Archaeology Data Service:

Repository Operations Documentation

Author/s:	This document is based on the AHDS Repository Operations documentation (draft 1.5) by Hamish James & Gareth Knight Created by Jenny Mitcham
Maintained by:	ADS Curatorial Officers
Version:	2.0
Created date:	2006
Last updated:	January 2011
Review Due:	January 2012 (unless significant change)

Conventions

A `Source Code block` is used to highlight text that represents identifiers, directory names, file names and similar concepts.

Words or phrases in curly braces ('{' and '}') are placeholders that should be replaced with the appropriate content. For example, '{`AIP-identifier`}' indicates that a valid AIP identifier, of the form `arch-{collection number}-{edition number}` , should be inserted, where, in turn, '`{collection number}`' and '`{edition number}`' must also be replaced with actual values, while the dashes, which are not enclosed by braces, are literal text to include in the final AIP identifier.

Square brackets ('[‘ and ‘]') are used to indicate a set of choices, from which one choice should be selected. Each choice is separated from the next by a pipe character ('|'). For example {file name}.[pdf|tif] indicates a file name that should finish with either the extension ‘.pdf’ or the extension ‘.tif’.

Introduction

This document summarises the requirements for preparing collections for the ADS repository, and for their management within the repository.

This document makes considerable use of terminology and concepts from the OAIS reference model.

An OAIS is defined within the Consultative Committee for Space Data Systems document as "an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community" (2002).

There are six mandatory responsibilities that an OAIS compliant archive should meet and they are summarised below:

- Negotiate for appropriate deposits
- Obtain sufficient control of resources
- Determine scope of community
- Ensure independent utility of data
- Follow procedures for preservation
- Disseminate to the designated community

The ADS carries out all of these core functions. We can also map both staff and archival activities to the OAIS data model illustrated below.

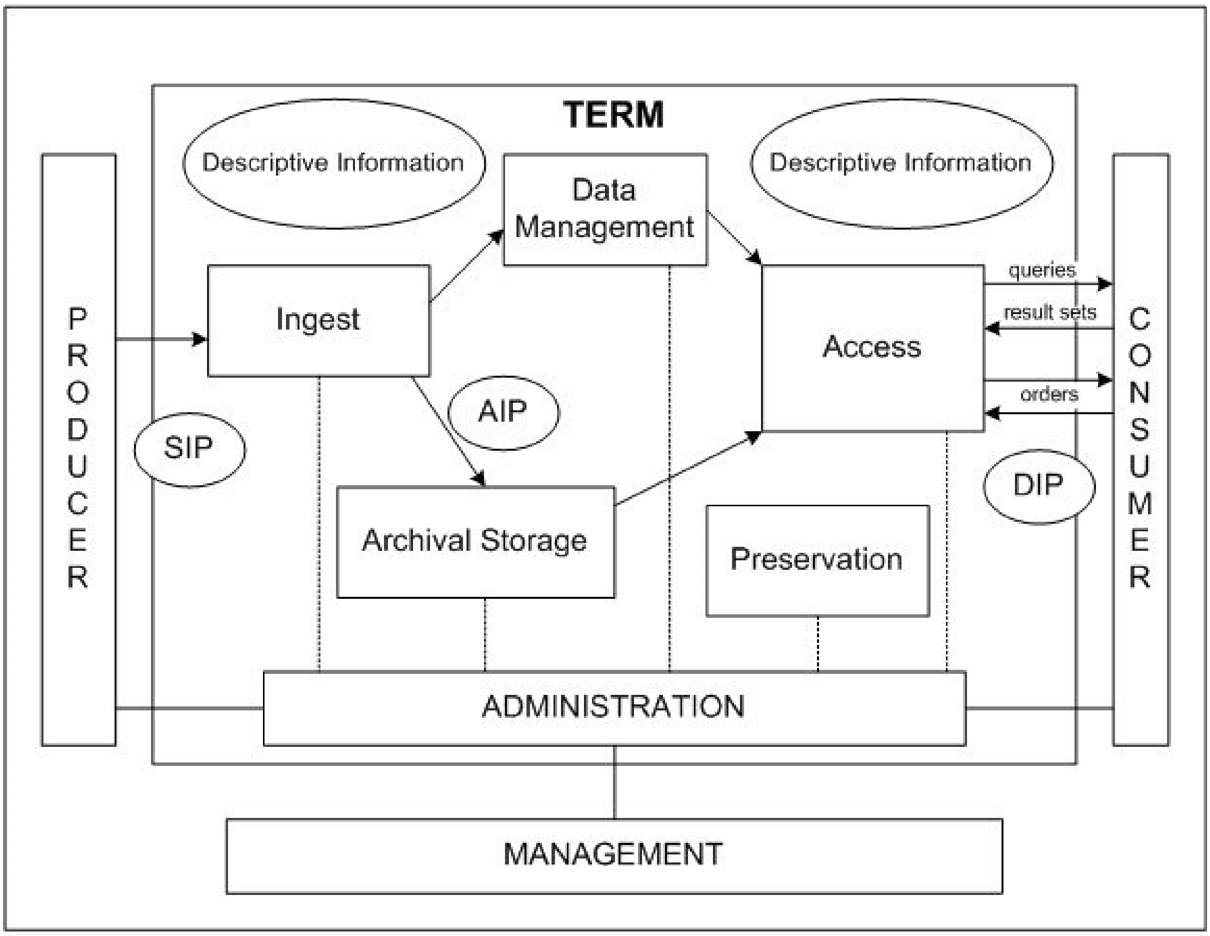


Fig. 1. Major functions of the OAIS Reference Model from Consultative Committee for Space Data Systems (CCSDS), CCSDS 650.0-W-1, Producer-Archive Interface Methodology Abstract Standard. (OAIS). White Book. Issue 1. Draft Recommendation for Space Data System Standards.

In the OAIS model, information packages move from producers through the OAIS and on to the data consumers. In the ADS, depositors send us data deposits (submission information packages), and each deposit is then accessioned into a collection and archived (as an archival information package). An online interface is created for this collection and in this way it can be disseminated and accessed by our users (as a dissemination information package).

Submission Information Package (SIP)

SIP is the OAIS term for a deposit of data. The ADS will accept SIPs that fit the requirements given in our:

- Collections Policy (http://archaeologydataservice.ac.uk/advice/collectionsPolicy)
- Guidelines for depositors (http://archaeologydataservice.ac.uk/advice/guidelinesForDepositors)

Archival Information Package (AIP)

AIP is the OAIS term for a coherent set of information that will be archived. SIPs accepted by the ADS will be added to an AIP that conforms to the requirements outlined in this document.

An AIP consists of electronic files containing the data, documentation, metadata and administrative material (scanned licence, checksums, correspondence etc.) for a collection.

The deposited data itself represents what the OAIS model defines as the content information, which is the actual material being preserved. Also a part of the AIP is the Preservation Description Information (PDI), which is the administrative metadata used to plan and manage the preservation of the content information.

The OAIS model allows AIPs to contain other AIPs. An ADS AIP may contain other AIPs where previous editions have been replaced with a new edition.

AIP Directory Naming

Each AIP must have a persistent identifier, taking the form:

```
arch-{collections number}-{edition number}
```

The collections number will be a unique identifier for the collection. This number will be automatically generated by the ADS Collections Management System (CMS).

The first edition of a collection should be given the edition number 1, and subsequent editions should be numbered sequentially (2,3,4 etc.). Editions must be numbered using positive integers. An ADS AIP may hold more than one edition of a collection. All previous editions of a collection should be stored in the `/previous/` directory under their AIP identifier.

AIP Directory Structure

Conceptually, the OAIS model describes an AIP as having two main parts, the content information and the preservation description information (PDI). For the ADS, these two parts of an AIP are both stored as files in a single directory tree. The AIP directory tree is a logical rather than physical structure, and its contents may be spread across multiple storage devices (for example the preservation data may be in a separate location to the dissemination). Keeping a consistent and tight directory structure allows us to store parts of an AIP in different locations without confusion. See appendix 2 for further information on this.

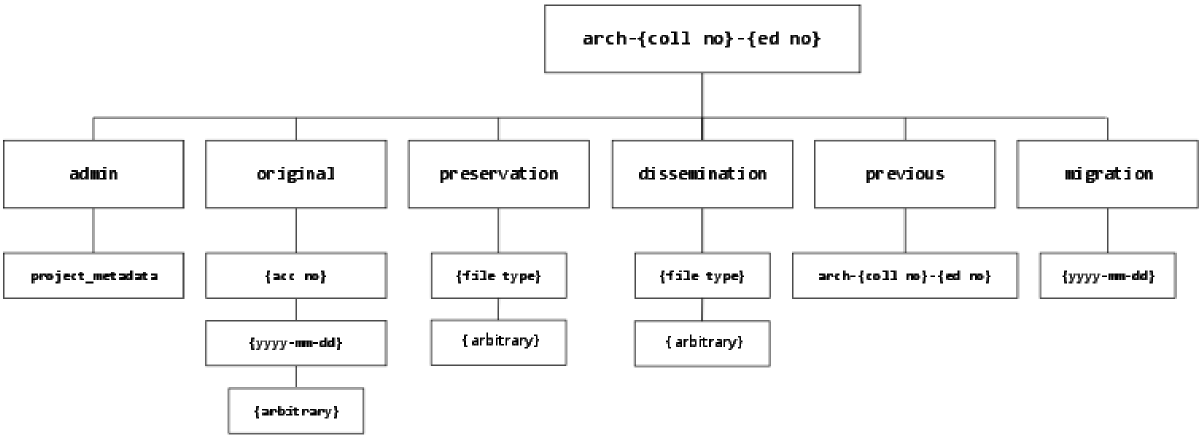


Figure 2: Logical directory structure for a collection

The 7 main subdirectories

The AIP directory structure has seven main subdirectories (summarised below). These neatly organise the material in an AIP according to its purpose and origin.

- **original** – Material received from the depositor as the SIP is stored here.
- **admin** – Administrative material, notably metadata about the contents of the AIP, the scanned licence agreement, correspondence and checksums are held here.
- **preservation** – The data directory for holding preservation files.
- **dissemination** – The data directory for holding dissemination files.
- **previous** – This directory can be used to hold previous editions of the AIP.
- **migration** – This directory can be used to hold old versions of files which have been migrated into newer formats.

These 6 directories must be named as listed above (in lower case).

This level of the AIP directory must not contain any other files or directories.

For further information about the content and structure of these directories see the relevant section below.

original

The `/original/` directory should contain the following:

All files received from the depositor and accepted as part of the final deposit

The /original/ directory should be structured as follows:

A subdirectory named by accession number must be created under /original/ to hold data from each individual accession (for example /770/)
A subdirectory named by date must be used under the accession number directory to distinguish tranches of material received from the depositor at different times under one accession. This directory should be named as follows: `yyyy-mm-dd` (for example `/2008-04-23/`).
The subdirectories of the SIP should be maintained in the structure they were delivered in. Other logical subdirectories may be created by ADS staff, for example where an SIP is delivered on multiple CDs, the subdirectories `/cd1/`, `/cd2/`, `/cd3/` may be used.
With the exception of replacing any spaces in file and directory names with underscores, changes to the names of directories or files provided by the depositor are discouraged. If other changes are necessary, they must be recorded in the Process section of the CMS. Guidelines on editing and updating archived data are available on the ADS staff wiki.

admin

The `/admin/` folder must contain the following files (see appendix 4 regarding reserved file names):

File name	Description	Comments
`licence.tif`	A licence form scanned and saved as uncompressed tif	Scanned licences must meet the Requirements for Scanned Hardcopy Material (appendix 3)
`checksum.txt`	A single text file containing checksums for all files within the AIP directory	Note: this may no longer be a requirement once the file level metadata database is fully populated and maintained
`{collection identifier}.xml`	A valid, METS compliant XML document holding CMF2 metadata for a collection.	
`{collection`	A valid XML document	

identifier}_gpmd.xml	containing the Generic Preservation Metadata (Process metadata) for a collection	
licence addendum	Only applicable where a deposit licence has been qualified	

The /admin/ folder may additionally contain the following items which should be stored in a directory called /project_metadata/:

File name	Description	Comments
dc_metadata.txt	A text file containing the Dublin Core metadata record that has been loaded into ArchSearch	This document may contain one or more Dublin Core metadata records
email_{yyyy-mm-dd}.txt	Copy of an e-mail sent to, or received from the data producer	Store only those mails which: - help document the SIP - give us permission to remove or edit files in the SIP - have been sent with original data to be accessioned into our archive - clear up copyright issues surrounding any of the data Any other e-mails can simply be printed and stored in the yellow folder
Other metadata files	Other files (normally supplied as part of the SIP) that help document or describe the project as a whole. They should be named as named within the SIP, or if created by ADS, in a logical and meaningful way. Do not store any files that are duplicates of data	If it relates to files of a particular file type, the /documentation/ directory will sit under the individual file type directory. For example a listing of captions for a set of jpg images should sit in /jpg/documentation/, or documentation for a database should sit in /csv/documentation/.

	and documentation that is already held in the /preservation/ directory.	
Introduction and Overview text	Introduction and Overview text for the web interface should be stored here	Where this information has been gleaned from other documents within the AIP, there is no need to store it separately here. As long as it is stored somewhere within the AIP that is adequate
`DepositReceipt.[docx|txt]`	Digital copy of the deposit receipt sent to depositor after accessioning of data.	This may either have been in e-mail form (store as txt file) or as a printed Word document (store as docx)

Note: it is important that files in the `/admin/` directory are in suitable preservation file formats

Data directories

The AIP Directory folder contains the following directories for holding the data component of the archive:

`/preservation/`	All files allocated for preservation
`/dissemination/`	All files allocated for dissemination

These 2 directories may contain the following subdirectories:

`/{file type}/`	Using these directories we can store files of the same file types together. You may have a directory named `/tif/`, `/jpg/`, `/pdf/` or `/csv/` for storing files of those types. Further examples can be found in appendix 1
`/{file type}/{arbitrary}/`	Other arbitrary sub-directory structures are permitted under the `{file type}` directory. Sub-directories used in the original deposit will often add context to the data and should be replicated within the two data directories. Other logical structures may be created by the ADS in order to help order the files within the online interface.

/{file type}/{arbitrary}/documentation/	This sub-directory under the file type or arbitrary directories is for documentation specific to the different file types. For example a listing of captions for a set of jpg images should sit in /jpg/documentation/, or documentation for a database should sit in /csv/documentation/.

These directories should not contain:

Thumbnail, preview or other images that we have created specifically for display within the web interface. These should be held in an /images/ directory within the relevant web directory itself. They should be seen as part of the web interface rather than a part of the archive.
Oracle loading files (applicable for special collection databases and monument inventory databases).

previous

The /previous/ directory should contain:

Files that were part of the AIP, but have now been replaced by a new edition. Files should be stored in /previous/ under a subdirectory named by the old AIP identifier (arch-{collections number}-{edition number}). Underneath this directory, a logical directory structure must be used based on the directories that the files have been moved from – this will help us see where these files have originated from. See example 3 for further clarification on usage.

The /previous/ directory should not contain:

Any files that represent part of the current state of the AIP

migration

The /migration/ directory should contain:

Files created by the ADS that were once part of the AIP but are now considered obsolete, having been replaced by newly migrated files in more up-to-date file formats. Files should be stored in /migration/ under a subdirectory named by the date that the files were replaced or moved. Underneath this date directory, a logical directory structure should be used based on the directories that the files have been

moved from – this will help us see where these files have originated from. See example 4 for further clarification on usage.

The /migration/ directory should not contain:

Any files that represent part of the current state of the AIP

Note: it may not be necessary (or possible) to keep copies of all files in /migration/ forever. Policy on this will be reviewed once migration strategy within the ADS has been more formally tested.

Bibliography

Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1 Blue Book

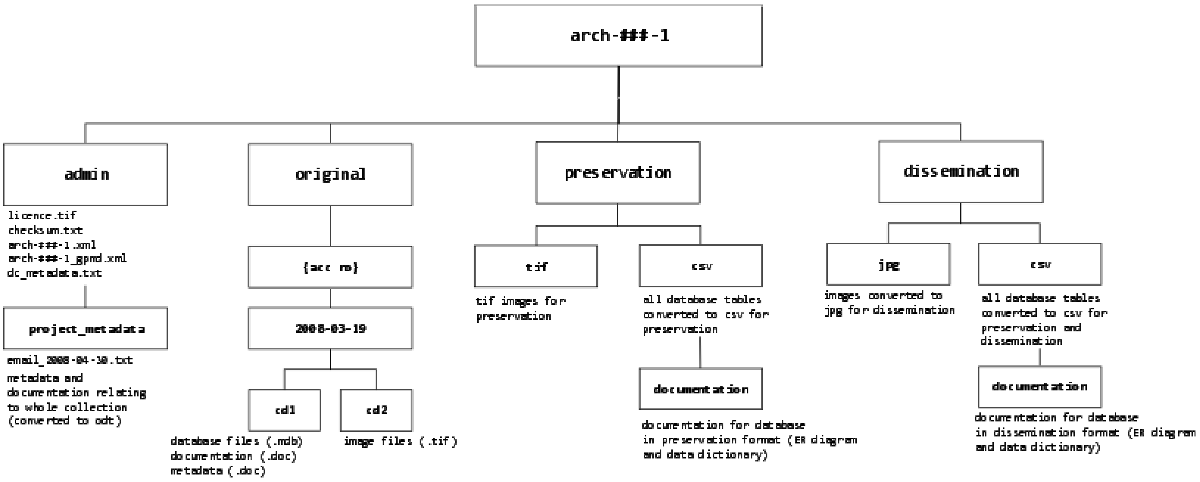
Examples of ADS directory structure

The following examples are intended to illustrate a number of different archiving scenarios showing how we should manage and organise the data within our directory structure:

Example 1: a 'simple' one-off deposit

In this example, a database with documentation and metadata and associated images were deposited on 2 cds and accessioned on the 19th March 2008.

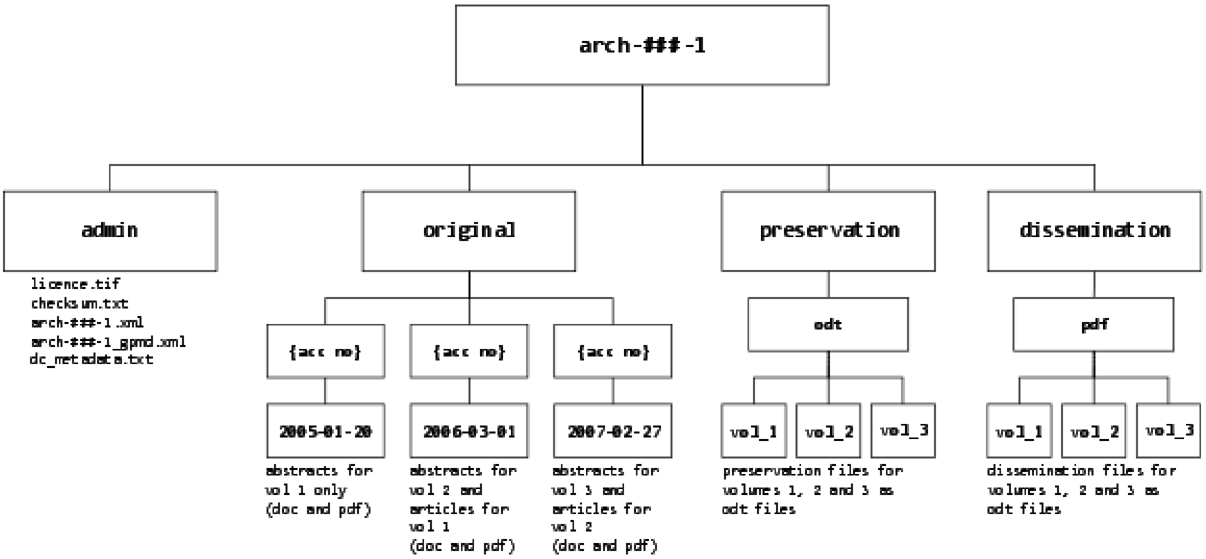
- In an e-mail sent on 30th April 2008 the depositor clarified some queries about the database and gave permission to ignore one of the data tables that contained only test data so this e-mail has been stored in `/admin/project_metadata/`. Documentation for the collection as a whole is stored in the `/admin/project_metadata/` directory as well and may include a completed ADS project metadata template
- Database tables have been saved as csv files under `/preservation/csv/` and `/dissemination/csv/`. As CSV is both a preservation and dissemination format the files are stored in the `/preservation/` folder and duplicated in `/dissemination/`.
- Likewise, the data dictionary and entity relationship diagram for the database are stored in the `/documentation/` directory under `/preservation/csv/`. These documents, where relevant, should be in suitable formats for the preservation.
- Tif images have been stored for preservation purposes under `/preservation/tif/` and batch converted to jpg images for dissemination (stored under `/dissemination/jpg/`)


arch-###-1
admin
licence.tif
checksum.txt
arch-###-1.xml
arch-###-1_gpmd.xml
dc_metadata.txt
project_metadata
email_2008-04-30.txt
metadata and documentation relating to whole collection (converted to odt)
original
{acc no}
2008-03-19
cd1
database files (.mdb)
documentation (.doc)
metadata (.doc)
cd2
image files (.tif)
preservation
tif
tif images for preservation
csv
all database tables converted to csv for preservation
documentation
documentation for database in preservation format (ER diagram and data dictionary)
dissemination
jpg
images converted to jpg for dissemination
csv
all database tables converted to csv for preservation and dissemination
documentation
documentation for database in dissemination format (ER diagram and data dictionary)

Example 2: a run of journals (we receive new data on a yearly basis)

In this example, MS Word documents and pdf files have been deposited for 3 volumes of a journal over the course of 3 years.

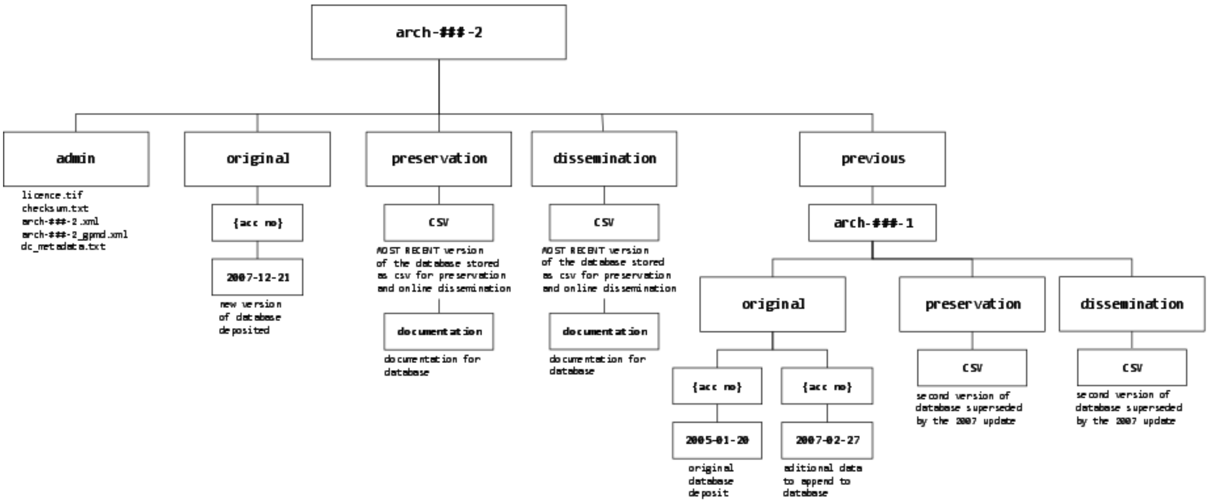
- 3 directories in `/original/` store the 3 deposits of data
- Files are converted to preservation format and stored in `/preservation/` under a file type directory followed by a volume number directory. Splitting files into separate volumes helps maintain order and structure
- PDF files are stored in `/dissemination/pdf/vol_{no}/` for dissemination online
- The usual files are stored in `/admin/`. Note the `checksum.txt` and the xml files will need to be regenerated each time a new data is added to this AIP


arch-###-1
admin
licence.tif
checksum.txt
arch-###-1.xml
arch-###-1_gpmd.xml
dc_metadata.txt
original
{acc no}
{acc no}
{acc no}
2005-01-20
2006-03-01
2007-02-27
abstracts for vol 1 only (doc and pdf)
abstracts for vol 2 and articles for vol 1 (doc and pdf)
abstracts for vol 3 and articles for vol 2 (doc and pdf)
preservation
odt
vol_1
vol_2
vol_3
preservation files for volumes 1, 2 and 3 as odt files
dissemination
pdf
vol_1
vol_2
vol_3
dissemination files for volumes 1, 2 and 3 as odt files

Example 3: an update to a database (showing use of `/previous/` directory)

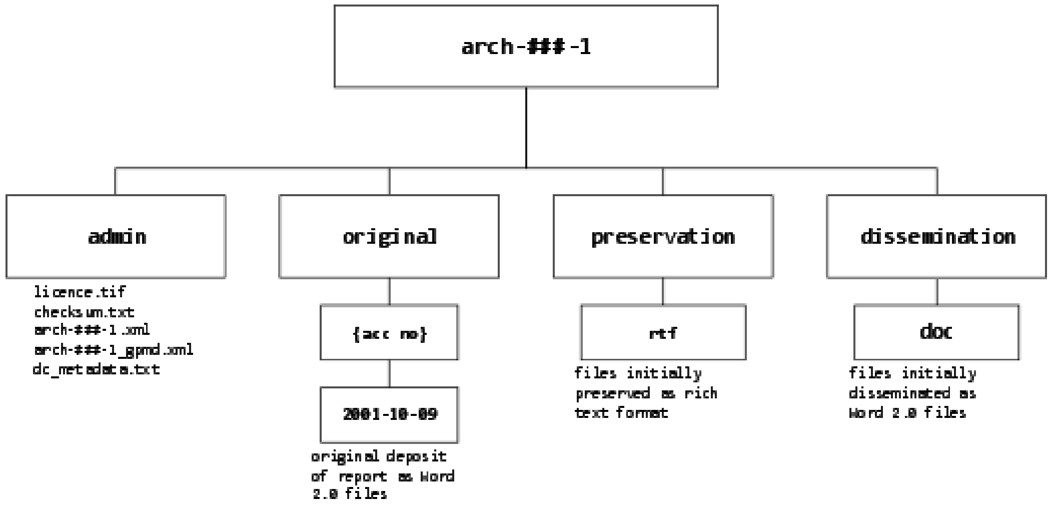
In this example, a database (and documentation) has been deposited in 2003 and has been made available as a series of csv downloads. Subsequently in 2005, additional records were deposited to append to the database and were accessioned separately. Then in late 2007 the final database containing updates and amendments was accessioned – this constituted a new edition of the archive

- 1st deposit: database is saved as csv files for preservation and dissemination and database documentation is stored in a `/documentation/` directory under the `/csv/` directory
- 2nd deposit: the additional records in the 2005 deposit are incorporated into the existing csv files and stored under `/preservation/` and `/dissemination/` as before
- 3rd deposit: the new and complete version of the database is converted to csv and stored under `/preservation/` and `/dissemination/` with the original documentation (still applicable as the data structure has not changed).
- The whole AIP is renamed to reflect the fact that it is now the second edition.
- The old original and csv files (and the directory structure that they sit within) from the 2003 and 2005 updated version are moved into the `/previous/` directory, stored under the original AIP name (edition 1)
- The `/admin/` directory remains unchanged though new checksums and xml files will be generated and stored as work is carried out on the AIP

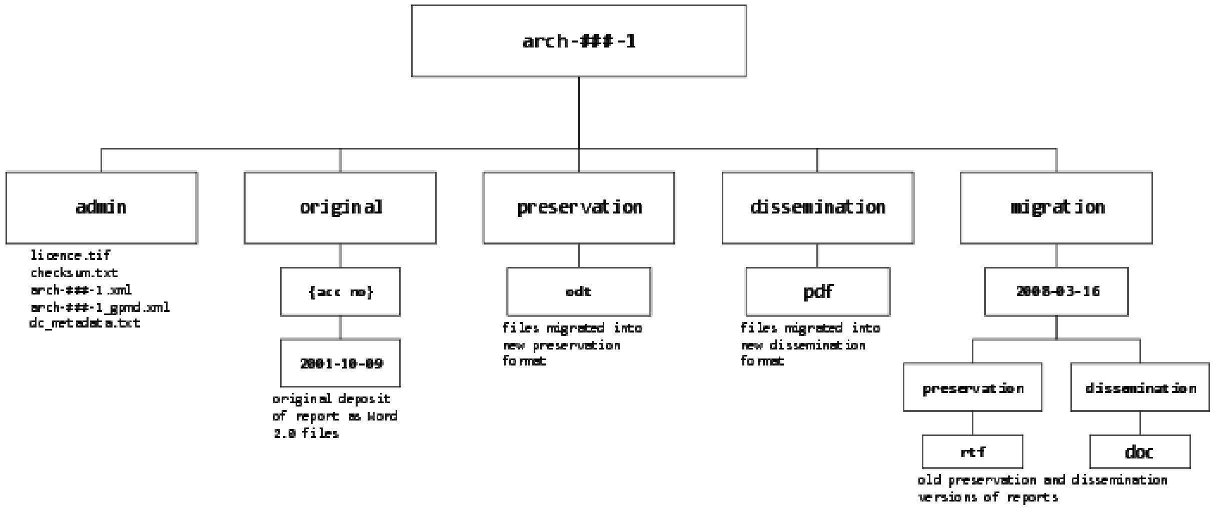

arch-###-2
admin
licence.tif
checksum.txt
arch-###-2.xml
arch-###-2_gpmd.xml
dc_metadata.txt
original
{acc no}
2007-12-21
new version of database deposited
preservation
csv
MOST RECENT version of the database stored as csv for preservation and online dissemination
documentation
documentation for database
dissemination
csv
MOST RECENT version of the database stored as csv for preservation and online dissemination
documentation
documentation for database
previous
arch-###-1
original
{acc no}
2005-01-20
original database deposit
{acc no}
2007-02-27
aditional data to append to database
preservation
csv
second version of database superseded by the 2007 update
dissemination
csv
second version of database superseded by the 2007 update

Example 4: what happens when we carry out a file migration

In this example a selection of archaeological reports were accessioned in 2001. They were submitted as Microsoft Word 2.0 files and we disseminated them in their original format and preserved them as Rich Text Format.



In 2008 we decided that we needed to migrate both the preservation and the dissemination files into newer formats. Here is how the resulting directory structure would look.

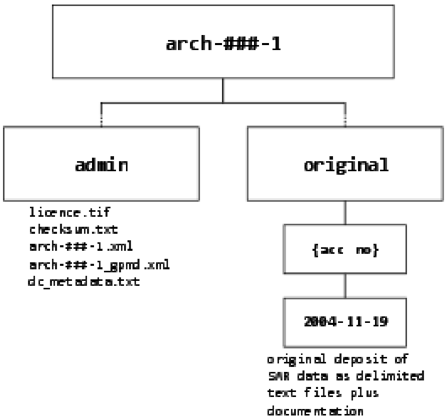


- Old doc and rtf versions of the files are moved into `/migration/` (stored under the date that the files were moved, and with the directory structure that shows the purpose and origin of the files)
- New odt and pdf versions of the files are created for preservation and dissemination and stored under the relevant data directories
- Metadata in the CMS is updated to reflect the changes and new xml files are generated to store in `/admin/` directory. New checksums are also generated for the whole AIP

Example 5: a single deposit of an SMR database

In this example we have received a database from a Sites and Monuments Record. They have signed a licence which is different to our standard deposit licence and have agreed that the data is for dissemination through ArchSearch alone. The ADS therefore have no commitment to preserve this data.

- Beyond preparing the data and loading into ArchSearch no action is needed
- Data files used to load the database tables into Oracle need not be stored here
- `/preservation/` and `/dissemination/` data directories are not necessary
- Metadata including GPMD must be created in the CMS and stored in `/admin/` directory in the usual way



Appendix 1: Example file type directories under data directories

Underneath the `/preservation/` and `/dissemination/` directories should be a list of sub directories related to the file type/extensions. Straight forward file types will simply be organized according to extensions (ie. pdf, csv). For more unique or non-standard data types, they should be organized in a sensible manner (ie. geophys). Examples are as follows:

`/dissemination/` directory:

pdf	
zip	for files such as GIS shapefiles that are disseminated zipped
jpg	

`/preservation/` directory:

gml	gis files converted to GML
tfw	image files with world files
txt	only to be used for plain text files, not for delimited data
pdfa	archival pdf files
tif	tif images
mpeg	this should include any mpeg file versions. Can further subdivide data into `/audio/` and `/video/` directories within this folder
docx	
dxf	
geophys	for all geophysics data. Can further subdivide data into `/mag/`, `/res/` and `/gpr/` for magnetometery, resistivity data and ground penetrating radar

Appendix 2: Where to store files

The ADS currently has a split AIP, with preservation files being held in /ADS_preservation/ and dissemination files in /adsdata/.

In order to avoid a situation where we are duplicating things, follow these guidelines when constructing directories in these two separate areas:

/ADS_preservation/ **should contain**:

- The /admin/ folder and all of its contents
- The /original/ folder and all of its contents
- The /preservation/ folder and all of its contents
- The /previous/ folder (where it exists) and all of its contents
- The /migration/ folder (where it exists) and all of its contents
- The only situation where the /dissemination/ directory may appear here are where there are dissemination files which are not available for download on-line (because of delayed release or large file size for example)

/adsdata/ **should contain**:

- The /dissemination/ folder and all of its contents

If we need to combine the AIPs into a single directory structure it will be easy to do so if these guidelines have been followed.

As a general rule, **avoid creating empty directories**. If there is nothing in a directory then it does not need to exist.

Appendix 3: Requirements for Scanned Hardcopy Material

The licence form, hardcopy documentation and other supporting information must be scanned at, or above these minimum standards:

Black and white documents, 200dpi 1 bit.
Greyscale documents, 200dpi 8 bit.
Colour documents, 200dpi 24bit RGB.

Scanned images must be saved as TIFF v6.0 (pref) or PDF/A. Image dimensions should be adjusted according to the size of the scanned paper (A4, legal etc.). Illegible images should be rescanned at a higher resolution and/or colour depth.

If scanned images are saved as TIFF, a separate file must be created for each page scanned rather than a single multi-page TIF file.

Appendix 4: Reserved File Names

A number of file names (such as licence.~[tif~|pdf~] and checksum.txt) are reserved for particular purposes. These file names must be used where appropriate.

Reserved file names can be modified. Modifications should be placed after the main name of the file, and separated from it by a dash.

A reserved file name should be modified to include sequential numbers if there is a need to save the information in multiple files. For example example, under /admin/, a three page hardcopy deposit licence scanned and saved as a series of TIF images could be named as follows: licence-1.tif, licence-2.tif, licence-3.tif.

A reserved file name may be further modified to include a date, in ISO 8601 format, if there is a need to distinguish between two versions of a file created on different dates that are both still current in some way (for example licence-2004-03-01.pdf) If both a date and a sequential number modifier are used, then files MUST be named as follows (using the example from above): licence-2004-03-01-1.tif, licence-2004-03-01-2.tif, licence-2004-03-01-3.tif.