Ingest Manual

Version 3.0

Created date:	2004
Last updated:	30 August 2017
Review Due:	January 2018 (unless significant change)
Authors:	Tim Evans and Ray Moore
Maintained by:	Tim Evans
Previous version:	**Ingest Manual 2.0**

Introduction

The ingest process at the ADS is a two-stage process. First of all, the data is accessioned. This should be carried out as soon as possible after receipt of the Submission Information Package (SIP). The second stage of the process, preparing the data for archive and dissemination, is more time consuming and will be scheduled in to the ADS work plan at a later date.

Accessioning the data

As a new SIP is received by the ADS it is a priority to get data stored safely on our systems and issue the data producer with an acknowledgement that the data has been received. Accessioning will be carried out at the earliest opportunity. Archival staff try to identify any potential problems with the data at this early stage, so that where more documentation is required, or if files are corrupt or contain viruses, the data producer can be informed as soon as possible.

An accessions checklist (a version of which can be viewed here) is available to guide curatorial staff through the accessioning process and should be used with reference to this Ingest Manual. This checklist is replicated within the ADS' Collections Management System (CMS) and must be completed. Only when all parts of the CMS checklist are completed can work begin on the creation of the AIP and DIP.

Data Transfer Session

Data is accepted in a variety of different file formats and can be delivered to the ADS in a number of ways (see Depositing Data at ADS). Further information on this is documented in our Guidelines for Depositors.

Virus Check

Before transfer of the data to our servers, anti-virus software is run on the SIP to ensure it is virus free. For deposits made by ADS EASY, virus checks are performed upon upload of the

file(s) to that system. If a virus is found, this must be removed either by disinfecting the relevant file, or requesting that the data producer provides a 'clean' copy. Further processing of the data files must not go ahead until the data is known to be free of viruses.

Manual checking will also highlight any files that are password protected. Data producers should be asked to re-submit copies of any files that are password protected.

Media and File Readability Check

If the SIP consists of only a small number of files these are opened to ensure they are readable and not corrupt. Unless batch processing is an option only a random sample of each file format present in a large archive will be checked at this stage. More thorough checks will be carried out in the data processing stage. The data producer should be contacted regarding any corrupt files and asked to re-submit.

Basic data validation may be carried out at the accessioning stage but it is recognised that it will not be possible to pick up on all potential issues with an SIP at this point. Fuller checks will be carried out at a later stage when the AIP and DIP are constructed. Checks will vary depending on the nature of the data and the file formats that have been submitted. Appendix 1 describes the sorts of checks that may be carried out here.

Check File Formats are Suitable for Deposit

The SIP must be checked to ensure that all files submitted are files that we can accept. Lists of suitable file formats are available in our Guidelines for Depositors and should be adhered to by data producers unless special provisions have been agreed in advance.

Where an SIP contains file formats that we are unable to deal with (for example because no suitable migration path exists, or that we do not have access to any software that can read them), the data producer must be contacted to see if they can provide the data in an alternative format. Where it is not possible to re-submit files in a suitable format, we may agree to accept them and archive on a ‘best efforts’ basis.

Documentation + Integrity Check

The SIP should be checked against any accompanying documentation such as file lists supplied by the data producer. The numbers of files, file names, formats and version should be checked here. The purpose of these initial checks is to highlight any missing files or major discrepancies between the delivered data and the documentation. Any discrepancies should be queried with the data producer.

All SIPs received by us should be accompanied by documentation. The ADS Guidelines for Depositors describe what documentation is required, and also contain links to metadata templates which can be completed and sent to us to supply the level of metadata that is required. Archives uploaded via the ADS EASY system also require metadata, and can either use the ADS EASY interface to manually insert this documentation, or should use the same templates from the Guidelines for Depositors to upload with the dataset.

Data producers should be requested to provide any missing documentation necessary for the processing and archiving of their data. Requests can be sent direct to the data producer.

As we will be creating a web interface to allow access to the data, the SIP should also include some introductory text about the project and the digital archive which can be used on the front page/s of the archive. It is also useful to have one or two images to illustrate the archive where appropriate. The data producer can be contacted at this stage and asked to supply these if necessary.

The Curatorial Officer should make a note of any files within the SIP that appear not be copyright of the data producer (or whoever signed the deposit licence). Should there be any unresolved copyright issues, the processing of the submission stops until relevant copyright clearances have been obtained. Any queries on the copyright of submitted material should go through the Collections Development Manager at the ADS.

Record Details of SIP in Collections Management System

Date of receipt of data should be recorded in the tracking module of the Collections Management System. This will allow the accessioning process to begin. All relevant details of the accession must be recorded in the Collections Management System, including details of data producer, media, file types (with versions) and quantities.

Copy to Data Server

The SIP should be copied on to our data server in the directory structure as specified within the ADS Repository Operations Documentation.

Replace Spaces with Underscores

Once this process has been carried out, the only editing of the SIP that is permissible is the cleaning of file and directory names. Guidelines on editing names are available on the ADS staff wiki, and replicated in brief within the ADS Repository Operations Documentation.

Create Checksums

Checksums must be created on the SIP and put in a text file to be stored in the Archival Information Package (AIP) under 'admin/original_accession_id.txt'.

Store Licence in AIP Directory

Work should not begin on creating an AIP until a signed deposit licence has been received.

A signed deposit licence should exist for every SIP. If a licence has been received by the ADS for this data, a scanned copy of it should be moved to the AIP directory as described in the Repository Operations documentation.

Scan Paper Documentation

Most documentation arrives in digital form. On the occasion where only paper copies are provided and we are subsequently unable to get a digital version from the data producer,

paper documentation will be scanned at an appropriate resolution (as described in the Repository Operations documentation). This should be stored in the appropriate place within the AIP directory structure.

Record SIP in the Object Management System

The archivist should use the DROID-based interface within the CMS to record the technical details of all files in the OMS.

Acknowledge Receipt of Data

A receipt or acknowledgment of the data received should be sent to the data producer. As well as noting how many files and which file types have been received in the SIP, it may also contain any queries that have arisen about the data during the accessioning process. A copy of this receipt should be stored in the AIP directory structure as described in the Repository Operations documentation.

Store Original Media

Once accessioning is complete, original media and associated paperwork are stored in the collections filing cabinets ordered by collection number. Both the CD/DVD case and any accompanying paper documentation should be annotated with the collection number and the accession number as recorded in the Collections Management System, plus the date data was received. Where the depositor has requested that original media is 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 only after the weekly backup to our deep storage facility has been carried out. A note should be made on the checklist for this accession that original media has not been retained.

Preparing the AIP and DIP for Archive and Dissemination

The process of migrating files and creating the AIP and Dissemination Information Package (DIP) is carried out by a member of the Archival Team at the ADS and involves the steps described below.

Though a certain amount of checking and validation of the delivered data will have occurred at the accessioning stage, it is often the case that previously undetected issues arise once we start preparing the AIP and DIP for archiving and dissemination.

A checklist (a version of which can be viewed here) is available to guide staff through the archiving process, and should be used with reference to this Ingest Manual. This checklist is replicated within the ADS' Collections Management System (CMS) and must be completed. Only when all parts of the CMS checklist are completed can work on the AIP be signed off.

Consistency Checks

Basic checks of data against accompanying documentation should already have taken place as part of the accessioning process, but once work on the data begins in earnest, there may be further checks that need to be carried out.

The nature and complexity of consistency checks depends on the data type under scrutiny. Databases, GIS and spreadsheets would usually require more checking than raster image files for example. Examples of the sorts of checks that may be carried out can be found in Appendix 1 of this manual.

Where problems with any of the data are highlighted, the data producer should be contacted, informed of the problem and encouraged to re-submit. In order to maintain the integrity and authenticity of the data, we would always prefer data producers to re-submit rather than carrying out edits in-house. Archivists should consult ADS guidelines on editing and updating archived data (available internally on ADS wiki) for further guidance.

Selecting Preservation and Dissemination File Formats

Refer to the Collections Management System to see which file types and versions have been deposited. Using this and the files themselves identify suitable preservation and dissemination formats for each of them. In carrying out this step it is important to consider the significant properties of each of the files and ensure that these will be preserved within the new formats. Refer to ADS Data Procedures documents (available internally on ADS wiki) to inform decision making about the most suitable preservation and dissemination file formats.

It may be that the formats the files were delivered in are suitable for preservation or dissemination with no changes. It may be that the same format is suitable for both preservation and dissemination, or it may be that two different versions of the file need to be created for the purposes of preservation and dissemination.

Develop a Conversion Plan

Once decisions have been made on appropriate file formats for preservation and dissemination, a conversion plan may need to be established. If the conversion into a preservation or dissemination format is a complex procedure, map out each of the steps to be taken and check that we have the software and expertise to carry out this work. Some conversions may be straightforward single-step operations which can be done as batch processes on large numbers of files. Other conversions may involve multi-stage processes using various different software packages.

Convert the Files

Following instructions contained within the ADS Data Procedures documents, copy and convert the deposited files as necessary until every file supplied is in the correct format and

location. Files intended for preservation or dissemination should be stored in the correct place within the directory structure as defined in the ADS Repository Operations Documentation.

Note that all work on files should be carried out on local copies of the data, not on the actual SIP that has been accessioned on to the data server.

Validate File Conversion

Ensure that the conversion of the digital resource has been successful and that the significant properties of each file remain unchanged. Though it is not practical to check each individual file, it is important to check a small sample of files against the original SIP to ensure that significant properties are maintained.

Record AIP and DIP in OMS

The archivist should use the DROID-based interface within the CMS to record the technical details of all files created for the AIP and DIP in the OMS. Further processes such as 'Match Objects' and recording of 'Parent-Child' relationships should also be undertaken.

Record Conversion and Editing Processes Undertaken

Once the conversion plan has been successfully implemented, the ‘Process’ section of the Collections Management System is updated to include details of these conversions. This will include detailed preservation metadata describing the conversion process, including the date, methodology, software and platform used.

Submit AIP for Checking

Once preparation of an archive is complete, it should be submitted for AIP checking by another member the Curatorial Team. The AIP checking will ensure that the recommended directory structure has been maintained, recommended preservation file formats have been used and required metadata and checksums are present. This double checking of an AIP ensures that all staff are working to a consistent standard and following the archival procedures as laid out in this Ingest Manual, the Repository Operations Documentation and Data Procedures documents. The AIP checklist (available to view here) is replicated within the ADS' Collections Management System (CMS) and must be completed within that system.

Once AIP checking is carried out and any problems or issues with the AIP are highlighted and corrected, the ingest process is complete. The AIP will remain unchanged until such a time as further data is accessioned in to it, or preservation or dissemination files need to be migrated into newer or more suitable formats.

An archive cannot be signed off and publicly released until the final AIP checks have been completed and signed off by the Collections Manager.

Appendix 1

Data Validation and Consistency Checks

Here are some examples of the types of checks which may be carried out as a part of the ingest process. This list is taken from the AHDS Archive Ingest Procedure Framework: HS Preservation Procedures Manual, working draft 1.3 prepared by Raivo Ruusalepp, Estonian Business Archives Ltd, December 2002/January 2003:

- Check that digital resources and their items adhere to the relevant formal definitions of their structure (e.g., an XML document conforms to its XML schema, a relational database conforms to its SQL schema, an image conforms to its stated image format – dpi, colour depth, compression, etc.).
- Image compression algorithm, dimensions, orientation, resolution, colour space, etc. correspond to the values stated in documentation.
- Digital audio compression algorithm, length of the recording, sampling frequency, bit rate, etc. correspond to the values stated in documentation.
- Digital video compression algorithm, length/duration of the recording, codec structure, frame rate, sound format, etc. correspond to the values stated in documentation.
- Linkages and dependencies between items within a particular type of digital resource should be checked for correctness (e.g., in a database, foreign keys having a matching primary key; in a spreadsheet, formulas refer to correct cells, etc.).
- Linkages and dependencies to other digital resources are correct (e.g., hyperlinks point to a currently valid URL, details of published works in a bibliography are correct, etc.).
- Items within a digital resource adhere to the relevant definition (e.g., a numeric field in a database contains a number, text strings do not exceed a stated maximum length, etc.).
- Items within a digital resource contain ‘sensible’ values that do not contradict relevant logical assumptions (e.g., age of a person should not be less than 0) and subject/resource type specific concerns.
- Documents (word processor files) should be checked for changes or errors in footnotes, tables of contents, links, auto-fields and formatting that may hinder the later use of the data resource.
- GIS, CAD and virtual reality data resources may require domain- or research area specific consistency checks to be applied (e.g., scale of different layers in a GIS, level of precision and sufficiency of co-ordinates in a CAD and VR data, etc.).
- Simple data types (numbers, text strings, dates, etc.) are not truncated, restricted in range, formatted or otherwise defined in a potentially confusing or ambiguous way (e.g., dates contain four digits for the century, date format, memo fields in a database do not contain embedded end-of-lines, etc.).

- Coded data must be checked that the data have been consistently assigned the documented code.
- Any codes that are used in data must be used consistently and according to the specified coding rules.
- Standardised data has been standardised consistently and according to specified rules or a recognised schema for the standardisation.
- Exceptions to particular standards, coding schemes, formats, etc. are documented and justified in the documentation for the data collection.