Archaeology Data Service

Monitoring Data and Data Integrity

Kieron Niven, Digital Archivist: Data Standards
6th December 2022

Important to know where your data is.
Monitoring data:

- during submission and ingest
- during archiving process
- ongoing basis

ads Archaeology Data Service

Types of check:

During archiving submission, ingest, and archiving:

- Unaccessioned holdings (backup)
- Teracopy (moving data, verifying files)
- data receipts
- DROID (while data being worked on)

CMS example DROID run:

05/09/2022	05/09/2022		Database Rows updated=7 File count=9; Droid count=9
14:51:54 05/09/2022	14:52:03 05/09/2022	FINISHED	Database Rows updated=5 File count=5; Droid count=5 /adsdata/arch-4159-1/dissemination/210701_CSJV_WP011_WA_1S19CONTT_archive/Spreadsheets/1S19CONTT_CNA_ZOOARCHAEOLOGY.csv old=95643e87254af24593a09f091c76a30f new=e28b02650a63aa5b4cb27665f5a177cd
14:49:40 05/09/2022	14:49:49 05/09/2022	FINISHED	Database Rows updated=5 File count=5; Droid count=5 /ADS_preservation/arch-4159-1/preservation/210701_CSJV_WP011_WA_1S19CONTT_archive/Spreadsheets/1S19CONTT_CNA_ZOOARCHAEOLOGY.csv old=95643e87254af24593a09f091c76a30f new=e28b02650a63aa5b4cb27665f5a177cd
14:42:05 05/09/2022	14:42:15 05/09/2022	FINISHED	Database rows deleted=2 Database rows inserted=2 Database Rows updated=8 File count=10; Droid count=10
12:18:35 29/07/2022	12:18:35 29/07/2022	NO PREVIOUS	
12:18:35 29/07/2022	12:20:09 29/07/2022	FINISHED	Database Rows updated=741 File count=741; Droid count=741

ads Archaeology Data Service

Types of check overview:

Once archiving finished (ongoing checks)

- **Location** - files exist in correct place
- **Integrity** - files are what they should be (content check)

ads Archaeology Data Service

Ongoing checks: methods

- Two main types: **simple** check and **fixity** check.
- Based on the **manifest file.**
- Carried out on a regular basis:
  - **simple** on 1st and 14th each month.
  - **complex**: once every 3 months.
- These fit in with backup schedule…

ads Archaeology Data Service

Types of check: Backups

- 28 days of snapshots.
- 90 day tape backup.
- Amazon S3 Glacier storage (AWS).

The Manifest file:

- Simple text file
- Stored with the preservation and original data in the */admin/* directory
- Contains:
  - Checksum
  - File path and name

The Manifest file: Basis for all checks

```
# MD5 checksums for ADS archive 1004927 (preservation)
# Generated Mon Nov 28 11:41:30 GMT 2022

214d40f2d9b4052f88a00c3541b1228b  admin/licence.pdf
d58ad8c462b2737bcb9469db0cb356d4  admin/project_metadata/20015545_receipt.txt
2113bf781028cdb988ebf571a5351cd4  admin/project_metadata/archive_release_email_1.txt
4437d5e6a12f0a67109c26c1721901ce  admin/project_metadata/deposit_receipt_1011995.csv
07be742b93fed587486443df0d7f588d  admin/project_metadata/email_deposit_1011995.txt
691ae9b8ac96893bb797302806f7288d  original/11995/2022-11-22/Z4012219.jpg
2528f70d463f76993e0e7c12595580b9  original/11995/2022-11-22/Z4012221.jpg
3741c4f43d11fde148cb3b5e5a2161f2  original/11995/2022-11-22/Z4012222.jpg
242be1ac9f8e424442c3a777b1583d3b  original/11995/2022-11-22/Z4012247.jpg
084634df10bf3118b9eb6e0557f681c4  original/11995/2022-11-22/Z4012257.jpg
d1a6f903fffd87f95aac1a85ade37ce5  original/11995/2022-11-22/Z4012287.jpg
05b6a62950d31235a8809ff6e4296a90  original/11995/2022-11-22/Z4012298.jpg
3687838dc5be9c60c5300f5de583c54b  original/11995/2022-11-22/Z4012304.jpg
1d5243690492a14d2317d082de141fdb  original/11995/2022-11-22/Z4012310.jpg
02b0a893f942ed6bc966d91510f07c33  original/11995/2022-11-22/Z4012320.jpg
9a14ffa4143a642fd45ebff2b14bd4cf  original/11995/2022-11-22/Z4012321.jpg
```

ads Archaeology Data Service

Types of check: Filename check

- Simple, quick.
- Compares files in file store to details in the manifest.
- I.e. identifies where files are missing or have not been properly processed (i.e. not included in the manifest).
- Initially produced a lot of results due to legacy data (non-standard file structures, old archives where DROID has not been run, or not been updated).

ads Archaeology Data Service

Types of check: Complex fixity check

- Slow to run: 5-6 days to complete.
- Some collections are so large they need to be scheduled separately.
- Uses MD5 checksum - based on legacy use and compatibility with DROID.
- Produces few results, considered serious (implies corruption, unauthorised change, etc.).

ads Archaeology Data Service

Both Checks are just that.

- Checks to highlight issues, not fixes.
- Require a backup strategy to address issues.
- E.g. ADS can restore data from snapshots or AWS, or tape (WCS!).
- Each option differs by cost, length of access, ease to restore.
- Be selective about checking and backups (ADS largely backup just original and preservation data).

Other Resources:

- DPC Handbook:
  - https://www.dpconline.org/handbook/technical-solutions-and-tools/fixity-and-checksums
- DPC Technology Watch Guidance Notes “Which checksum algorithm should I use?”
  - https://www.dpconline.org/docs/technology-watch-reports/2399-twgn-checksums-addis/file
- Results of the NDSA 2021 Fixity Survey and Fixity Case Studies:
  - https://ndsa.org/2021/11/03/results-of-the-2021-fixity-survey-and-fixity-case-studies.html

ads Archaeology Data Service

Monitoring Data and Data Integrity

kieron.niven@york.ac.uk

Archaeology Data Service
Department of Archaeology
University of York
The King's Manor
Exhibition Square
York, YO1 7EP

www.archaeologydataservice.ac.uk

help@archaeologydataservice.ac.uk