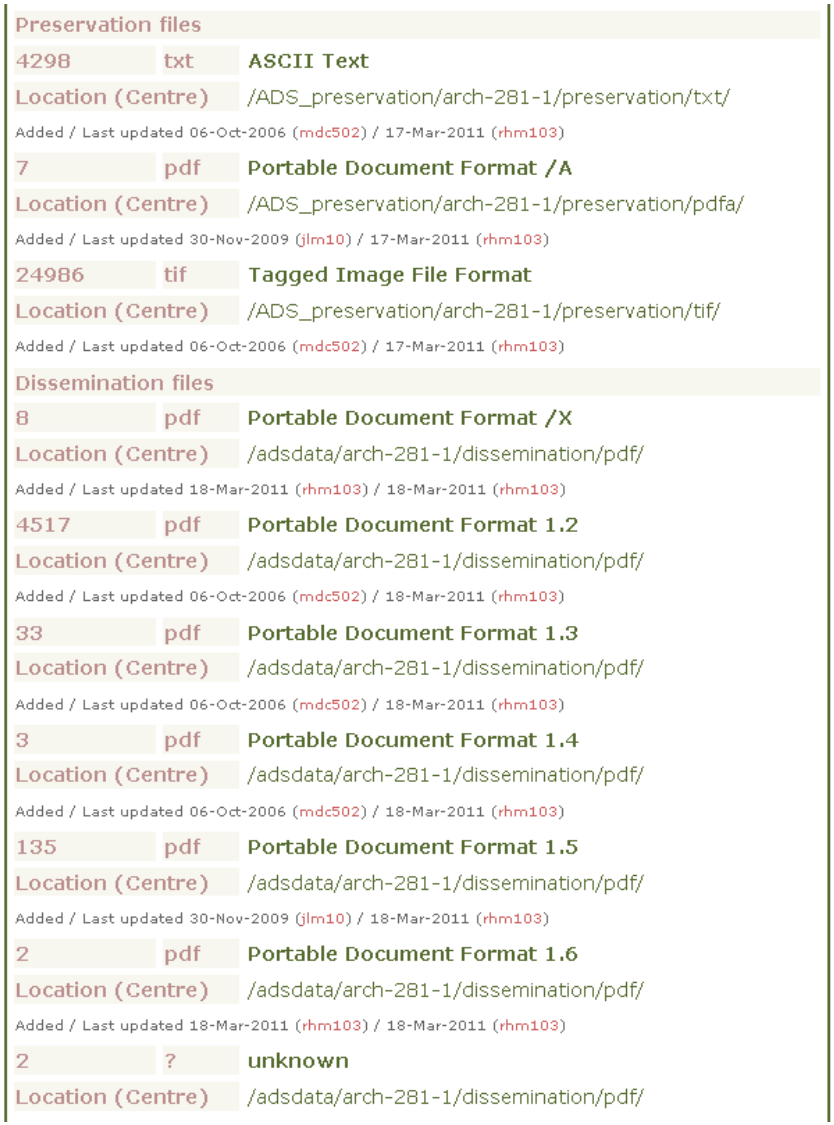
http://archaeologydataservice.ac.uk/



Metadata

- In the early days of the ADS/AHDS we recorded:
  - collection level metadata:
    - about the project as a whole (who, where, what, when)
  - metadata for groups of files:
    - a batch of GIS files
    - a set of photographs
- NO file level metadata

How we used to record files...



Preservation files

4298 txt ASCII Text

Location (Centre) /ADS_preservation/arch-281-1/preservation/txt/

Added / Last updated 06-Oct-2006 (mdc502) / 17-Mar-2011 (rhm103)

7 pdf Portable Document Format /A

Location (Centre) /ADS_preservation/arch-281-1/preservation/pdfa/

Added / Last updated 30-Nov-2009 (jlm10) / 17-Mar-2011 (rhm103)

24986 tif Tagged Image File Format

Location (Centre) /ADS_preservation/arch-281-1/preservation/tif/

Added / Last updated 06-Oct-2006 (mdc502) / 17-Mar-2011 (rhm103)

Dissemination files

8 pdf Portable Document Format /X

Location (Centre) /adsdata/arch-281-1/dissemination/pdf/

Added / Last updated 18-Mar-2011 (rhm103) / 18-Mar-2011 (rhm103)

4517 pdf Portable Document Format 1.2

Location (Centre) /adsdata/arch-281-1/dissemination/pdf/

Added / Last updated 06-Oct-2006 (mdc502) / 18-Mar-2011 (rhm103)

33 pdf Portable Document Format 1.3

Location (Centre) /adsdata/arch-281-1/dissemination/pdf/

Added / Last updated 06-Oct-2006 (mdc502) / 18-Mar-2011 (rhm103)

3 pdf Portable Document Format 1.4

Location (Centre) /adsdata/arch-281-1/dissemination/pdf/

Added / Last updated 06-Oct-2006 (mdc502) / 18-Mar-2011 (rhm103)

135 pdf Portable Document Format 1.5

Location (Centre) /adsdata/arch-281-1/dissemination/pdf/

Added / Last updated 30-Nov-2009 (jlm10) / 18-Mar-2011 (rhm103)

2 pdf Portable Document Format 1.6

Location (Centre) /adsdata/arch-281-1/dissemination/pdf/

Added / Last updated 18-Mar-2011 (rhm103) / 18-Mar-2011 (rhm103)

2 ? unknown

Location (Centre) /adsdata/arch-281-1/dissemination/pdf/

- Record what types of files we have
- Record versions where we know them
- ...but not consistent
- ...and how do we know which files are which?
- ...and we don't always have all the answers

Weird and wonderful file formats

- GIS (e00, shp, dbf, gml, tfw, rrd, aux, xml)
- Photogrammetry (dxf, wrl, cmr, tif, jpg, obj, x3d)
- Geophysics (txt, csv, dat, rep, tif, grd, sta, his, plm, lst, gip, cip, cmd, cms, tem)
- Marine survey (bag, xtf, dat, txt, xyz, gf3, gsf, mgd77, ddf, segy)
- Laser scanning (xyz, obj, mtl, dxf, dwg, sdts)

Solution to many problems... 'The Big Table'

- Better handle on what we've got
  - Preservation planning
  - What needs migrating?
- Better integrity checking
  - Automatic checking of files by comparing checksums
- Better resource discovery
  - GIS files showing Roman sites
  - Databases of Medieval pottery
- Better control of on-line access
- Standardisation of downloads

ADS+ project

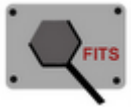
- Arts & Humanities Research Council (AHRC) funded project (DEDEFI)
- Re-design of our old Collections Management System (CMS)
- Incorporating file level metadata
- Allow deeper and richer object-level access to our collections

Why DROID?

JHOVE2

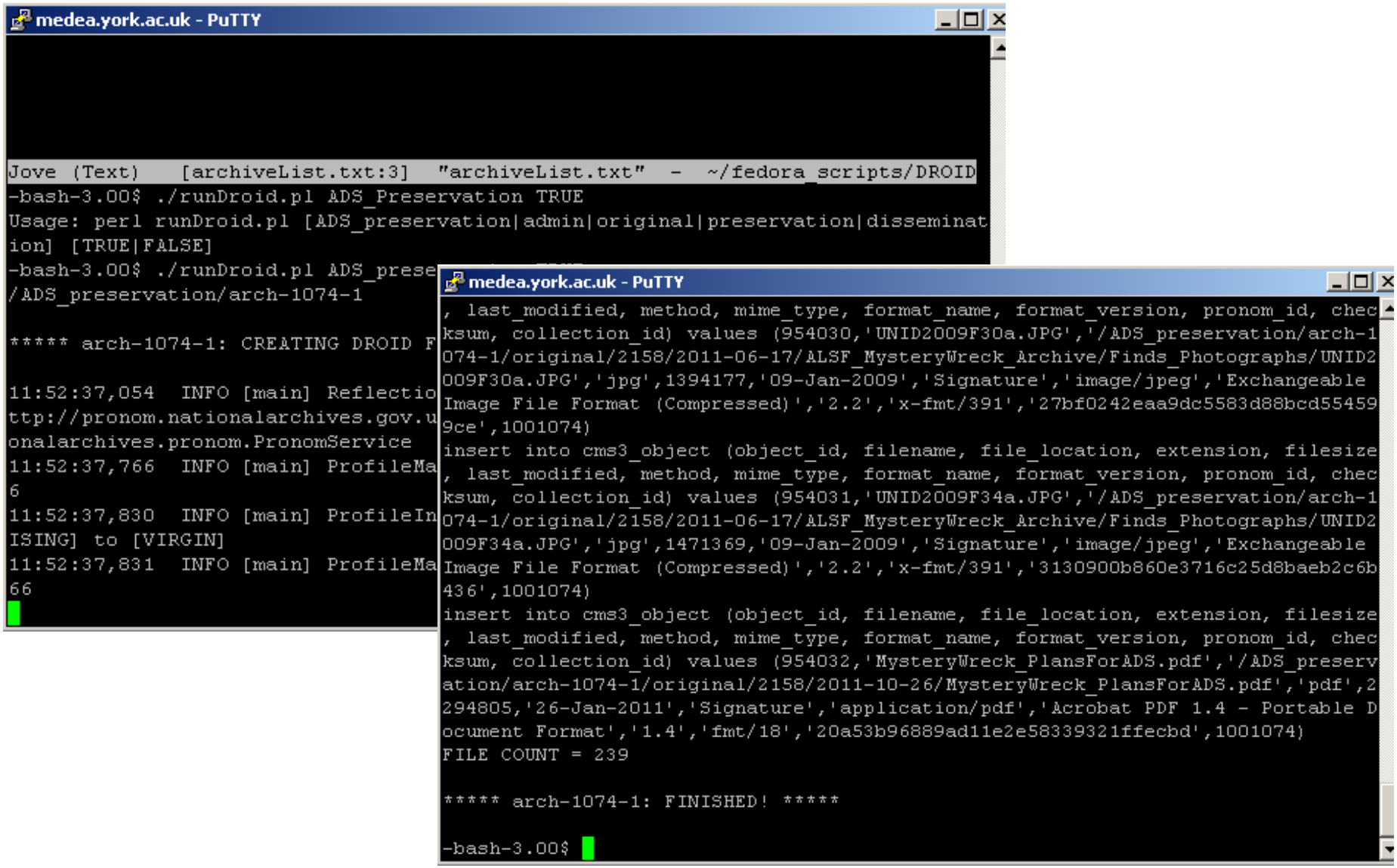
DROID
Digital Record Object Identification

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND



- Active development
- Good support
- We can work with TNA to extend the range of file types

Running DROID: command line



Displaying file level metadata

dissemination:

File	Info	Filename	Method	Pronom Id	MIME Type	Format Name	Version
1	?	02201001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
2	?	02201002.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
3	?	02201003.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
4	?	02202001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
5	?	02202002.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
6	?	02203001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
7	?	02204001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
8	?	02204002.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
9	?	02205001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
10	?	02206001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
11	?	02207001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
12	?	02208001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
13	?	02208002.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
14	?	02209001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
15	?	02209002.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
16	?	02209003.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
17	?	02210001.pdf	Signature	fmt/16	application/pdf	Acrobat PDF 1.2 - Portable Document Format	1.2
18	?	022tl001.pdf	Signature	fmt/16	applicat		
19	?	022tl002.pdf	Signature	fmt/16	applicat		
20	?	02301001.pdf	Signature	fmt/16	applicat		

Filename: 02201001.pdf

File Location:	/adsdata/arch-281-1/dissemination/pdf/022/02201001.pdf
Filesize:	205552
Checksum:	6330e2ef25bddd98d7807c918c6afbb8

close

Submitting files to TNA

	send Andrew quite a lot of detail about the formats		
DT1/HD	GPR files from Sensors and Software hardware/software. We have been sent some of these in collection 498 (Hadra archive) and DROID doesn't know what to do with them. Fortunately we already did some work on these file formats as part of the Big Data Project which is documented on the wiki so have been able to send Andrew quite a lot of detail about the formats	a selection of files from the /profiles/ directory in the buspark folder from coll 498 (we can send more if he wants!)	29th June 2011
SEG-Y	GPR files - good preservation format for ADS though I don't think we have any files as yet. Sent Andrew link to technical specification and suggested they concentrate of version 1	none	29th June 2011
E00 (and X00)	GIS files - ArcInfo Interchange Format - used to be our preferred preservation method but now dropped in favour of GML. We still have a few of them in our archive though	5 sample files sent from 731, 401, 373	29th June 2011
TIF/TFW /AUX/RRD	Tif world file	2 sets of sample files from 342 and another collection	29th June 2011
MP4	Digital video. We have 3 of these as part of the VENUS archive (that Jon migrated from Xvid format) but Droid won't identify them (even though PRONOM knows about mp4 files	sent the smallest 2 of the 3 files from VENUS archive	29th June 2011
M2V	Digital video. We have a few of these as part of Stuart's phd archive (that we migrated from avi format)	sent 2 of the smallest ones from Stuart's phd	29th June 2011
HTML	Most html files are identified by DROID but we have a few that aren't. It seems to id files that start with `<html>` but not fragments of a file that don't start at the begining or files with an early doctype declaration at the start, like this one in Assemblage `<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN//4.0">`	Sent a batch of files that do and don't work - some from Assemblage and some from Tell Brak / Kilise Tepe archive	29th June 2011
GRD/GRS /DAT/STA/HIS	Geoplot files	Sent a batch from collection 410 and also pointed him to dissemination pages for Trent-Soar archive	29th June 2011
DOC	Word perfect files. DROID seems to id by extension only and	Sent a sample file from collection 1001	14th July

...and TNA's response

The National Archives

MyPage (not signed in)

Search the archives | Search

Advanced search

About us | Education | Records | Information management | Shop online

You are here: Home > Information management > Our projects and work > Digital preservation > PRONOM > Release notes

The technical registry
PRONOM

Welcome : About Add an entry
Search ? Help Information resources

Release notes

7th September 2011
DROID_SignatureFile_V52.xml

The following changes have been made to PRONOM and the DROID signature file:

New Records

fmt/360: pulse EKKO data file. Outline entry added. Submission from Archaeology Data Service.
fmt/361: pulse EKKO header file. Outline entry added. Submission from Archaeology Data Service.
fmt/362: GSSI SIR-10 RADAN data file. Outline entry added. Submission from Archaeology Data Service.
fmt/363: SEG Y Data Exchange Format Generic. Outline entry added. Submission from Archaeology Data Service.
fmt/364: National Imagery Transmission Format 1.0. Outline entry added.
fmt/365: National Imagery Transmission Format 2.0. Outline entry added.
fmt/366: National Imagery Transmission Format 2.1. Outline entry added.
fmt/367: ESRI World File Format. Outline entry added. Submission from Archaeology Data Service. Format information supplied by the Geospatial Multistate Access and Preservation Partnership (GeoMAPP).

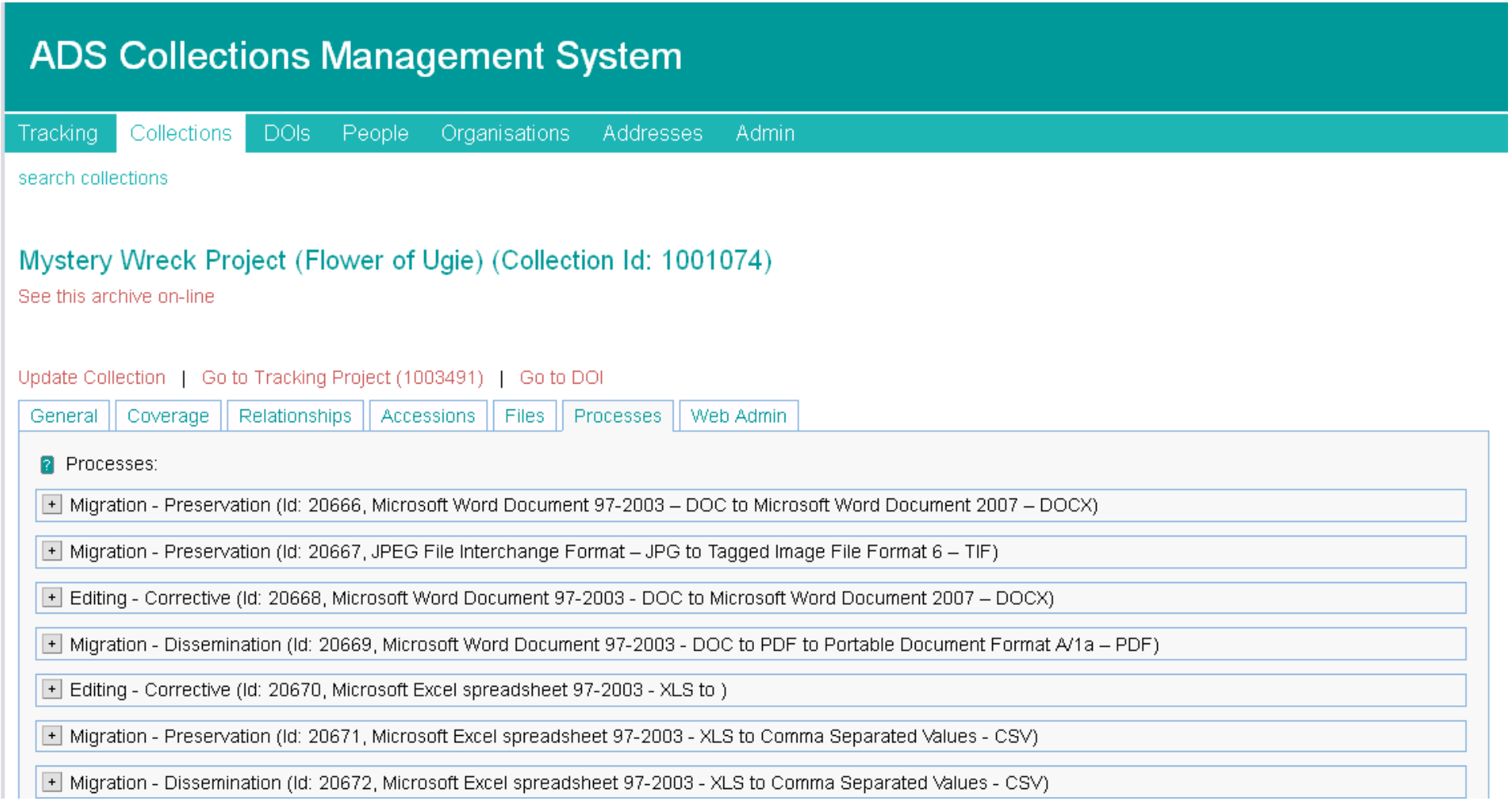
Issues with our implementation of DROID

- Highlighting files not identified
- Highlighting files with multiple identifications
- What is unique identifier of object table?
- Link object table to resource discovery metadata for files
- Make use of file extension mismatch warning?
- Inclusion of button in CMS to run DROID
- Create summary display of files in CMS so grouped by file type / version
- Start actually *using* the object data...

How DROID could be improved...

- Some files not identified
- Multiple identifications
- Problem files (that aren't what they claim to be)
- Handling of zip files
  - We need to know that we have a zip file
  - We need to know what's inside the zip file
- More case studies of DROID in action
- Advice and sharing of knowledge on obsolescence of formats/versions
- Software feature
- PRONOM data as web service/SPARQL?

Recording processes in ADS CMS



Keep in touch

Thank-you for listening!

Follow us on Twitter:

@ADS_Update

Follow us on Facebook:

http://www.facebook.com/archaeology.data.service

E-mail: jenny.mitcham@york.ac.uk
Website: http://archaeologydataservice.ac.uk/