ADS Ingest Manual

Archaeology Data Service

November 2022

Document Control Grid

Title:	ADS Ingest Manual
File name:	ADS_ingest_manual_ v7
Location:	Internal Drive
Status:	LIVE
Version:	7
Last updated:	19/11/2022
Created date:	01/08/2004
Review due:	01/09/2023
Authors:	Kieron Niven, Olivia Foster, Tim Evans Historic Contribution by Ray Moore
Maintained by:	Tim Evans
Required Action:	None

Glossary

ADS	Archaeology Data Service
CTS	CoreTrustSeal
DPC	Digital Preservation Coalition
DPC RAM	Digital Preservation Coalition Rapid Assessment Model
Metadata	Descriptive information about data
OAIS	Open Archival Information System
SIP	Submission Information Package

1. Purpose of this document

This document outlines the general process used to ingest data submitted to the Archaeology Data Service (see Figure 1). Ingest is a deliberately broad term that covers multiple discrete workflows, from negotiation with the Depositor (i.e. data producer), through assessment of delivered data, and then to a formal act of accession which results in a defined SIP.

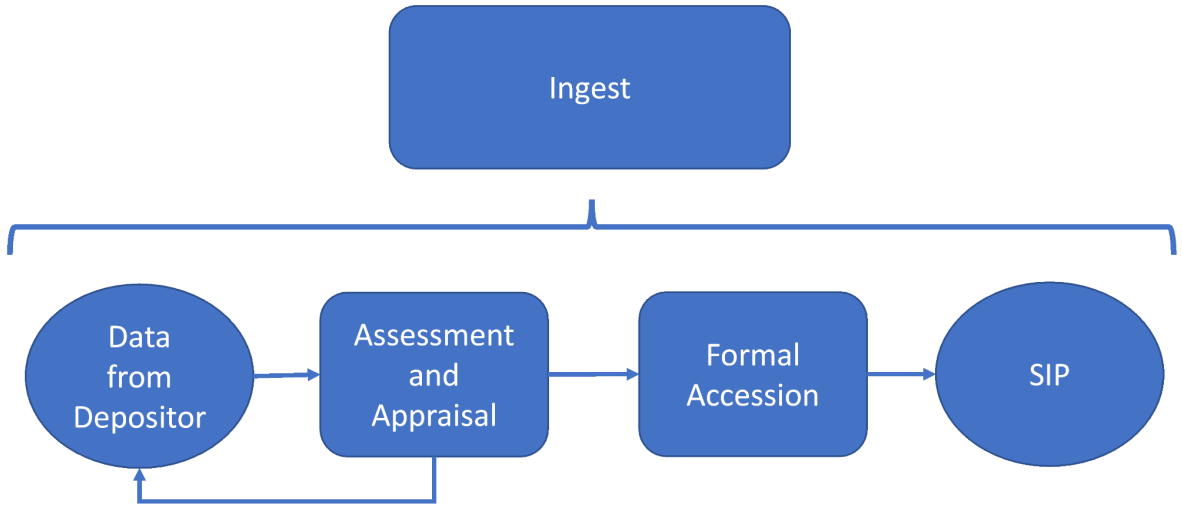

Figure 1: the broad framework of the 'Ingest' phase at the ADS.

From the OAIS:

"Ingest is the set of processes responsible for accepting information submitted by Producers and preparing it for inclusion in the archival store. Specific functions performed by Ingest includes receipt of information transferred to the OAIS by a Producer; validation that the information received is uncorrupted and complete; transformation of the submitted information into a form suitable for storage and management within the archival system; extraction and/or creation of descriptive metadata to support the OAIS's search and retrieval tools and finding aids; and transfer of the submitted information and its associated metadata to the archival store. In short, the Ingest function serves as the OAIS's external interface with Producers, managing the entire process of accepting custody of submitted information and preparing it for archival retention."

It is not intended to reflect every single micro-process and procedure undertaken by the ADS, but to give ADS staff and Data Producers a clear definition of key events, stages and responsibilities. This includes information on the workflows for submissions from ADS-easy, OASIS and other, external digital exchange services. Alongside these, the ADS continues to accept data through the exchange of physical media. Whilst

ingest is broadly similar for each of these discrete workflows, there may be subtle differences which are documented below. Once accessioning into the repository is complete, the workflow for all submissions is broadly consistent.

The Ingest manual outlines all steps undertaken as part of this phase. A separate policy for **Assessment and Appraisal** covers the rationale, and more precise steps undertaken in that particular workflow.

2 Overview of pre-Accession workflows

2.1 ADS receives data

The ADS receives data from depositors through a series of submission streams that allow the exchange of data, alongside technical and contextual metadata. Depositors use one of the following repository deposition streams:

- OASIS: an online form for the reporting of archaeological projects in the UK, and maintained by the ADS on behalf of national heritage agencies. Data producers can upload unpublished reports (but no other data) for long-term preservation and access with the ADS. Metadata for files is captured within the form.
- ADS-easy: an online application for the transfer of data to the ADS for long-term preservation and access. ADs-easy uses FTP to allow Data producers to transfer most accepted data types and requisite metadata. ADS-easy is synchronised with the OASIS system, allowing re-use of relevant metadata and signposting between relevant outputs and archives.
- External data exchange service: These include, but are not restricted to, the University of York DropOff (file sharing) service, alongside other commercial file sharing services DropBox, GoogleDrive, etc. The ADS also utilises SFTP transfers with depositors where necessary.
- Exchange of digital media: including CD-ROM, DVD, USB or portable hard-drives in person, or through the postal service. Older forms of media (e.g. 8-inch, 5¼-inch, and 3½-inch floppy disks) can be deposited only after consultation with the ADS to ensure media can be read using current tools.
- Physical retrieval: in very rare cases ADS staff are authorised to move data from in-situ machines (e.g. a PC in an office building) as part of arrangements to 'rescue' data.

Where data is supplied manually (external data exchange service, exchange of digital media, Physical retrieval), data is moved to a designated space within the ADS

systems demarcated as non-accessioned holdings. In singular cases, such as a very large dataset, a local file store may be used for ease of access.

For the **ADS-easy** and **OASIS** applications, files uploaded by the external user are held on a separate VM or section of the ADS NFS, where they reside until formally accepted via accession (see below). Again, it should be clearly noted that data uploaded to ADS-easy and OASIS **has not been** formally accessioned, and thus cannot be considered preserved until the full creation of an AIP.

For the sake of consistency in this and other ADS documents, all of these spaces named above are referred to as **Temporary Holdings**.

2.2 ADS Assess Data

Any data received by the ADS will be subject to an assessment by a member of the Curatorial and Technical Team. This is detailed in a separate **Assessment and Appraisal document.** The aim of this phase is to ensure that the deposit meets the requirements of the ADS. The ADS website provides detailed instructions for preparation and submission of data through the **Instructions for Depositors**. This includes specific information on accepted formats, collection-level documentation, and file-level metadata requirements. The ADS also provides general advice for depositors through the following resources:

- Data Management and Sharing Plans
- Guidance on the Selection of Material for Deposit and Archive.
- Digitising Journal Articles and Grey Literature Reports.
- Policy and Guidance on the Deposition of Personal, Confidential and Sensitive Data.
- Archaeology Data Service / Digital Antiquity Guides to Good Practice.

The major criteria checked during this stage of ingest are:

File formats are suitable for deposition

All files adhere to the **Instructions for Depositors**.

Files are virus free

Deposits made through ADS-easy and OASIS applications are submitted to virus checking *during* the upload process. Digital deposits outside the above submission streams, and those accessions involving the exchange of physical media, are subject to the checks outlined in the **Security Overview**.

Files can be opened

All media and data submitted to the ADS are scrutinised to ensure that it has not become corrupted prior to or during transmission. Checks also seek to identify password protection or encryption that adversely affect preservation and dissemination pathways.

Files are accompanied by full and appropriate metadata

Each file, no matter what its source, should have the same minimum level of metadata that allows it to be preserved, found, and re-used. File-level metadata requirements are documented in the **Instructions for Depositors**.

Collections are described using full and appropriate metadata

Each Collection of data, no matter what its source, should have the same minimum level of metadata that allows it to be preserved, found, and re-used. Collection-level metadata requirements are documented in the **Instructions for Depositors**.

Sensitive data

Data meets ADS standards for deposition of sensitive data.

Data is deposited with a Licence permitting preservation and access

It is the responsibility of depositors, as per the terms of the deposit licence, to ensure that they have the 'rights', as data owner or copyright holder, to submit the dataset to the repository. Similarly, it is the responsibility of the depositor to ensure that the submission adheres to current legal and ethical guidelines. The repository does provide additional guidance on the deposition of personal, confidential, and sensitive data allowing depositors to plan for the submission of data.16 In those instances where repository staff identify data that they believe infringes the licence or policies, these concerns are directed towards the depositor and, where necessary, the Collections Development Manager. A clear record of any such discussions is maintained within both the preserved dataset and the ADS' Collection Management System (CMS) ensuring clear documentation of any issues.

Outcome

The outcome of this phase is either:

- Move to formal accession and SIP

- Deposit is rejected, and the Depositor is provided with a list of issues that need to be addressed.

3 Overview of Accession workflows

3.1 ADS record administrative metadata in Collections Management System (CMS)

Once assessment of the SIP is complete, repository staff document the accession using the CMS.10 This records the date of accession and depositor; the creation of a digital checklist is included to track the collection as it moves through the ADS workflow.

3.3 Standardise file names and directory structure

Generally, depositors should follow the file naming guidance provided in the Instructions for Depositors.

In the case of depositions submitted via OASIS and ADS-easy stringent programmatic controls are in place to ensure that depositors adhere to the file naming policy.
For other digital depositions outside of these submission streams, or involving the exchange of physical media, manual checks ensure adherence to the file naming policy.

Repository staff ensure that all files have been checked and, where required, updated in accordance with the file naming policy. In accordance with Repository Operations Manual documentation of any changes or updates within the CMS ensures a clear record of any changes to the dataset.
As noted in the Instructions for Depositors. and Repository Operation Manual the ADS does not make any specific stipulations regarding the directory structure used within the SIP. However, as the Guidelines for Depositors note, we do ask that "a logical file structure" is used that "allows data to be easily retrievable". Data that is poorly structured may make preservation activities difficult, and will certainly affect the dissemination of data and restrict reuse. Where SIPs use a poor, or overly complex, data structure, it may be necessary for repository staff to make changes to the data structure of the submitted dataset. Appropriate documentation, using the CMS, ensures a clear record of any such changes.

3.2 Data transfer and temporary storage

Once the data assessment has been complete and all necessary file name and structural changes have been made to the SIP, then it should be copied from the local drive to the preservation server. All file data transfers use a dedicated client and follow the SFTP protocol.

Where the dataset has been deposited digitally using ADS-easy or all data will be stored on a dedicated server until accession; at this point data will be transferred to the preservation server during a semi-automated ingest process initiated in the CMS. These data transfers use the SFTP protocol, with checksum validation to ensure the successful transfer of files.

3.3 Record details of SIP in the CMS

Once assessment of the data deposit is complete, repository staff document the accession using the CMS. This records the date of accession and depositor; the creation of a digital checklist is included to track the collection as it moves through the ADS workflow.

Deposition of data can sometimes be a protracted process requiring negotiation with the depositor. It is important that any correspondence is preserved and attached to the collection record in the CMS.

Any correspondence pertinent to the preservation or dissemination of the dataset should also be stored alongside data. This should be stored in accordance with the structure outlined in the Repository Operations Manual.33

All file names for correspondence should follow the guidance outlined in the Repository Operation Manual.

3.4 Create a new 'version' of the dataset

In those instances where the SIP includes a new 'version' of the entire dataset then repository staff should ensure that the directory structure is as outlined in the Repository Operations Manual. This is a programmatic process initiated from the CMS. Any creation of a new version of the dataset should be correctly documented, and checked, within the CMS.

3.5 Create fixity values and file signature

The ADS uses the National Archives' file characterisation application, DROID, to create technical metadata for each file within the SIP. This includes a checksum, used to monitor the 'health' of the file within the archive, alongside file format, file format version and MIME-type that facilitate the ongoing management of the dataset. All metadata is stored within the OMS.

3.6 Create data receipt

Once the accession is complete, repository staff should acknowledge the receipt of all data by email. The email should highlight any issues, queries or concerns regarding the dataset. Repository staff should ensure the depositor is given an opportunity to address or respond to any problems, and any such responses should be clearly documented within the CMS, and emails retained within the file store in accordance with Repository Operations Manual.

A 'data receipt' should be attached to the acknowledgement email.

All emails, and the deposit receipt, should be stored in accordance with Repository Operations Manual.

3.7 Store original media

Once an accession is complete, physical media should be labelled with the correct collection number and accession id(s). Media should then be stored in the collections filing cabinets in the ADS offices.

Where the depositor has requested that original media be returned to them (this may happen if data has been delivered on a memory stick or portable hard drive) this should be returned to them. A note is added to the CMS to document that media has been returned to the depositor.

In circumstances where data has been deposited electronically through ADS streams (ADS-easy or) or external file-sharing services, these should only be deleted once data has been copied to the preservation servers.

3.8 Document completion of accession process

Once the depositor has acknowledged that the submission is 'complete' and as intended, and a 'signed' deposit licence has been returned, the accession can be 'signed off'. Repository staff should ensure that the 'accession' process is marked as complete on the internal workflow checklist.

4 Preparation of the AIP and DIP

Once accession is complete, repository staff will begin the process of normalising data, and the creation of both AIP and DIP. A dedicated checklist, available through the ADS wiki, guides digital archivists through the archiving process.19 Alongside these checklists, the ADS' Data Procedures provide information on the processes and procedures for the preservation and dissemination of discrete data types.

Only once preservation and dissemination activities are complete, and the creation of both AIP and DIP finished, will the process of AIP checking begin.

4.1 Check and assess the significant properties of files to be preserved and establish conversion plan

Repository staff should check and assess the significant properties of the dataset alongside any associated documentation and metadata.

In some circumstances previously undetected problems may arise once we start working with the data and preparing the AIP and DIP. Where issues are noted, the digital archivist should contact the depositor and, where necessary, encourage them to re-submit replacement data in order to maintain data integrity and authenticity. It may be possible for repository staff to undertake edits on behalf of the depositor, but any such changes require documentation within the CMS.

Repository staff should develop a conversion plan for all archives. The ADS' Data Procedures provide detailed guidance on appropriate formats for preservation and dissemination. In some circumstances, data may already be in formats suitable for preservation or dissemination and may not require normalisation, but in others, conversion may be more complex and require multiple steps. Therefore, it may be worthwhile mapping out each step to ensure the appropriate software and expertise is available. While the Data Procedures provide guidance on typical preservation and dissemination pathways, it may be necessary to consult with colleagues over the most

appropriate course of action for your dataset. In instances where the dataset is large, or the number of files significant, batch processing may provide a reliable and consistent approach to normalisation. Digital archivists should ensure that all normalisations are consistent with the guidance outlined in the Data Procedures.

4.2 Preserving and disseminating data

The creation of a working or 'local' version of the dataset, on which all normalisation and updates can be carried out, will ensure the integrity of the original data (AIP). All data should be normalised to suitable preservation and dissemination formats in line with the preservation plan and Data Procedures.

Repository staff should ensure that the normalisation of data, alongside any other changes, are successful and that the significant properties of each file remain unchanged. As noted above, it may not always be practical to check each individual file within a dataset, but it is important to check a representative sample of the dataset.

Once all normalisation processes are complete and the resultant files validated to ensure the preservation of all significant properties, the transferral of data from local directories to the preservation and dissemination servers can begin. All file transfers use a dedicated client, and follow the SFTP protocol. Fixity checks ensure that transfers of data have been successful. Digital archivists should ensure that data is stored in accordance with the original directory structure and as outlined in the Repository Operations Manual.

4.3 Create technical metadata

Once copied to the ADS servers the use of DROID39 allows the creation of technical metadata for all files and normalised data within both the AIP and DIP. The transferal, and storage, of all technical metadata to the OMS is a semi-automated process initiated through the CMS.

4.4 Creating and matching objects

Repository staff should use the 'match objects' functionality, initiated from through the CMS, to link all related files into notional 'objects'. This is largely programmatic, although it may be necessary to group files and objects manually. Checks should be made of all objects to ensure that they are accurate, including the creation of

'parent-child' objects where relationships between ingested and preserved files are complex.

4.5 Ensuring the correct data types have been assigned

Digital archivists should ensure that all files within the collection have been assigned a 'data type' as part of the accession process (outlined in Section 3.10), or, in the case of depositions through the ADS-easy and submission streams, created by the depositor during data upload.

At the same time, the creation of AIP and DIP may result in the creation of new files and 'objects', it may, therefore, be necessary to assign new data types to these outputs.

Repository staff should ensure the accuracy of all assigned data types. Where necessary, functionality within the CMS allows 'updates' to be made to the data type.

Other files, such as the deposit licence or deposit receipt, may have an internal and administrative function; repository staff should ensure that these objects have been assigned the correct data type. Again, 'updates' to the assigned data type can be made through the CMS.

4.6 Ensuring the correct resource types have been assigned

To facilitate the internal management of data, repository staff should also ensure that the correct 'resource type' is allocated. This allows greater granularity in characterising the content of files and objects. The CMS allows the correct resource type to be assigned.

4.7 Ensuring the correct resource types have been assigned

Digital archivists should ensure that all relationships between files, or 'objects', are documented within the OMS where the nature of that relationship, or 'relationship type', can also be expressed.

4.8 Ensuring the correct resource types have been assigned

ADS-easy In instances where data has been submitted through ADS-easy metadata is completed through a series of online forms. These services provide some quantitative checking of metadata during upload, with additional qualitative checks made as part of the accession process. The transferal of this metadata to both the CMS and OMS respectively forms part of the accession process for these submissions, however repository staff should always carry out checks to ensure accuracy and completeness.

The identification and resolution of any gaps or problems with the dataset form part of the accession process. However, in instances where the identification of new/other issues with collection or file level metadata are noted, or where extant problems have not been addressed to the satisfaction of the repository staff, it may be necessary to contact the depositor (again) and updates to the data or metadata sought. In some circumstances, repository staff may carry out updates or enhancements to metadata as part of the archiving process; any such changes require documentation within the CMS.

In circumstances where documentation has been submitted through the online forms within ADS-easy metadata will be stored within the CMS and OMS. While it is not necessary to extract any metadata from the OMS for inclusion in the AIP for preservation reasons, as standard practice all file level metadata is exported from the OMS and stored within the DIP as a separate file in a suitable format. The inclusion of this downloadable version of the file-level metadata within the archive interface provides data users with the necessary documentation to facilitate data reuse. All metadata extracted from the OMS should be stored in accordance with the Repository Operations Manual. Repository staff should ensure that the correct data type has been assigned and any relationships between metadata and data documented.

The submission of the SIP digitally (outside of ADS-easy) or through the exchange of physical media requires collection-level and file-level metadata to be submitted using the ADS' dedicated metadata templates. Collection and file level metadata submitted using these templates should be transferred to the CMS and OMS respectively. As noted above, care should be taken to follow the guidelines concerning problems or issues with data or metadata. Following transfer of all documentation from the templates to the OMS, checks ensure parity between the metadata stored in the two locations. All completed metadata templates should be stored, in accordance with the Repository Operations and disseminated in appropriate formats outlined in the Data Procedures. Repository staff should ensure that the correct data type has been assigned and any relationships between metadata and data documented.

For all types of submission, depositors may upload additional or supplemental documentation, beyond the required standard metadata, as separate, discreet files. The accession of additional metadata should follow the workflow outlined above for all depositions. Supplemental metadata is subject to the same qualitative checks carried out for all data and metadata. All supplemental metadata should form part of both the AIP and DIP in appropriate formats outlined in the Data Procedures. It should also be stored in accordance with the Repository Operations Manual. Repository staff should ensure that the correct data type has been assigned and any relationships between metadata and data documented.

4.9 Record conversion and editing processes undertaken

All conversions, or processes, carried out on the dataset should be documented in the 'Processes' section of the CMS. A semi-automated system allows repository staff to generate most of these programmatically following file matching. This process makes assumptions about the nature of the conversion, so all processes require checking to ensure accuracy. Repository staff can make edits or updates to processes through the CMS interface. In some cases, it may be necessary to add processes manually.

The CMS documents the type of process carried out, and includes a detailed record of each process. A record of any problems, or other information, should be included in the comments section of each process.

4.10 Interface creation

Following the creation and documentation of both AIP and DIP, the creation of a dedicated interface for each collection should follow. Guidance on this is available from the ADS wiki, and in consultation with the Collections Development Manager.

For 'standard' depositions the ADS uses a series of templates and examples to facilitate this process. These templates can be adapted and amended where required.

For 'special collections' a bespoke interface must be created. Agreement on the form and content of the interface is generally during negotiations for deposition, and documented within the CMS – whether it includes a database search, a map interface, 3D viewer, etc.

Once completed, whether the interface is a 'standard' deposition or 'special collection', repository staff should ensure that the interface meets the required standards in terms of accessibility, validation, and compatibility.

The CMS allows repository staff to keep a clear record of any interface functionality, queries, directories, database tables, and repository templates.

4.11 Updating associated documentation

The CMS can be used to document any 'notes' concerning the dataset and the preservation process, while it also allows the storage of any associated documentation and correspondence relating to the dataset.

Any documentation relating, specifically, to the preservation or dissemination of the dataset should also be included in the AIP. All documentation should be stored in accordance with the Repository Operations Manual and named in accordance with the File Naming Policy.

Repository staff should ensure the completion of all required collection-level metadata within the CMS.

Any correspondence or documentation pertinent to the deposition, maintenance or management of the dataset should be preserved within the CMS. These can be added as 'notes' or attached to the CMS record.

4.12 Updating associated collection metadata

Repository staff should check that the transfer of all collection-level metadata from the requisite template submitted as part of the SIP, or, where data has been deposited digitally via ADS-easy or, has been transferred to the CMS.

As noted above, typically, the identification of any problems or gaps in the metadata occurs during accession with any updated and additional metadata received from the depositors. In rare circumstances, however, repository staff may notice issues whilst working with the dataset, in such circumstances they are required to contact the depositor directly for additional information. As observed above, it may also be necessary for repository staff to augment and enhance metadata; the CMS allows the creation of a clear record of any such processes and activities.

The dissemination of the collection metadata through the 'metadata' page within the archive interface ensures the data users have clear access to the documentation for the archive.

4.13 Submit AIP for checking

On completion of both the AIP and DIP, AIP checks ensure that the outcomes meet the requirements outlined in this Ingest Manual, the Repository Operations Manual and Data Procedures. An assigned digital archivist initiates a semi-automated process of checks, alongside visual inspections of the AIP (and DIP), following completion of the preservation and dissemination procedures. Historically, all archives had AIP checks carried out, but as workload has increased, it has become impossible to carry out the assessments for all archives. In cases where the collection is relatively simple, i.e. it contains a small number of raster images and some metadata, formal AIP checks may not be necessary and the visual inspection of the archive, carried out by all repository staff, should highlight any issues. This is often the case for archives submitted through the ADS-easy or submission streams. The repository still regards AIP checking as an important part of the preservation process, particularly as it ensures the correct adherence to recommended archival procedures and policies. Instances where AIP checking may be necessary include, but are not restricted to, the following:

- where a collection is large, or particularly complex
- where the digital archivist is unsure whether preservation and dissemination activities have been completed successfully
- where the archive includes 'new' data types or formats
- where digital archivists are undergoing training, or to ensure digital archivists are consistently applying current policies and procedures
- where Collections Development Manager feel checks may be necessary

Collections requiring AIP checks are marked within the CMS, and the Collections Development Manager allocates the task to a Digital Archivist. A dedicated checklist provides guidance on the assessment carried out as part of the AIP checking processes,19 with additional information provided in this Ingest Manual, the Repository Operations Manual and the Data Procedures. Once complete, any problems or issues with the AIP and DIP are highlighted and appropriate action taken to address them. As noted above, checks of the archive interface also form part of the AIP process.

At the same time, the sharing of the archive interface with the depositor allows their input into the archive interface. Any comments or requests by the depositor should be addressed where appropriate and within reason. The formal submission of archives for "sign-off" by the Collections Development Manager follows.

4.14 Archive release

Following sign-off by the Collections Development Manager, the archive is ready for release. An agreement for the date of release is sought from the depositor.

Embargo: The repository does allow depositors to request an embargo date for those archives that contain sensitive information, or to allow full publication of any outputs. Following appraisal by the depositor, and completion of the AIP checks (where carried out), the removal of all pages and files of embargoed content from the production server ensures the protection of any associated data. In instances where an advance DOI is required, for publication say, the creation and continued support of a landing page may be necessary. Once an embargo has passed, the return of all pages and files to the production server, and the necessary updates to DOI, ensures the dataset is publicly accessible.

Detailed documentation for the release process is available in the ADS wiki, although a shorthand version is available within the Procedure Checklist. A final check is made of the archive (AIP and DIP) which ensures that all procedures have been followed, all processes documented, and all metadata completed.

Repository staff run DROID to ensure that all files within the AIP and DIP are recorded within the OMS and all files have been correctly linked together into notional 'objects' using the 'match objects' functionality. Checks should be carried out to ensure that all 'data types' have been correctly recorded. These actions ensure that an accurate record of the archive contents, and up-to-date technical metadata, form part of the documented dataset, facilitating the ongoing management of the dataset.

The 'date of release' (and associated 'ready for release') should be added to the CMS. Part of the release process involves the minting of the persistent Digital Object Identifier (DOI) for the collection. All releases are publicised through the ADS' 'Collections History' page, resource discovery metadata transferred to the ArchSearch catalogue, and the archive is included in the archive index. In instances where the archive includes formal reports or other 'library-worthy' documents, citations are also added to the ADS Library. Where the archive is 'marine based', metadata is added to the Marine Environmental Data and Information Network (MEDIN) data portal.

The release process also involves the transfer of the complete AIP to off-site, 'deep storage' where the dataset is backed up to ensure the long-term preservation of the dataset.

Once released, the repository publicises the publication of the collection through its website and social media.