Please cite this as: Evans, T., Hollander, H., Jakobsson, U. Gilissen, V. and Wright, H. 2024 Understanding current practice through case studies from established repositories, Internet Archaeology 67. https://doi.org/10.11141/ia.67.3

Archiving Archaeological Data: Understanding current practice through case studies from established repositories

Tim Evans, Hella Hollander, Ulf Jakobsson, Valentijn Gilissen and Holly Wright

The article describes and compares key policies and workflows from three long standing digital repositories for archaeological data: Data Archiving and Networked Services (DANS) of the Netherlands, the Swedish National Data Service (SND), and the Archaeology Data Service (ADS) from the United Kingdom. The workflows examined are representative of operational workflows around data assessment and data removal that are common to all data repositories, as well as specific strategies for dealing with Microsoft Access databases, a commonly deposited file format used by archaeologists for assessment and analysis.

The article presents each workflow as a succinct case study with an emphasis on why the decisions have been made to follow a certain method. This is followed by a discussion on similarities and differences in approach and implementation in order to bring together core recommendations that can be used by others who are currently building or scoping the development of new digital repositories in the archaeological sector. Links are also provided to copies of the detailed workflows and policies deposited within the Community Owned Digital Preservation Tool Registry (COPTR), a finding aid for helping find tools to deal with practical data management issues.

1. Introduction

At the time of writing, the importance of digital preservation for archaeological and heritage-based data is more prominent than at any other point in the history of our discipline (Novak *et al*. 2023; Richards *et al*. 2021). Recent collaborations and projects, most notably SEADDA, have helped highlight the success stories but also the many and imminent risks and threats that still exist (Jakobsson 2021). One of the recommendations of SEADDA is that in order to counter these risks, we should establish communities of practice so as to better share tools, approaches, policies, and workflows. This is not a simple task of course; those organisations with the responsibility of curating archaeology will vary in size, may be autonomous or part of a wider infrastructure or organisation, have different funding models (including limited funding), and have to deal with variations in localised practice and legislation. Accordingly, there can never be a one-size-fits-all approach to digital preservation in our sector that can be held up as the “correct” way of working. However, thanks to the endeavours of organisations such as the Digital Preservation Coalition (DPC), there is now a more unified sense of what good digital preservation comprises and guidance on the routes to best practice (DPC 2015). This best practice extends into the world of skills and knowledge, with a recent and renewed emphasis on review and refreshment to guard

against any erosion of capacity and to ensure that skills, technology, and policy remain fit for changing purposes (Currie and Kilbride 2021).

However, the amount of digital preservation tools is ever increasing and the landscape of information vast and dispersed amongst websites, blogs, conference papers, and social media updates (Cushing *et al*. 2021 53-54). A reaction to this has been the emergence of registries, principally the Community Owned Digital Preservation Tool Registry (COPTR), which aims to curate information on tools and workflows so as to aid finding, understanding, and implementation (Mita 2015). One of the potential weaknesses of this approach is the lack of discussion and reflection on these mechanisms and protocols that may guide a user to understand why something was implemented and its relevance to their specific scenario and needs (Cushing *et al*. 2021 59). This is a challenge that should be met by the community of archaeological data curators and principally those with long-standing policies and workflows that have changed or evolved over time. Organisations such as the Archaeology Data Service (ADS) have large portfolios of documents that describe their policy and strategy, often to very specific detail, and over time have developed in-house applications or utilised an array of proprietary and open source tools to achieve these aims and objectives. However, for an outsider to understand the "why", it is imperative that efforts are made to better define and articulate these case studies so that they can be assessed relevant to the needs of the user.

The following paper is thus an attempt to introduce and discuss examples of workflows from three digital preservation practitioners in our sector. These case studies comprise three workflows, two procedural and one technical:

- **Assessing information:** This is an example of a workflow that is undertaken for all datasets regardless of size or origin which is specifically stated as CoreTrustSeal requirement R08 - *the repository accepts data and metadata based on defined criteria to ensure relevance and understandability for users* (CoreTrustSeal Standards and Certification Board 2022)
- **Deaccessioning data:** This is an example of a less frequent workflow (for example, for 2022/23, SND deaccessioned two datasets, and ADS one dataset) and one that also may not be immediately anticipated by organisations setting up or developing new repositories. Although not an overarching CoreTrustSeal requirement, evidence of the repository's approach to deleting/removing data and metadata from collection/holdings, including the impact on persistent identifiers (PIDs), is nested under requirement R09 - *the repository assumes responsibility for long-term preservation and manages this function in a planned and documented way (CoreTrustSeal Standards and Certification Board 2022)*.
- **Process for dealing with Microsoft Access databases:** Access databases are a common format across all three repositories. Having defined strategies and workflows for specific file formats is an essential part of a number of CoreTrust Seal (2022) requirements, including:
    - R09: *File formats and metadata schemas for long term preservation*.
    - R11: *How workflows are adjusted for different types of data and metadata*.

The specific and detailed workflows for assessing information, deaccessioning data, and preserving Microsoft Access Databases have been deposited in the COPTR registry, and the case studies themselves are designed to provide an overview of key points, each followed by a short commentary.

Swedish National Data Service (SND)

SND started as SSD (Swedish Social Science Data Service) in 1981. After a call from the Swedish National Research Council, SSD became SND in 2008 and added Humanities and

Health medicine to its scope of Social Science. In 2016, SND became a consortium of seven universities, and by 2018 it was running as a functional consortium of nine universities (and a new funding period). SND's metadata catalogue currently holds around 2,700 datasets. Approximately 2,000 of these are accessible directly via SND, the rest via external sources. Out of these, roughly 650 relate to the field of archaeology and history. On average, SND accepts 100 depositions per year from researchers at universities, research infrastructures, and some local authorities (health data from municipalities).

Archaeology Data Service (ADS)

The ADS was established in 1996 and currently holds over 5 million files in 348 unique formats. The ADS receives on average 700 data collections per year, the majority through an online deposition tool called ADS-easy. A separate online reporting system called OASIS is also the route for accessioning around 7,500 grey literature reports each year. Depositors are primarily commercial archaeological units, but also include higher education institutions and independent researchers.

Data Archiving and Networked Services (DANS)

DANS was established in 2005 (see Hollander 2021), and currently holds *c.*150,000 archaeological datasets, comprising over 2 million files. In 2022, DANS launched the domain-specific Data Station Archaeology. This Data Station receives around 2,500 manual deposits a year, with 800 automated deposits (via the provincial depot system), and on average 18,000 Portable Antiquities of the Netherlands (PAN) deposits of metal finds found by members of the general public. Depositors are commercial archaeological organisations, either directly or via the provincial depot, academics, and PAN.

2. Assessing Information

COPTR Workflow: Assessing information

A key workflow for any repository is the checks and processes carried out on data when initially deposited by the data producer (Lavoie 2014, 12). Often these checks are - by the very nature of repositories - a technical exercise focussed on criteria including virus scans, corruption of data, and file authenticity and integrity. Another facet is more human-orientated and focussed on the usability (quality) of the (meta)data, and thus one that requires expertise and judgement of the repository staff involved. Repositories may take different approaches to how and when these assessments are made, as well as the course of resolution when issues arise.

The following case studies present each repository's method and approach to accepting and assessing data within their jurisdiction.

2.1 Case Study SND: assessing information

SND currently uses an in-house built management and documentation system called DORIS (DataORganization and Information System) for accepting and assessing data. DORIS is the only way to deposit data to SND. It can be used by universities using local storage that still desire to use SND's catalogue to describe and disseminate the project data. It is possible to make metadata searchable and accessible via SND's web catalogue without submitting data to SND. It is possible to describe data via DORIS that has already been published elsewhere if the data has a PID at that place.

The system is accessible by researchers and support staff from almost all universities in Sweden, researchers from other countries if they register an account, and SND staff. From SND's perspective, the researcher is the person who deposits the data and can be any project member appointed to perform this role. This is normally the support staff at most universities organised into what SND calls Data Access Units (DAU), however not all universities have such a unit at the time of writing. SND's role is classified as Research Data

Advisor. When using DORIS, (meta)data can be published in SND's research data catalogue.

The workflow for the use of DORIS is described in two documents (both in Swedish): *Kravbeskrivningen* [PDF] ("requirements description") and *SND's policy för granskning av data och metadata* [PDF] (SND's policy for reviewing data and metadata), and in the *DAU-handboken* (wiki in Swedish for DAU staff)). When a researcher wants to describe data in DORIS, they can choose between several profiles (Earth and Related Environmental Sciences, Engineering and Technology, General, History and Archaeology, Language Resources, Medical and Health Sciences, Natural Sciences, and Social Sciences) that consist of several optional or mandatory metadata fields.

Depending on what type of data and possible legal limitations, the researcher either deposits a copy of the data to SND CARE (SND's CoreTrustSeal-certified repository) or to their own university's storage solution, and then describes their data in DORIS. If data is deposited to SND CARE, it is SND staff that assesses the (meta)data. If the researcher is affiliated with a university that has a DAU, the DAU participates in the assessment. If data is deposited to the university's own storage, it is the DAU that assesses the information and checks the data to make sure that there is enough metadata and that the data file is readable, usable, and understandable. Only one entity (researcher/DAU/SND) can make changes to the metadata at a time and according to the Status of the record (see Figure 1). Communication between the researcher and the DAU, the researcher and SND, or the DAU and SND is documented in DORIS. When an entity considers itself finished with the editing, it passes on the editing right to the next entity or returns the right to a previous one for completion (see Figure 1).

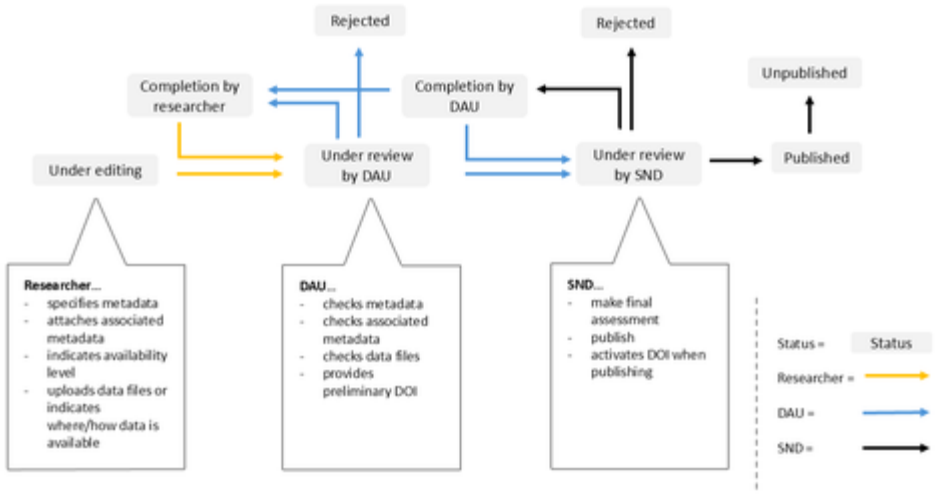

Figure 1: Status flow chart for publishing a new data description in DORIS (SND)

Data descriptions of data stored at SND CARE can only be published by SND. When SND assesses that the data description and any file(s) can be published, the status is changed to Published. If the data is stored locally, it is the local DAU that publishes the data description and the data. Each published dataset is assigned a digital object identifier (DOI) upon publication. However, a preliminary DOI is set at an early stage, but is only visible to SND staff. This can be shared with the researcher when needed if a journal requires a DOI before publication of an article. The preliminary DOI is not valid before the data description is published.

When assessing the data description (metadata in the web form based on different profiles), documentation, and data file(s) for data deposited to SND CARE, SND starts with the assumption that the data, attached documentation, and metadata are sufficient to allow a

potential secondary user to find the data, assess its reusability , and re-use the data without needing to contact the researcher(s). The data must “stand by itself”. This means that SND staff check the data description, read all attached documentation to make sure that there is enough information, and open the data file(s) to check variable labels, possible abbreviations, etc. to make sure that it is understandable. Most importantly, SND staff check for any information that can relate to a living person since this affects where the data can be stored and whether or not it can be openly downloadable. As of today, SND have a list of preferred formats to use when depositing data. If data is in a common proprietary format, SND staff normally also convert the data to an archival friendly format.

A data description that is published can be updated by the researcher or DAU independently from SND and a new version is created. The changes that a researcher makes is assessed by the DAU, but must follow SND policies and the *Kravbeskrivningen* [see https://doi.org/10.34894/ZFRPU1 and PDF] ('requirements description'). If for any reason a new version of the data needs to be published, it is either the DAU or SND (depending on where the data is stored) that assesses the new meta(data) and publishes a new version of the data which also receives a new DOI. Information on the different versions can be read in SND's catalogue.

2.2 Case Study ADS: assessing information

Data can be deposited with the ADS via several methods:

- An online archive submission tool (ADS-easy).
- Grey literature reports from the OASIS system.
- Submission via physical media such as CD-ROM or USB devices.
- Submission via FTP clients/sites such as Dropbox or WeTransfer.

Regardless of the manner of deposition, all data deposited with the ADS are the subject of detailed evaluation in terms of scope, completeness, and quality. At the ADS, this is undertaken **prior to formal accession** (i.e. the act of accepting data). This is a distinct stage of ingest known as assessment and appraisal.

The assessment and appraisal event is undertaken by a member of the ADS Digital Archives team. The member of staff runs through a checklist to ensure that:

- Data deposit contains no malware (all files).
- Digital objects are in correct formats (all files).
- Data deposit has collection-level metadata.
- All digital objects have core descriptive metadata (all files).
- Digital objects have additional technical metadata (all files).
- Digital objects can be opened, are valid, and can be reused (all files, 10% sample for large datasets).
- The data deposit has no sensitive data concerns (all files, 10% sample for large datasets).
- Content is appropriate and complete (all files, 10% sample for large datasets).

Usually, these checks are made for *all* files submitted. However, for collections comprising over 1,000 images and/or text documents, the Digital Archivist only checks a representative sample of 10% of the file type.

In instances where issues are identified such as data is submitted in formats outside of the list of accepted formats or lack of metadata, the ADS will contact the Depositor with a full list of issues and required actions and ask that data is re-submitted or otherwise resolved. If the Depositor is unwilling or unable to address said issues within a set time limit, then that part of the deposit may be refused or the deposit refused entirely. Any issues and resolutions identified during assessment and appraisal are stored as documentation (usually TXT files) within the archive information package (AIP). In certain cases, the Digital Archivist may make these fixes themselves. This is a judgement call and is only made in clear-cut cases of human error and as long as it did not impact the significant properties of the object in question.

Confirmation that the repository has accepted the dataset is made via the issue of a formal deposit receipt listing the data accepted and any notes or comments. This receipt is stored within the AIP as administrative metadata. Only once the deposit receipt has been issued and the data formally accessioned has the transfer of responsibility from Depositor to repository been confirmed.

After the point of formal accession, the ADS policy is generally not to allow Depositors and Curators to make amendments to datasets. There are exceptions in this workflow however (see Repository Operations for more detail).If a Digital Archivist spots an issue with the data after appraisal and accession, but before completion and public release of the dataset, then two options are available:

1. Small corrections to metadata or data, such as typos in metadata or the inclusion of empty tables that are obviously the result of human error (e.g., a table called “test1”) may be made as an 'Editing - Corrective' event. This is performed on the AIP and dissemination information package (DIP) versions, and logged as a formal process allowing another Curator to understand the provenance.

2. Larger corrections such as significantly incomplete metadata or issues that affect data usability are passed back to the Depositor. The Depositor is asked for a new object, which is recorded as a new accession, with the older object removed according to the Deaccession and Data Disposal Policy. These actions are logged in the ADS Collections Management System. Any correspondence is archived as part of the administrative metadata in the AIP.

If a Curator or Depositor spots an issue with the data after completion and public release that requires replacement of an object with a revised to significantly corrected version, then this usually requires a new version of the archive.

1. The original dataset (submission information package (SIP), AIP and DIP) are retained.

2. The revised dataset is submitted but recorded as a new version with appropriate metadata to record provenance (particularly *when* this occurred).

There are exceptions to this rule, which, as with unreleased data, are usually simple corrective procedures or redaction of information under sensitivity concerns or UK General Data Protection Regulation (GDPR). In these cases, the DIP and AIP versions of the object are amended as an 'Editing - Corrective' event. This is performed on the AIP and DIP versions and logged as a formal process allowing another Curator to understand the provenance. It is a core policy that after accession, the SIP cannot be modified by the Curator or Depositor, only deaccessioned according to the procedure described above.

2.3 Case Study DANS: assessing information

Datasets can be deposited at DANS via the following routes:

- The Depositor directly deposits their dataset via the discipline-specific Data Station Archaeology, entering metadata and uploading the files themselves. When the Depositor is done, they submit the dataset for review by the DANS Data Processing Team.
- A 'front office' is established, for example at a university, where the data is collected and undergoes local assessment before being submitted to DANS via a general account. These datasets will still be reviewed by the DANS Data Processing Team as well.
- A machine-to-machine connection is set up in order to send datasets directly into the DANS archive. These datasets are directly published and are not reviewed by the DANS Data Processing Team. DANS will discuss the dataset quality and evaluate pilot dataset submissions while setting up a machine-to-machine connection and will gradually monitor samples of incoming datasets.

DANS allows Depositors and Curators to both make amendments to datasets (including data and metadata), as described below, and will record the identity of the person making the change in its provenance records. Changes can only be made to dissemination copies of the dataset (including data and metadata). These are managed in accordance with 'Data Management' provisions and will result in new versions of the dataset.

The SIP cannot be modified by the Depositor, and can only be modified in exceptional circumstances by the Curator. The AIP can be amended from time to time by the Curators pursuant to bit-level and format preservation, and cannot be altered by the Depositor.

With the exception of machine-to-machine deposits (see above), every dataset submitted to DANS will undergo basic curation according to the Data Stations Policy and an internal Data Processing Team manual. Basic curation includes the following verifications:

- Verification of validity of the dataset if the dataset is the result of scientific research.
- Verification of the presence of privacy-sensitive data, both in the files and in the metadata. The Depositor is ultimately responsible for the correct handling of personal data, but the Data Manager will check and may give advice to the Depositor. For example, if the dataset includes personal data in files with open access, the Data Manager may double-check with the Depositor if the settings are as desired.
- Verification of completeness of the dataset with regard to both the data files deposited and the accompanying documentation files.
- Verification of the description of the dataset for completeness, accuracy, and understandability. The Data Manager may improve the metadata before publishing the dataset.
- Verification of the readability of the files.
- Verification of the validity of the files, for example, concept versions, temporary software, or system files should not be included in the dataset.
- Verification of the file format. The verification is performed on the basis of a list of preferred file formats, guidelines for offering the best long-term guarantees for usability, accessibility, and sustainability of data files. If files are deposited in formats other than preferred formats, the Data Manager may consult with the Depositor if preferred formats can be added to the dataset, or, depending on the tools available to the Data Manager, opt for enhanced curation. In case of the latter, converted files in the preferred formats are added to a new version of the dataset. The original submission will be published as the first version of the dataset. DANS will encourage as much use of preferred formats as possible, but ultimately all formats will be

accepted. However, DANS cannot guarantee the long-term sustainability of non-preferred file formats.

- Verification of the clarity of the directory structure. If this structure is not sufficiently clear, it will be adjusted.

In the case of large datasets, it may not be possible to check the readability and validity of every single file. A representative selection of files may be checked instead, with the Data Manager using their own insight to assess what sample is satisfactory. Larger datasets are often large because they come with many images which can all be scanned in a thumbnail display.

2.4 Commonalities and differences

In each case study, the repository takes a leading role in ensuring that data meets a minimum standard, with a shared emphasis on ensuring that metadata is present and that it adequately describes *all* the data objects deposited. A possible divergence is on sensitive data, with DANS and the ADS including appraisal within their checks. Another notable difference is that for large collections, the ADS adopts a 10% sampling strategy for *some* aspects, such as the aforementioned scan for sensitive data issues. Whilst reducing the time taken for assessment, this increases the risk that problematic data moves through ingest, and thereafter may only be detected during the creation of the AIP and DIP and after publication.

Aside from what is being assessed, the differences mainly relate to how this assessment is made and the paths to issue resolution. For DANS and SND, the online management system/portal for deposition is used as both the means of delivery and assessment. In the case of SND, the use of DORIS to facilitate and record a conversation with themselves as the repository, the data producers *and* the DAU. While the ADS follows a similar workflow of checks (often performed locally) and issue resolution to SND, this is currently undertaken outside of any one system and the decisions and issues that go into the final SIP are documented within the AIP itself as administrative metadata. The decision (and ability) to capture such conversations within or outside a management system is perhaps moot as the process itself is documented in either case. It is however for repositories to consider how these decisions, which go a long way towards dictating what is in the SIP, are documented, archived, and ultimately reflected within the final AIP/DIP accessed by data consumers. In simple terms, repositories should help an end user understand how a dataset came to be in a format and what was potentially *omitted or changed*.

It is also interesting to note the different approaches to issue resolution between the case studies. The ADS approach places the onus on the data producer to meet repository requirements, with Digital Archivists only making corrections or amendments on rare occasions. Through the hierarchy implemented in DORIS, these requirements and issue resolution are perhaps better overseen through the assistance of the DAUs resulting in a more iterative process. In slight contrast to the ADS, DANS staff do take a more proactive approach to dealing with smaller issues within datasets, such as the editing of metadata and directory structures.

3. Deaccessioning data

A universal consideration for all repositories is the removal of data. A repository may be asked to remove, or themselves identify the need to remove, data for a myriad of reasons including but not restricted to:

- Intellectual property rights.
- Legal requirements and proven violations.
- Research integrity.

- Privacy/confidentiality/ethical concerns.

COPTR Workflow: Deaccessioning data

It is essential that a repository has a procedure in place to deal with cases where users can identify issues or otherwise request that data is removed. The repository must ensure that procedures are in place to ensure a clear responsibility for decision making, that staff understand the protocol for removing data, and that this event is documented.

In the current age of open access data publication and citation, the removal of data that may have formed the basis of research or decision making is not something to be approached lightly. Links to removed objects or collections should still resolve to an appropriate page, ideally with the provenance recorded (i.e., that the material has been removed and at what time). These workflows and procedures form an integral part of being a 'trusted repository' and ensure that any privacy or confidentiality concerns can be reasonably addressed, but within a framework where such events are adequately documented and understood by users.

3.1 Case Study SND: deaccessioning data

De-publication of data can be made on request by the researcher, the DAU, or by SND. The reasons for de-publication should always be documented in DORIS. SND follows DataCite's recommendations regarding de-publication and a so-called tombstone page for data is saved. It is always SND that manually removes the data. If data is stored at SND then they make the decision, otherwise the university is responsible. (See an example of a deaccessioned dataset from SND).

3.2 Case Study ADS: deaccessioning data

The ADS makes a distinction between disposal of non-accessioned data, for example a deposit that stalled or was otherwise cancelled, and the formal removal of data that has been accessioned. In most cases, accessioned data has been issued a DOI and potentially cited, but variations do occur. The following scenarios are detailed in the workflows deposited in COPTR:

- Removal of data from the SIP (deaccession) only: There are some circumstances where the accession of data has been completed, but before work on the archive is completed and the archive publicly released, data must be deaccessioned. Such requests are typically instigated by the Depositor, although in rare circumstances, the need for deaccession may be the result of a lack of funding to carry out the preservation work or identification post-accession of a serious issue within the archive that has not been responded to by the Depositor.
- Removal of individual files from the AIP and DIP: This workflow covers scenarios where data has been archived and publicly released, but a request to remove single objects has been received, including individual reports deposited via OASIS. Requests may come from the original Depositor or a user who wishes to raise a complaint about the data itself, for example, breach of copyright.
- Complete removal of a released dataset: The deaccession of a collection following the completion of preservation activities and publication of the dataset is rare and only carried out in exceptional circumstances. In these circumstances, the process of deaccession is a more complex problem necessitating updates to archive web interfaces, DOIs, and any catalogues where the dataset is exposed.

In all cases, a request to remove files from an ADS archive will be made in writing to the ADS, clearly stating the reason for said removal. The Collections Development Manager will issue a holding response acknowledging receipt of the request. This response will also outline that repository staff will conduct an appraisal of the record and the associated dataset

and that a response will be forthcoming. A Digital Archivist will conduct an appraisal of the request and report the likely impact to the Collections Development Manager. Where the Depositor(s) agree with the repository's findings, an agreement will be sought, and a plan arranged on the best approach to address the issue.

In some instances, the Collections Development Manager may decide that consultation with identifiable stakeholders who have a legal or ethical interest in the dataset is required. This may include:

- Any data creators responsible for the creation of the dataset.
- Any funders of the research who require it to be preserved in perpetuity.
- Any organisation(s) under whose auspices the research was carried out and who may continue to hold responsibility for the dataset.
- Any other individual who is setting out the requirement for deposition or preservation.

These discussions will commence prior to any data removal or formal deaccession.

It is ADS policy to ensure that anyone that has citations to resources that have subsequently been removed are retained, and furthermore, that the provenance of the deaccession or removal is recorded within all associated metadata.

Persistent identifiers (i.e., DOIs) for any removed content will remain in place, and will in no case be removed. The DOIs will resolve to a landing page of the original metadata, with a further description making it clear that the dataset has been withdrawn. The ADS will update all required internal metadata to indicate that the resource has been removed. This will be replicated in DataCite metadata and any external aggregators or metadata services.

3.3 Case Study DANS: deaccessioning data

Datasets deposited in the Data Station are potentially cited either in the original research or in derived research. In both cases, the datasets should remain available for validation (reproducibility) considerations. This means that datasets are only deaccessioned (removed from publication and open availability) or deleted in special circumstances.

Datasets can be deaccessioned when:

- Errors are discovered in a published dataset that renders it unusable for research. In such cases, DANS will endeavour to publish a corrected successor dataset in collaboration with the Depositor.
- If the content is unlawful or the dataset is fraudulent from a scientific point of view.
- If one or more of the authors or rights holders did not give permission for publication.
- If, after contact with the Depositor, personal data appears to have been deposited without a legal ground such as permission of research subjects.
- If the content consists of personal data and a research subject rightfully objects to the preservation of a digital object with an appeal to the GDPR (e.g., right to be forgotten; revocation of informed consent).

Datasets and deposits can be deleted:

- If there is a legally binding maximum preservation period for the content.
- In cases where there are other compelling reasons, to be decided by the director of DANS.

In all cases where access to a published dataset is terminated, a notice will be added to the landing page of the PID associated with the dataset to indicate that the dataset is no longer available.

If there are sufficient grounds to decide to deaccession a dataset, the deaccessioning is done by a Data Manager. Dataverse enables the Data Manager to select one or more versions of a dataset to deaccession. A reason needs to be stated for the deaccessioning, which can be selected from a drop-down list of valid reasons (e.g., 'Legal issue or Data Usage Agreement') or otherwise described.

The DOI will always lead to the most recent available version of a dataset. If one or more versions of the dataset are deaccessioned, it will be clear that there are deaccessioned versions on the versions tab of the dataset, along with the reason given for deaccessioning. The user will not be able to view or access the deaccessioned version at all. If all versions of a dataset are deaccessioned, the DOI will lead to a tombstone page informing them that the dataset was deaccessioned, with the reason given.

3.4 Commonalities and differences

Although each repository has a similar policy, primarily the ability to react to notification of an infringement of copyright or GDPR, there are some interesting differences. The DANS policy clearly states that removal may occur where “content is unlawful or the dataset is *fraudulent from a scientific point of view*”. In the case of the ADS, the reasons for removing data are not explicitly stated, only that any issue (presumably including a complaint against a fraudulent dataset) is made to the Collections Development Manager. The ADS workflow is also highly varied. Differing scenarios on whether an object or collection has to be removed create differences in how this process is handled within the team and thereafter documented. This is also complicated by the theoretical requirement to other parties that may have required preservation of a dataset as a condition of funding or other legal consideration. This potentially adds a significant time overhead to the ADS as various steps and procedures are worked through. This has the advantage of ensuring that removal of data is documented and transparent, but represents a hidden cost that is often borne by the repository themselves.

The core similarity is that each organisation maintains the PID for the removed dataset.

4. Dealing with Access databases

COPTR Workflow: Preserving Access databases

A database is a collection of data items and links between them, structured in a way that allows it to be accessed by a number of different applications programs. The most common forms of databases used in archaeology are flat file and relational databases, although there is a growing movement towards the use of object-oriented database models. The ADS Guides to Good Practice state:

In flat file databases there can be an inherent looseness in the way that data is defined and recorded along with a significant duplication of sets of information from record to record. The relational model addresses these and other issues by requiring a data structure to be pre-defined and by splitting related groups of attributes into separate tables which are then linked together through key fields (Primary or Foreign keys). In contrast to spreadsheets and many flat file databases, most database applications allow (and in fact require) the strict specification – in terms of field length, data type (numeric, etc.) of the data types to be recorded. Databases can potentially consist of more than just data values. Forms, used for data entry or for running queries, are often the only way in which many users interact with databases and can be viewed as part of the database but separate from the data itself. Likewise, the queries and results or reports that result from user interaction may also be considered as 'non-data' components of a database” (Archaeology Data Service 2023).

The preferred format requirements of all the repositories in this case study, particularly the ADS, often place the onus on the data producer to supply the data within a database in a non-proprietary format such as series of delimited text files. That being said, depositing in proprietary formats is still common. A recent study by DANS of deposits between 2000 and 2020 has shown that Microsoft Access databases are still deposited in relatively large numbers, with nearly 300 deposited as MDB files in 2020 (DANS 2022). Of interest is the presence (but relatively lower numbers) of databases deposited in the newer ACCDB format. The low use of this newer files format is attributed to cultural attachment to models of templates created in earlier formats as well as a more general trend to use spreadsheets (DANS 2022). The same trend is present at the ADS, with 1,190 Access databases (1137 MDB and 53 ACCDB) deposited since 1996. In contrast, SND holds 350 databases exported from the Intrasis system. The reliance of data producers on a proprietary format such as Access, particularly the older version of MDB, poses a significant risk when considering the recommendations for database preservation (DPC 2021 3).

4.1 Case Study SND: dealing with Access databases

Microsoft Access databases deposited to SND and stored in SND CARE are stored in their original form. The data is shared as MDB files, however the data is also stored and shared as separate CSV files and XML files. Data is exported from the Access database via a self-developed Python script. The script was made in Python 2.7 and converts, checks the format, renames files, defines projections in the GIS files, and creates and names new folders. Together with the csv files, a schema.ini file is provided with information regarding the column's name and size (remnant from the mdb export to csv). When the format is upgraded, a new version with information is published. Access databases stored locally at a university are managed by that university in accordance with local archival policies.

4.2 Case Study ADS: dealing with Access databases

In deposits, the ADS accepts the following versions of Microsoft Access:

- Microsoft Access 97 and above (MDB files)
- Microsoft Access 2007 onwards (ACCDB files). It is also worth noting that, although this format is the default in Access 2007 and 2010, files created in Access 2010 may not be completely compatible with Access 2007. The format, as with previous MDB files, continues to be based on the JET Database Engine.

In the assessment phase, the Digital Archivist will check that the following metadata is present for any Access database deposited:

- Table Documentation
    - Table name
    - Table description
    - Primary keys
    - Foreign keys
    - Row count
- Field Documentation
    - Field name
    - Field description
    - Field data type
    - Field length

- Supporting Documentation
    - Entity relationship diagram
    - Supporting documentation: explanation and/or definition of codes used, units of measurement used in specific fields if not already defined in descriptions.

The Digital Archivist will:

- Check for the presence of forms or Structured Query Language (SQL) statements. Functionality such as forms is not retained in the AIP or DIP, however these are left in the SIP.
- Check for orphaned tables and records or empty or unrelated tables. Depending on extent, these may be removed in AIP and DIP versions, or if unclear as to the purpose, a query raised with the Depositor.
- Check that referential integrity has been enforced for related tables. The Digital Archivist will run queries looking for duplicates and orphan records and highlight any issues with the Depositor. Where controlled vocabularies have been used to complete fields, the Digital Archivist will make sure that the vocabularies are indeed controlled! If the database is using online vocabularies, the Digital Archivist will make sure the Uniform Resource Identifiers (URIs) resolve!
- Check tables for duplicated rows.
- Check text fields where the length of the data is the same as the field length - it may indicate truncated values.

Following assessment, data can then be accessioned and the original Access database is stored within the SIP.

The next stage is the creation of the AIP and DIP, which comprises normalisation to CSV format. The ADS use two main workflows:

1. For MDB files, the ADS has a Java Package that exports all tables as CSV with “” text qualifiers.
2. For ACCDB files, the Digital Archivist manually exports each table to CSV with “” text qualifiers.

After creation of the CSV files, the Digital Archivist will:

- Check that all tables have been exported.
- Check that row counts after export match what is in the metadata.
- Check text fields where the length of the data is the same as the field length – it may indicate truncated values.
- Check for embedded newlines, tabs, and quotes as these may corrupt exported delimited text files.
- Check that any special characters have been preserved. Scan text fields for CP1252 characters (ASCII values between 128 and 161 - smart quotes, some accented characters, em dashes, etc). These are not preserved in CSV files that use ANSI encoding. If present, the Digital Archivist will ensure the file encoding to UNICODE UTF-8.
- Scan text fields for characters beyond ASCII 167. If present, then accented or other characters exist (as used in French, Gaelic, Ancient Greek), and the Code Page or

language of the original data must be determined. ANSI files preserve these characters, but it is worth recording that they exist within the documentation.

Where possible, CSV files should retain the same name as the original database with the table name appended. However, it may be necessary to change the table names. Where possible, CSV files should retain the same name as the original database with the table name appended. However, it may be necessary to change the table names in the case of obvious spelling errors or cases where the table name is vague or cryptic. For example 'desrciptions' (sic), 'table1', or 'mypottb'.Any changes to database or table names within the preservation or dissemination versions should be recorded in the process metadata as 'Editing - Corrective' events.

In all cases, the CSV versions will also form the basis of the DIP, made available through the ADS website. In specific cases, the original Access database will also be made available alongside the CSV versions. The file is made available under a separate statement warning the user that this is a proprietary format, which raises issues over long-term support. Cases such as these are rare, and usually occur when the Depositor has flagged items such as the forms or SQL as being useful to the target audience. The ADS will host these upon request, but recommend that SQL statements or other forms of 'content' are best included within a piece of supporting documentation supplied with the database that helps users with specific queries.

4.3 Case Study DANS: dealing with Access databases

The preservation package (AIP in OAIS-RM terms) may differ from the SIP with regards to format as files may have been converted by DANS to guarantee long-term readability by humans and machines. DANS maintains a list of preferred file formats and will ensure that these reflect the needs of the designated community and the requirements of current technology. When DANS performs file format conversions, a new version is created and the original submission will be retained.

DANS guarantees the long-term sustainability of data in formats which DANS lists as preferred formats. Whenever DANS makes changes to the guidelines to the effect that a preferred format becomes a non-preferred format, DANS is responsible for the migration of all archival files in non-preferred formats to preferred formats. DANS accepts non-preferred formats, but cannot guarantee their long-term sustainability nor is it responsible for keeping these formats sustainable. DANS will ensure that the preservation packages retain integrity through measures aimed at preventing bit-level information loss. DANS will evaluate formats from time to time and make adjustments to preservation packages to guarantee continued readability by humans and machines. When outdated formats are migrated to successor formats, the archival metadata is updated accordingly. Files in outdated formats are preserved to maintain the chain of provenance.

In practice, the non-preferred file formats for databases (MDB, ACCDB) created by Microsoft Access are widely used. However, the MDB and ACCDB formats are very poorly supported outside of Microsoft Access, which is proprietary. Due to the different versions of these formats, it is possible that different versions of Microsoft Access itself do not always support the files properly.

For the time being, for many databases created with Microsoft Access, DANS has provided sustainable and accessible formats by storing the tables from the databases as separate CSV text files. Tables can be exported from Access via an Access-form. This is a form within an Access database which is available via the DANS website. The download includes a readme file with information about the use of this tool. The form can be used to export all tables within the database to standardised, preferred CSV files. The form can be copied and pasted into another Access database to use for exporting tables from that database. It is also possible to import an external data table or spreadsheet (such as an Excel spreadsheet or a DBF file) into this database, then export that table to CSV.

Storage of the tables as CSV files only retains the tabular data from a database. Any overarching documentation needs to be described in a separate document accompanying the CSV files. Within Microsoft Access, the “Database documentation” function can be used to generate a document with column descriptions and table relationships. This document can be saved as PDF/A as formatted text and supplied with the tables of the database. In addition, care must be taken that all codes and variables used are explained, which can also be done by providing further descriptions in a separate document (“codebook”).

4.4 Commonalities and differences

It is notable that out of all possible formats and flavours of text, each repository uses CSV as a format for preservation and dissemination. At the ADS, the decision to use CSV (instead of Tab Separated Values (TSV) or even TXT files with Tab or Pipe delimiters) is driven by user familiarity (both internal and external), that most standard software applications such as Microsoft Excel can open CSV with minimal human intervention, and the fact that CSVs do not require additional metadata to document the delimiter. That being said, it is interesting to note that SND and DANS also provide the original Access database and SND additionally provides an XML as an alternative machine-readable format. This potentially offers a greater variety of re-use scenarios, subject to the support for MDB/ACCDB being in place. This highlights a dichotomy within the AIP: on one hand, the concept of a disposable/shorter-term format that can still provide a role in facilitating access and re-use, and on the other, the resolution to break the entity down to its component parts that require an end-user to rebuild. As noted in the introduction, both solutions are equally valid in presenting data to users, but it is arguable that a user of SND may have less steps to rebuilding a database than a user of the ADS. Conversely, the potential overheads for providing access to a single format is lower than providing multiple. It would be interesting to measure this 'cost' against the benefit of reuse - for example, are databases more likely to be downloaded if available in a variety of formats?

5. Discussion

The three case studies in this paper demonstrate that while some technical and methodological aspects of digital preservation in archaeology may vary, our problems and goals are reassuringly the same. Indeed, the differences in methods and approaches documented herein are related more to the in-house technical platforms and solutions implemented over the historical development of the organisations than to any conceptual divergence, with the three repositories unified by a singleness of purpose (c.f. Ahmad *et al.* 2023). While some variance is due to the context of the organisation and the role which they play within a national framework for digital data in their regions, many of the decisions share a common theme of fulfilling and balancing the needs of both the designated community, the archive itself, and end users. The continued sustainability of all three approaches is in contrast to any hypothetical concern over diversity in the implementation of digital preservation solutions.

A key theme within each case study is the importance of the ability of each repository to adapt tools and solutions that fit their needs. It is notable that none of the case studies uses a third party digital preservation system, but have instead developed solutions in house. This is particularly important when looking at the aspect of actually “doing” active digital preservation as exemplified by the case studies on dealing with Access databases. Each organisation has developed a home-grown solution for normalising files into CSV, with the primary objective to get data out by whatever means works. As demonstrated, there is no need for overly developed workflows or technical solutions when a relatively low-tech solution fulfils the objective. For a nascent or developing digital archive, it should, we hope, be reassuring that digital preservation does not need to be time consuming to implement or always require a third party system (Rieger *et al*. 2022).

Conversely, it is significant that all repositories have invested time and resources in developing an ingest and accession mechanism for datasets. Each system has been designed to both simplify the act of depositing an archive - with all the considerations of file formats and metadata that entails - but also to facilitate and expedite the archival process. As the ADS - at the time of writing - still accepts data by other mechanisms including physical media, it would be interesting to examine in more detail the cost/benefit of an online deposit mechanism versus other means so as to identify what core services are essential for a digital archive. For example, if a certain level of data (in terms of deposits but also size of data) is expected, is an online deposition tool therefore *essential* for success? If so, this would have ramifications for any new service that may be assessing requirements for their digital preservation system. That being said, there are still different viable scales and methods of ingest application: DORIS and the Data Station arguably represent more developed and streamlined workflows of curation, whereas ADS-easy presents a mechanism for transfer with the in-depth review coming after the event rather than in the system itself.

Streamlined or not, each case study requires human resources to appraise and assess data, task that can require a certain level of familiarity and knowledge of the format, but also of the content itself. It could be argued that for smaller organisations with less staff resources, this approach may not be sustainable. However, as demonstrated by the case studies, we would suggest that simply having a policy that reflects what you are able to do is more important than a policy which aspires to the unreachable. To quote an old adage, one should cut one's coat according to one's cloth. Less does not necessarily mean a lesser archive. On this note, it should not be overlooked that policies and workflows themselves are cultural products and relics of people's concerns and priorities, which themselves may be based on their work with the designated communities in their regions. We should remind ourselves that this is fine, but also that reflecting and sharing can be good for re-evaluating our own ways of working.

Bibliography

Ahmad, R., Rafiq, M., and Arif, M. 2023 'Global trends in digital preservation: Outsourcing versus in-house practices', *Journal of Librarianship and Information Science* **56**(4), 1114-1125. https://doi.org/10.1177/09610006231173461

Archaeology Data Service 2023 *Databases and spreadsheets: A guide to good practice*, Zenodo. https://doi.org/10.5281/zenodo.7740647

CoreTrustSeal Requirements 2023-2025 (V01.00), Zenodo. https://doi.org/10.5281/zenodo.7051012

Currie, A. and Kilbride, W. 2021 'FAIR Forever? Accountabilities and responsibilities in the preservation of research data', *International Journal of Data Curation* **16**(1). https://doi.org/10.2218/ijdc.v16i1.768

Cushing, A., Burchmore, T., Conroy, S., Doyle, P., Hegarty, N., Kelly, R., Kufeldt, P., McGann, M., Ormond, C., Quine, G., Reba, M. and Woods, R. 2022 'How do users discover digital preservation tools? Report on a survey of professionals' in *Proceedings of the 18th International Conference on Digital Preservation*, Glasgow, Scotland. 53-60. http://doi.org/10.7207/ipres2022-proceedings

DANS 2022 *Monitoring van bestandsformaten 4: formaten in gebruik bij Data Archiving and Networked Services (DANS)* https://kia.pleio.nl/groups/view/4fc4e83a-f55b-4000-b1cb-3fe9a16d3f93/kennisplatform-preservation/blog/view/b92f7a1e-fd6a-4c88-875c-5bccc470554c/monitoring-van-bestandsformaten-formaten-in-gebruik-bij-data-archiving-and-networked-services-dans (Last accessed 21/09/2023).

DPC 2015 *Digital Preservation Handbook*, 2nd edition. https://www.dpconline.org/handbook (Last accessed 21/09/2023).

DPC 2021 *Preserving databases. Data types series,* DPC technology watch guidance note. http://doi.org/10.7207/twgn21-06

Hollander, H. 2021 'Digital Dutch archaeology: Future perspectives', *Internet Archaeology* **58**. https://doi.org/10.11141/ia.58.28

Jakobsson, U. 2021 'Digital archaeological archiving in Sweden: the Swedish National Data Service perspective', *Internet Archaeology* **58**. https://doi.org/10.11141/ia.58.18

Lavoie, B. 2014 *The Open Archival Information System (OAIS) reference model: Introductory guide*, 2nd Edition, DPC Technology Watch Report 14-02. http://dx.doi.org/10.7207/twr14-02

Mita, A. 2015 'Community Owned digital Preservation Tool Registry (COPTR)', *Technical Services Quarterly* **33**(3), 332-333. https://doi.org/10.1080/07317131.2016.1156969

Novák, D., Oniszczuk, A. and Gumbert, B. 2023 'Digital archaeological archiving policies and practice in Europe: the EAC call for action', *Internet Archaeology* **63**. https://doi.org/10.11141/ia.63.7

Richards, J.D., Jakobsson, U., Novák, D., Štular, B. and Wright, H. 2021 'Digital archiving in archaeology: The state of the art. Introduction', *Internet Archaeology* **58**. https://doi.org/10.11141/ia.58.23

Rieger, O., Schonfeld, R.C. and Sweeney, L. 2022 'The effectiveness and durability of digital preservation and curation systems', Ithaka S+R [web], 19 July 2022. https://doi.org/10.18665/sr.316990