



AHDS Archive Ingest Procedures Framework

AHDS Preservation Procedures Manual

Final Version

prepared by

Raivo Ruusalepp
Estonian Business Archives, Ltd.

March 2003

Table of Contents

1. INTRODUCTION	2
2. SCOPE	2
3. STRUCTURE	3
4. DETAILED PROCEDURES GUIDE	4
4.1 PRELIMINARY OR PRE-ACCESSION STAGE	6
4.2 ACCESSION / SUBMISSION	9
4.2.1 Data transfer session	9
4.2.2 Update metadata	9
4.2.3 Virus check	9
4.2.4 Media and file readability check	9
4.2.5 Data resource completeness / integrity check	10
4.2.6 Documentation completeness check	10
4.2.7 Copy to processing area	10
4.2.8 Authenticate original version	11
Output	11
4.3 ADMINISTRATION AND METADATA MANAGEMENT	12
4.3.1 Check all forms	12
4.3.2 Check copyright and confidentiality clearance	12
4.3.3 Update metadata databases	12
4.3.4 Scan paper documentation	13
Output	13
4.4 QUALITY ASSURANCE	14
4.4.1 Review data and prepare validation	14
4.4.2 Consistency checks	14
4.4.3 Metadata update	17
Output	18
4.5 CREATING THE PRESERVATION VERSION	19
4.5.1 Choosing the preservation method	19
4.5.2 Develop a conversion plan	20
4.5.3 Convert the files	21
4.5.4 Validate file conversion	21
4.5.5 Authenticate the preservation version	22
4.5.6 Metadata update	22
Output	22
4.6 PREPARE THE ARCHIVAL INFORMATION PACKAGE	24
Output	26
4.7 SUBMIT THE ARCHIVAL INFORMATION PACKAGE FOR PRESERVATION	27
Output	27
APPENDIX I	28
FLOW CHART OF INGEST PROCEDURES	28

1. Introduction

The ingest of data resources into an archive — starting from negotiating the submission to storing the verified and documented resource for long-term preservation — must be well defined, transparent and documented. The ingest process must be efficient and smooth, yet discriminating to reject any data and documentation that is not fit for long-term preservation.

Any archive can be seen as offering a service — preservation of data, information, material, etc. over time. Preserving and providing access to material in an archive can only be successful if the deposited material fulfils certain criteria (e.g. suitable physical condition, sufficient documentation, etc.). A digital archive must ensure that the quality of digital resources it ingests for preservation is sufficient for them to be retained over long term and that access to them can be maintained. The quality control performed in the course of ingest of archival digital resources must be:

- thorough to detect any gaps and errors in data;
- systematic and complete to include all files and formats;
- well documented to ensure the accountability of preservation process.

In order to lower the risks of lost access to resources over time and undocumented features in data, the AHDS Service Providers [shall] adhere to common archive ingestion processing procedures as described in this report.

The AHDS has adopted a migration-based preservation strategy. Through successive conversions from one format to another, this strategy retains the information content of data resources, but not necessarily the original experience of using the information content. Migration of digital resources can only be successful if sufficient technical detail about data resources is recorded as their metadata at the stage of their deposit.¹ Validation of data resources in common file formats must be performed by all AHDS Service Providers using the same benchmarks and same processing rules. An OAIS reference model-based AHDS centralised preservation facility (Archival Storage) can be implemented efficiently only if the Archival Information Packages deposited by each Service Provider adhere to common rules. Control and harmonisation of data ingest procedures is the first step towards this goal.

The aim of this manual is to help standardise the data resource ingest processing by the AHDS Service Providers. Through the list of procedures the report provides a meaningful sequence of processing steps for ingest of digital resources into the AHDS collections. The manual makes references to the OAIS model and serves as a preparation for implementation of a common preservation facility for all AHDS Service Providers.

2. Scope

This report describes procedures and processing steps in the management of ingest of a digital resource into an archive. The focus is on accession and processing of new collections from the stage of transfer to the Service Provider until the preservation version of the resource is complete.

¹ cf. 'AHDS Preservation Metadata Framework', 2002

The report is based on:

- material collected from the Service Providers throughout the AHDS preservation consultancy (public documents, e-mail correspondence, visits, interviews, etc.);
- HDS “Collections Manual”;
- OTA “Documenting the Resources at the OTA”;
- OAIS Reference Model (ISO draft standard) and associated materials.

3. Structure

The ingest procedures are divided into following stages:

- > Preliminary or pre-accession stage
- > Accession / Submission
- > Administration and metadata management
- > Quality assurance
- > Creating the preservation version
- > Preparing the Archival Information Package
- > Submitting the Archival Information Package for preservation

Most stages are divided into sub-stages that are described in detail.

The report ends with appendices that present the procedures listed in chapter 4 as a flow-chart and processing rules for some common file formats.

4. Detailed procedures guide

The OAIS Reference Model Ingest entity provides the services and functions to accept Submission Information Packages (SIPs) from Producers and prepare the contents for storage and management within the Archival Storage. Ingest functions include receiving SIPs, performing quality assurance on SIPs, generating an Archival Information Package (AIP), extracting Descriptive Information from the AIPs for inclusion in the archive database, and coordinating updates to Archival Storage and Data Management. The ingest process transforms the SIPs received into a set of AIPs and Package Descriptors which can be stored and accepted by the Archival Storage and Data Management functional entities. The Ingest is described in several phases that lead to the creation of the AIP (see Figure 1).

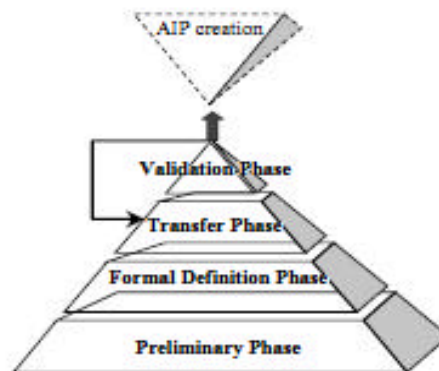


Figure 1. OAIS Ingest interface phases.

Starting with negotiations with prospective depositors, the AHDS Service Providers collect and manage substantial amounts of information about each data resource in their collections. Most of the information that is vital for preservation and dissemination of the archived resources, is created at the ingest stage of the data resource's life cycle. Managing and documenting the ingest, and the processing carried out throughout it, forms the foundation of the accountability and efficient work of the Service Providers.

The AHDS Service Providers currently manage the submission of data resources according to same rules, which largely match the OAIS ingest functions. After the Service Provider has received a SIP transmitted by a depositor in an Submission Session, the SIP is processed by the Service Provider ingest function. This processing includes decompression, disaggregation, verification of all appropriate file components, validation by format and content, and transformation to normative formats, if appropriate. Once all ingest processing is complete, the Service Provider sends a submission confirmation to the depositor via email indicating the success or failure of the ingest. Invalid SIPs that fail during ingest processing are, as a rule, not added to the archive and the depositor must resubmit the SIP with all necessary corrections as indicated in the negative confirmation message.

There is some divergence among the AHDS Service Providers in the methods used for data validation and consistency checking, which are handled differently and carried out on different levels. This manual will suggest common principles for consistency checking of ingested data resources.

Similar to the OAIS functional model, valid SIPs successfully passing the ingest function are transformed by the AHDS Service Providers into AIPs, which are then transferred to the Archival Storage function.

The AHDS preservation strategy defines three logically (but not necessarily physically) distinct renditions of every deposited data resource which are created and retained:

- The original version that was deposited;
- The preservation version created to preserve the information content of the data resource;
- The dissemination version created to be used with currently available software.

The preservation version is created with the following considerations in mind:

- Minimise the need for future migration;
- Minimise dependency on proprietary file formats;
- Maximise software and hardware independence;
- Ensure documentation supplied by the depositor is preserved.

It is the preservation version of each submitted data resource that will become the primary AIP produced by the AHDS Service Providers. The AHDS policy is also to retain the original submission, without any processing or changes made to it and, therefore, the original version must also be compiled into an AIP for submission to the Archival Storage. Although the central AHDS Archival Storage is still in the design phase, each Service Provider has made its own provisions for preserving its collections and submitting the preservation version for storage does not differ in principle from submitting an AIP to Archival Storage.

The depositor can be assured that the AHDS Service Provider has assumed active archival responsibility for the deposit submitted in the SIP only after receiving an affirmative archival confirmation that is issued once the processing of an AIP by the Archive Storage function has been completed.

The AHDS ingest processing procedures and creating the preservation version have been divided into logical groupings as presented in the next chapter. The explained processing tasks must be completed and usually followed in the listed order (cf. Appendix 1). Optional tasks are represented with italicised headings.

Recommendations

- The AHDS should unify the methods of consistency checking that the Service Providers are using for quality assurance of ingested data resources.
- The AHDS should ensure that all Service Providers define and create (where necessary) the three versions/renditions of data resources as a standard practice.

4.1 Preliminary or pre-accession stage

All AHDS Service Providers currently require that depositors sign a deposit agreement or licence, and submit a filled in data and documentation transfer form. These provide the AHDS with information about both the data resources and their creator(s) and correspond to what the OAIS model calls a ‘submission agreement’. The Submission Agreement identifies the SIPs to be submitted (a SIP includes all necessary documentation about its data contents) and the Data Submission Sessions.

Negotiations with prospective depositors and collecting information about the future deposits, which in the OAIS are called the ‘preliminary phase’ (first contact, feasibility assessment, preliminary agreement) and the ‘formal definition phase’ (formalisation of contractual and legal aspects, definition of transfer conditions, validation definition, delivery schedule), are currently not consistently managed and recorded by all AHDS Service Providers. A few simple databases (‘negotiations database’, ‘accessions management database’) are in use, but their structure and management has not been co-ordinated among Service Providers.

At present, the AHDS holds a relatively open collection policy and is prepared to accept for submission virtually any digital resource that falls within its collection remit. With infrequent deposits this strategy is viable and has a low risk level, however, when the flow of submitted data resources increases, the importance of preparing and managing the ingestion of data into the archive will become a necessity. Indeed, the existing accessions management databases have been set up by Service Providers with larger average number of deposits.

The technical characteristics of potential deposits (e.g., file format, complexity of a data resource, usability requirements, etc.) will unavoidably become a more significant argument in the future for acceptance or non-acceptance of a data resource for submission. It is in this light that the AHDS Submission Agreements (Data and Documentation Transfer Forms) must be reviewed and updated to allow for quicker decision-making and better collections management. Asking the depositor to provide as much technical metadata about the potential deposit and a schedule for submissions will enable the planning for preservation by AHDS Service Providers at an early stage — knowing the data resource file formats, usability requirements and data volume in advance forms the foundation for acquisition decision and preservation management. The Submission Agreement could also include a best practice requirement that the depositors do not delete their data resources until been given a confirmation from the AHDS Service Provider that all files have been transferred successfully and can be processed.

The AHDS also receives deposits where the original data creator is unknown or is no longer available to provide all the details and metadata required for correct management of submissions. In such cases, the Service Providers must ensure that as much surviving documentation as possible about the deposited data resource and its context is submitted.

It is also common to assess the feasibility of archival preservation of a digital resource by first conducting a number of checks on a test-sample of the whole collection or resource. The test transfers and preservation feasibility assessments are also practised by National Archives of many countries and have proved their value in efficient management of archives’ collections. Archives usually reserve the right not to accept a collection for submission if based on such feasibility tests it was deemed that the data resource is unsuitable for preservation or insufficiently documented. The AHDS may consider for the future conducting preliminary

feasibility studies with the aim of assessing the risks and defining more precisely the scope of the deposit and archiving project. The AHDS should also explicitly define reasons for refusing a deposit (e.g., when it does not fall under relevant funding criteria, when it is unlikely to serve the AHDS user communities, etc.).

It would be prudent for the AHDS Service Providers to initiate a new entry in their collections management database/preservation metadata database for every prospective deposit and populate the entry with as much technical detail as possible and as early stage as possible. There may be a significant lapse in time between when the initial negotiations were started and when the data is transferred to the Service Provider and knowing the original (or intended) technical details about the data resource will be helpful in conducting the ingest processing. Once the data resource is actually transferred to the Service Provider, the database/metadata can be validated and updated, but the decision-making for suitable preservation formats and procedures for creating the preservation version can be made already earlier.

Each new acquired data resource means a new long-term responsibility for the AHDS. Hence, assessing the future costs associated with each new submission (beyond the ingest operation time) is part of efficient collections management. The analysis of “future impact” on the archive must consider:

- The permanent data volume to store, which in some cases may be triple or more the volume of original deposited data resource. Significant increase in storage requirements may imply for the Service Provider: purchasing additional storage space; change in required IT support; new costs for software, hardware, support, network connectivity, etc.
- The long-term migration requirements and the precautionary measures to avoid the loss of data.
- The security requirements, which may include access control mechanisms, authenticity (fixity) requirements, required security of the preservation system (e.g., number of safety copies, storage vaults, etc.).

Output

The preliminary stage should provide the Service Provider with at least the following set of information about the prospective deposit:²

No.	Goal
A.1	Identify the primary information which the Service Provider must preserve.
A.2	Establish a preliminary definition of the different data objects that will be transmitted.
A.3	Analysis of all aspects of feasibility of preserving the data objects.
A.4	An estimate of the required resources (both short- and long-term).
A.5	A preliminary Submission Agreement that includes the transfer schedule.

In the OAIS standard the preliminary phase is followed by the formal definition phase which results in signing the formal Submission Agreement — the equivalent of the AHDS Deposit Agreement or Deposit Licence. The Submission agreement formalises contractual and legal

² cf. ‘Producer-Archive Interface Methodology Abstract Standard’, CCSDS 651.0-R-1 (Draft), 2002, pp. 17-30

aspects (e.g. intellectual property rights) of the service that the archive is offering to the depositor, it includes the delivery schedule and agreed data transfer methods.

The OAIS also assumes that the Archive has developed a ‘data validation plan’ during the preliminary phase for the given deposit and this plan is included in the Submission Agreement as an informative point for the data creator, as explaining the possible reasons for the Archive rejecting the data resource. A separate AHDS guidance document lists significant properties and known problems with common file formats that could be used as example cases or typical scenarios of processing when negotiating with potential depositors.

Recommendations

- The AHDS should update the Data and Documentation Transfer forms to allow more technical detail and metadata relevant for preservation to be collected from the depositor at the ingest of data resource.
- The AHDS should unify its management and documentation practices of ingest negotiations (e.g., through an agreed structure for an acquisitions or collections management database).

4.2 Accession / Submission

This section describes the processes involved in accessioning a data resource. It assumes that the negotiations with the depositor have been completed and the data transfer agreed upon. Stages 4.2.1-4.2.6 are sequential, but an omission discovered or error encountered at any stage will halt the processing until the issue has been resolved through the depositor providing the necessary data and documentation.

4.2.1 Data transfer session

The data submission session must be managed technically (no cut-offs or transfer problems in network transmission) and the depositor must receive a confirmation of the successful delivery of all files listed in the Submission Agreement. In the case of transmission anomalies the files must be re-transferred and the problems recorded in metadata.

The security restrictions set on the data and stated in the Submission Agreement must be followed and managed throughout the data transfer process.

The OAIS model also recommends using initial transfer tests before the beginning of data delivery, to discover any arising problems and adjust the transfer parameters.

4.2.2 Update metadata

The accessions management metadata must be updated to reflect the transfer of data to the Service Provider. If no prior entry exists for the submitted data resources, a new entry must be initiated.

4.2.3 Virus check

All submitted files must be checked for viruses. If viruses are found the depositor must be alerted and replacement files must be requested and transferred. If the depositor cannot provide replacement files, the files must be disinfected by the Service Provider. All details of any viruses found must be recorded in the metadata.

The processing of submitted files must not proceed until the infected files have been replaced with “clean” files or disinfected.

4.2.4 Media and file readability check

File and media access problems may have been discovered already by the virus check, but additional readability checks are recommended for all media and files submitted.

If the data resource was submitted on a removable storage media, the media must be checked for corruption and if found not readable, replacement must be requested from the depositor.

If the data resource was submitted in compressed or encrypted form, the file check must ensure that the Service Provider has the necessary tools to uncompress or decrypt the transferred files.

The submitted files must be checked for corruption and if found not readable, replacement must be requested from the depositor. All anomalies and corruption cases must be recorded in metadata.

4.2.5 Data resource completeness / integrity check

The submitted data resource must be saved in a dedicated disk area (e.g., acquisitions area) where all sub-directories necessary to mirror the original directory structure of the deposit must be created and files copied into them.

The number of files in each directory must be checked against the submitted documentation that describes the data resource. The comparison with documentation must discover discrepancies at least in the number of files, file naming and file formats. The completeness and integrity of the data resource must be assured.

If some files are found to be missing, the depositor must be contacted and transfer of the missing files requested. If there are surplus files that are not described in the accompanying documentation, the depositor must be contacted requesting additional documentation. If the depositor is not available, or does not have the file(s) or documentation, the omissions must be recorded in metadata.

Example: the HDS is creating a text file called “filelist.txt” that lists all deposited data and documentation files by name and lists all items of hardcopy documentation deposited.

The creation of file lists and integrity checking can be automated by creating script or a macro to carry out the task.

4.2.6 Documentation completeness check

The verification of documentation submitted with the data resource must also be carried out before any data validation processing can commence. The documentation must be checked for completeness (i.e., it must cover all submitted files and all required aspects of data). If the accessions/collections management database includes an entry for the data resource, the current documentation must be compared to the earlier recorded metadata to discover any discrepancies (e.g., different file formats, different level of description, different table structures, etc.).

If any omissions or serious discrepancies are discovered, the depositor must be contacted requesting additional documentation or explanation for changes in documentation and/or data.

It is also useful to make a note (e.g., in metadata) about parts of documentation that have not been submitted digitally and that may need to be digitised.

Provided the documentation format has been standardised, the process of documentation validation can be automated up to a degree.

4.2.7 Copy to processing area

Once all files and documentation are safely transferred to the Service Provider and verified as readable, the data resource can be copied to the processing area for data validation and creation of preservation and dissemination versions. The directory structure of the original deposit must be preserved.

The dedicated processing area must not be publicly accessible and the files must be protected from deletion.

4.2.8 Authenticate original version

Once the deposited data resource has been saved using its original structure, it is possible to use automated authentication methods as a security measure to protect the files from (unintentional) corruption during further processing. Calculating checksums and recording their value in the data resource metadata is an example of such authentication.

Output

At the end of this stage, the Service Provider will have:

No.	Goal
B.1	Complete set of files that belong into the deposited data resource.
B.2	Complete set of documentation describing the deposited data resource.
B.3	All files are accessible and readable.
B.4	All files are stored in a dedicated server area for processing.
B.5	<i>Fixity metadata for the deposited files.</i>

Recommendation

- The AHDS should unify its practices of documenting the process of accessioning and data transfer during ingest.

4.3 Administration and metadata management

A number of administrative tasks need to be completed either in parallel with or after the transfer of the deposited data resource. Documenting the transfer process, checking that the necessary licences have been signed and creating additional metadata are all part of this stage. It is particularly important that all legal aspects of the deposit (i.e., copyright and confidentiality) have been cleared before processing of the submission proceeds.

4.3.1 Check all forms

The AHDS procedures require that the Deposit Agreement/Licence Form must be signed by the depositor. The copyright issues and methods of access to the deposited data are settled through this form, so it is vital for the AHDS to get it approved by the depositor. If the Deposit Agreement/Licence Form has not been completed and signed, the processing of deposit cannot proceed until it has been approved.

Equally, the Data and Documentation Transfer Form must be as complete as possible. Complementing the stage 4.2.6, the documentation provided in this Form should be checked for adequacy. The provided documentation must be sufficient for describing the deposited data resource (including codebooks and user guides). Documentation is 'sufficient' if it provides enough information to generate an AHDS catalogue record, and enough information for a user to fully explore the data resource and understand the process by which it was created. The depositor must be contacted if documentation is inadequate.

4.3.2 Check copyright and confidentiality clearance

The Deposit Agreement/Licence Form must be checked for potential problems with copyright. The depositor must obtain, where necessary, permission to deposit and/or disseminate the data resource. If there are unresolved copyright issues, the processing of the submission must stop until the copyright clearance has been obtained.

The data resource must also be checked for information that could enable the identification of living individuals. As a rule, the AHDS has a remit to disseminate only anonymised data, but the depositor must have obtained all the necessary permissions to deposit and/or disseminate the data resource. If it appears that the depositor does not have these permissions, the processing of the deposited data resource must be halted and the depositor contacted for obtaining these clearances.

Part IV of Schedule 8 of the 1998 Data Protection Act lists conditions under which collected data are exempt from Data Protection clauses for the purposes of research. According to this chapter, the data custodian must, nevertheless, ensure that personal data are not passed to a country without legal protection for personal data equivalent to that in the UK, unless the custodians first assure themselves that the data will be adequately protected in practice.

4.3.3 Update metadata databases

After all the documentation and legal issues have been checked, the processing of the submitted data resource can proceed. If the Service Provider uses an accessions management database that is separate from its collections management database, the accessions management database must be updated at this point to reflect the end of negotiations and transfer of the deposit.

If the deposited data resource does not have an entry yet in the collections management database, it should be (at the latest) be entered into the database now. Entering the deposit into collections management database should also assign a unique ID (e.g., collection id, study number, etc.) to it. The collections management database will also require a title for the deposited data resource and a deposit date.

The depositor could also be informed at this stage that the data resource has been transferred successfully and has been approved (has received an ID in the AHDS collection) and will be taken into validation processing.

4.3.4 Scan paper documentation

Any documentation deposited on paper as part of a data resource must be scanned and stored as digital images alongside the data files. A new directory may be created for documentation images. Same preservation processing procedures apply to these images as for the data files.

Documentation should be scanned at a resolution and colour depth that ensures the legibility of the text, tables, graphs, artwork and other elements present on each page.

Example:

The HDS recommends the following scanning settings:

- Black and white documents, 200dpi 1 bit A4 page images
- Greyscale documents, 200dpi 8 bit A4 page images
- Colour documents, 200dpi 24bit A4 page images

Illegible images should be rescanned at a higher resolution and/or colour depth.

As a rule, the original paper documentation should be retained also after their scanning and stored together with other hard-copy materials deposited as part of the given submission.

Output

At the end of the stage 4.3, the Service Provider will have:

No.	Goal
C.1	Complete and signed Deposit Agreement/Licence Form.
C.2	Signed and as complete as possible Data and Documentation Transfer Form.
C.3	Cleared copyright permissions.
C.4	Cleared confidentiality permissions.
C.5	Closed negotiations for the deposit.
C.6	New collections management entry for the deposit.
C.7	Deposited data resource has a title and an ID.
C.8	The depositor has been notified of progress of the ingest processing.
C.9	Relevant paper documentation has been scanned and stored with the data resource.

Recommendation

- The AHDS Service Providers must ensure they are meeting their obligations under the data protection legislation. A statement to that effect should also be publicised (e.g., on the web pages).

4.4 Quality assurance

Checking the conformity of the data resource to its model described in the accompanying documentation serves from the Service Provider's point of view, the purpose of a quality control. The Archive must ensure that it *can* provide the service it claims to, with the data resource that has been submitted. Thus, the archives usually carry out a number of checks to discover any problems, anomalies and potential difficulties with the files and data they have been given. The consistency check is also important for reducing risks of introducing errors through conversion and processing when creating the preservation and dissemination versions of the data resource.

The validation stages involves three steps: preparation of the validation plan, in-depth validation and metadata update.

4.4.1 Review data and prepare validation

The validation stage should begin with a review of all data and documentation included in the deposited data resource, with the purpose of identifying all data files that need to be validated and the documentation that can be used for validation checks. The validation plan should include:

- all files or parts of files (e.g., tables) that need to be checked;
- documentation that can be used for validating each file or its part;
- validation method to be used for each file or its part;
- method of documenting the validation process.

The validation plan can be a part of preservation metadata schema or collections management database.

The validation plan serves primarily Service Provider's own, internal work management purpose, but if included in the data resource metadata, it also serves as a data authenticity and archive accountability proof. Where many similar data resources are deposited, it would be efficient to develop validation plan templates that could be simply filled in for each new deposit.

4.4.2 Consistency checks

The validation checking must follow the established validation plan or strategy and the results of validation must be documented. The consistency checks may be (semi)automated, in which case the validation must be logged or its results documented in metadata.

The nature of consistency checks depends on the data type under scrutiny, but as a minimum, the checks should be performed on three levels:

- 1) File level
- 2) File structure level
- 3) File content/data level

4.4.2.1 File level consistency

On the file level, it is important to ensure that:

- > The file names described in the documentation supplied by the depositor match the actual file names that were transferred.
- > The file names must not contain white spaces and must have a file extension. Any filenames that do not meet this convention, or use characters not permitted by the computer systems used for processing and preservation must be changed. Original names and new names should be recorded in metadata.
- > The formats described in the documentation supplied by the depositor (e.g., in the Data and Documentation Transfer Form) match the actual file formats that were transferred.
- > The file format check should (where relevant and possible) include format version checking (e.g., MS Access 97 or 2000).

4.4.2.2 File structure level consistency

The file structure will vary according to data type and file format, but a few generic checks that should be performed can be listed:

- > The adherence of plain text files (e.g., ASCII text, delimited text, plain text mark-up, etc.) to the appropriate character encoding should be checked.
- > Structured text files should be checked against the relevant formal definition of the structure (e.g., field and record delimiters for a 'table', for more complex structures it may be a DTD, XML schema or similar method of definition).
- > Any plain-text data files that have a syntactic description of their structure available should be checked for conformity with the structure description.
- > For binary files, open them in a suitable software package, preferably using the same software and version that was used for creating the data file, and check for structure and format consistency. This check may include, for example:
 - database structure conforms with the described structure (e.g., number and names of tables in a database);
 - table structure conforms with the described structure (e.g., number of columns in a spreadsheet);
 - image resolution conforms with the described resolution;
 - number of layers in a GIS corresponds to the documented number;
 - the digital audio and video standard (e.g., MPEG-2) corresponds to the documented standard;
 - all described constituent parts of a virtual "world" are present in a virtual reality data resource;
 - etc.

The AHDS has developed guidance consistency issues found with popular software packages. Based on these, new tests and checking routines will have to be created for individual data resources.

4.4.2.3 File contents / data level consistency

The in-depth validation of data must be undertaken when it is known that the data file will undergo conversion for creating the preservation or dissemination version. In other cases thorough checks of file contents should be carried out as best practice.

Methods for validating the informational contents of a file depend on the data type and the file format. The source materials/subject area of the digital resource may also help AHDS Service Provider staff identify key aspects of consistency (e.g., consistent colour reproduction is more important in scans of art works than scans of 19th century statistical publications). As a rule, more work is involved with consistency checking of databases, GIS, spreadsheets and text files. Most of this checking has to be done manually or can be only semi-automated. Although not easy to define, the consistency of data resources like images, digital audio and video, CAD and virtual reality, should also be checked to ensure that the collections the Service Provider is ingesting contain what they are said to contain.

For application-specific data consistency checks consult a separate AHDS guidance document. A list of general categories of checks that should be performed is given here:

- > 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.

If any inconsistencies are found, attempts should be made to resolve the problems, if this is possible with simple methods (e.g., correcting a file name, changing name of fields in a table, re-creating a missing relationship, etc.). For serious problems that need clarification, the depositor must be contacted for help, additional documentation or re-submission of data files.

All validation checks applied, results of validation testing and any remedial action or changes made in data must be systematically recorded in metadata. If a problem cannot be satisfactorily resolved also after contact with the depositor, a comment to alert users of the data resource must be added to the data resource metadata.

4.4.3 Metadata update

Once the content of the deposited data resource has been validated and checked for consistency with submitted documentation, the Service Provider can begin with compiling the catalogue entry for the data resource. The in-depth validation of data will have provided understanding of the data resource and its structure, based on which the end-user oriented metadata can easily be written. Knowledge of structure, codings, compression algorithms, ranges, quality, etc. of the original data resource are a vital information resource for the users of the data. If any details of documentation are still found to be missing at this stage, the depositor should be contacted.

All changes made to the data during the validation, even the minor ones (e.g., correcting a data type of a field in a table, adjusted orientation of a picture, correcting a misspelling in a mark-up tag, etc.) must be recorded in the metadata of the data resource as part of the processing performed by the AHDS Service Providers. Classes of minor change, for example changing all the misspelled mark-up tags in an XML document, need not be recorded individually. Although usually not serious threats to the authenticity of the submitted resource, however,

such changes do alter the original before any further versions are created from it. Thus as they will be replicated in future conversions of the data resource it is important to ensure they are recorded.

The second important outcome of the data consistency check is information about all files and their components that will become objects of digital preservation. As already stressed, the data validation process must be thorough and cover all files and data that were submitted, also to discover any undocumented features or components (e.g., embedded objects) that need to be saved and preserved separately. These — effectively changes in the submitted data resource — must be recorded and feed into the decision-making for choosing appropriate preservation version (cf. stage 4.5.1.2 below).

Any changes of file names and/or extensions made in the course of the consistency checks (either because of mismatch between the documentation and the submitted files, or to match the file extension with its data type, etc.) must also be recorded in metadata, as these changes affect the authenticity of the original submitted data resource.

Output

The quality assurance will end with:

No.	Goal
D.1	A documented validation plan (i.e., methods) for the submission.
D.2	Validation tools (e.g., automatic scripts) tailored for the submission.
D.3	Validated and checked files.
D.4	Validated and checked data.
D.5	Documented validation results.
D.6	Documented changes made in data.
D.7	User-oriented metadata/catalogue record.
D.8	Documented features of data that are significant for preservation.

Recommendations

- The AHDS should develop common principles and guidance for consistency checking for widely used data resource types, and implement them in practice throughout all the AHDS Service Providers.
- When problems are discovered during the quality assurance, they should be resolved and documented in the following order of preference:
 - issues are resolved and actions taken recorded in metadata with assistance of the depositor;
 - issues are resolved and actions taken recorded in metadata, relying on subject expertise of the AHDS Service Provider in question;
 - issues are recorded in metadata, but left unresolved.

4.5 Creating the preservation version

According to the OAIS reference model, once the Submission Information Package (SIP) is within the OAIS, its form and content may change. An OAIS archive is not always required to retain the information submitted to it in precisely the same format as in the SIP. Indeed, preserving the original information exactly as submitted may not even be desirable.

The AHDS Service Providers have identified preservation file formats for the data types that they accept for submission. Where possible, these formats are open standards and pose little risk for long-term preservation. However, if several preservation formats are available, the choice of one or another format for preservation is currently not based on fixed rules or guidelines. Equally, the conversion methods and procedures for creating the preservation version are not fixed, recorded, nor documented by the AHDS Service Providers. This section of the manual is suggesting procedures for a more transparent and accountable conversion strategy when creating the preservation version of the submitted data resource.

The choice of preservation methods and formats may begin already in the pre-accession stage, when the AHDS Service Provider is negotiating the deposit. Provided sufficiently detailed information about the potential submission is available, the assessment of risks associated with file formats to be deposited can begin early on and a conversion plan for each file may already be completed by the time the processing reaches the validation stage. However, with more complex or larger data resources it may be impossible to devise a complete conversion plan in advance, without analysing the data resource first and, therefore, this processing stage is described here in detail.

4.5.1 Choosing the preservation method

Data resources that are deposited in formats not suitable for long-term preservation, must be converted to one or more additional formats that can be retained for long term and disseminated to the users. The choice of appropriate formats must approach all files and formats submitted as part of the data resource individually and assess risks associated with preserving the original files without converting them. The following steps should be involved in choosing the appropriate preservation method for each file:

4.5.1.1 Create a copy of all files

The original version of the digital resource is an integral part of the AHDS's preservation strategy, thus, it is important that any processing necessary to create a preservation version of the data resource is carried out on a temporary copy of the deposited files, in order to avoid any chance of changing the original version.

All files should be copied to a separate, dedicated disk or server area and the directory structure of the original must be replicated.

4.5.1.2 List all file formats

All file formats that comprise the original data resource must be listed. This data may have been gathered already at the quality assurance stage (cf. stage 4.4.3 above), but may be reviewed and updated at this stage.

A list of "default" preservation formats should be added to file formats found in the data resource. If there are several suitable preservation formats, all possibilities should be listed.

4.5.1.3 Analyse file formats and choose the best preservation format

Using the compiled list of file formats in the submission and their suitable preservation formats, an assessment should be made to find the best fitting preservation format for each file. This should be based on experience, expertise and background knowledge of the AHDS Service Providers, and should utilise the AHDS Technology Watch information. The choice of a preservation format must consider the significant properties of the deposited data resource that need to be preserved. The most important of these will be listed in the agreements with the depositor and would have been discussed during the negotiations at the pre-accession stage. However, using their expertise of data processing, the AHDS Service Providers may maintain a list of previous or potential problems with file formats, and such a list should be used as a bench-mark or checklist that the deposited file formats are compared against.

- > If the submitted original file is in a format that the AHDS Service Provider has deemed suitable for long-term preservation, no processing or conversion is necessary.
- > If the submitted original file is in a format that has only one “default” preservation format, risks and losses occurring by converting the original into preservation format must be assessed. If these risks are low or non-existent, the “default” preservation format must be confirmed. If there are significant risks to the integrity of the data resource or the conversion outcome is likely to be un-verifiable, the problem should be recorded and further analysis for finding a different preservation format or conversion tool must be started.
- > If the submitted original file is in a format that has several possible preservation formats, analysis must be performed to choose a preservation format that carries least risks and has best conversion tools available. If none of the preservation formats is suitable, further analysis is required to find a different format for the retention of the data resource.
- > If the submitted original file is in a format for which there is no preservation format selected yet, the Service Provider must turn to the AHDS Technology Watch and other information sources to collect sufficient information to allow a confident decision for conversion to a new preservation format.
- > If no preservation format can be identified for a file format, the file should be preserved in original format and converted to several currently widely available file formats (e.g., ASCII, XML, MPEG, etc.) that will preserve at least the informational content of the original file. This strategy must include the caveat that the AHDS must monitor the file format development for the particular data type and chooses a suitable preservation format as soon as one becomes available. Conversion and/or creation of a preservation format may be delayed also when it is known that a new standard format will be released shortly.

The list of chosen preservation formats for all file formats included in the submission (including documentation files) should be approved and recorded or included in the metadata of the data resource. It will become part of the authenticity audit trail.

4.5.2 Develop a conversion plan

Using the approved list of preservation formats, conversion paths for every file format must be developed. The conversion of the original file format into the chosen preservation format may involve several stages and multiple software packages. The Service Provider must ensure that

it has access to these software packages and their required versions. When developing the conversion plan, the following aspects must be borne in mind:

- the long-term preservation needs of the data resource;
- significant properties of the data resource;
- the funds and resources available;
- the technical infrastructure available;
- the knowledge and expertise available.

Conversion plans for common file formats can be re-used once they have been developed and reliably tested for non-lossy conversions. However, if a new conversion path, that has not been used before, is to be used, then the conversion must be performed on a test sample of the data. Once the results of the test have been verified as accurate and matching the original, the conversion path can be approved and used for the whole data file. Results of testing the conversion paths, in particular the multi-stage conversions involving several software packages or when open source software is used, should be documented as part of the data resource metadata.

The approved conversion plan should become part of the data resource metadata as proof of authenticity of the preservation version. Its value is not significant until the original submitted version of the data resource remains usable, but the conversion method used will gain the equal value with, for example, description of the original data collection methods, once the original deposit has become obsolete. The original conversion plan will also be used when time arrives for next conversions as part of the data migration strategy.

4.5.3 Convert the files

Using the approved conversion plan, all files that need to be converted into their chosen preservation formats must be converted. The conversion process should be documented (e.g., date, person responsible, names of new files created, etc.). A copy of the files that do not need converting must be made to the same disk area where the new preservation version is being assembled.

Result of the conversion can be one of the following:

- file in a more recent version of the same file format;
- file in a new (different) file format that can retain the authenticity and functionality of the original over long term;
- file in an open standard file format that can be retained over long term.

Of these choices, the third is preferable.

4.5.4 Validate file conversion

Methods of validating the conversion results and the depth of validation depends on the data type and file format that was converted. The preservation version of every file must withstand the same kind of quality and consistency checks as described in sections 4.4.2.2 and 4.4.2.3 above. After the conversion, the significant properties of the digital resource must still be present and unaltered, other features of the resource, however, may have been lost or changed in the course of conversions. As a general principal as much of the original characteristics and functionality as possible should be maintained in the conversions.

A comparison of data values in the original version and the created preservation version is in most cases the easiest method of validating the conversion (e.g., make sure that tables have retained their original structure; text has retained its original formatting where necessary; links to objects have not been broken; footnotes in text have been retained; etc.).

When reliably tested conversion paths are used, spot checks of conversion results may be acceptable, but the extreme cases should always be checked (e.g., first and last row/line; high and low values; numeric and text values; length of file; etc.).

Validation results must be documented as part of the data resource metadata.

4.5.5 Authenticate the preservation version

As soon as the preservation version files are ready and validated, authentication information must be attached to them. This can be done using a check-sum algorithm, a digital signature, etc. The authentication information must be recorded in the preservation metadata³ and will be used for demonstrating the authenticity of the data resource via audit trail throughout the long-term preservation.

4.5.6 Metadata update

At this point, the Service Provider should have all the information available to complete the preservation metadata schema for the submitted data resource and its preservation version. The preservation metadata set will be started by the Service Provider that received the data resource, it will be deposited together with the data resource in the AHDS archival storage/central preservation service (as part of the Archive Information Package), and will be used and amended throughout the subsequent preservation processing of the data resource.

Updating the preservation metadata set could be undertaken as a separate stage, or be included in the compilation of the Archival Information Package (see section 4.6 below).

Output

The creation of a preservation version stage will end with:

No.	Goal
E.1	List of file formats included in the deposit and their corresponding preservation formats.
E.2	Conversion plan for creating the preservation version.
E.3	Preservation version of all files included in the deposit.
E.4	Documentation on the conversion of files.
E.5	Documentation on the methods and results of validating the conversion results.
E.6	Authentication information on the preservation version files.
E.7	Updated preservation metadata schema for the preservation version of the data resource.

Recommendations

- The AHDS Service Providers must ensure that any file processing when generating the preservation version must be done using a copy of the original data resource.
- The AHDS must unify the list of preservation file formats it recommends and uses.

³ see Fixity Method and Fixity Information (elements 8 and 9) in the 'AHDS Preservation Metadata Framework', 2002

- The AHDS should ensure that conversion plans are developed in advance of migration needs arising, and that these plans include validation methods for conversion results.

4.6 Prepare the Archival Information Package

According to the OAIS reference model the generation of Archival Information Packages (AIPs) involves one or more Submission Information Packages (SIPs) being transformed into AIPs that conform to the Archive's data formatting and documentation standards. The descriptive information that needs to accompany data in an AIP should be agreed in advance with the Archive.

In the AHDS context this means that the Service Provider must prepare an AIP of the preservation version of a data resource and all its accompanying documentation together with all relevant metadata, and then submit it to the AHDS archival storage for long-term preservation. The precise content and format of the documentation and metadata required for submission into the AHDS archival storage is yet to be defined as part of the Service Provider/AHDS archival storage interface design. However, it is clear that responsibility for creating the preservation metadata, structural metadata for end users and any catalogue records lies primarily with the AHDS Service Providers.

The Service Providers are also responsible for defining the collection, object and file levels in the data resource and applying the preservation metadata schema⁴ accordingly. Different data types included in a composite object (e.g., components of a web page, embedded objects in a word processing document, etc.) will be treated separately for preservation processing, but will be kept together 'intellectually' through their metadata records. Defining such composite objects and describing them accordingly is the task of the Service Providers prior to transferring the data resources into Archival Storage.

The AHDS follows the policy of preserving the original deposited data resource aside the preservation version it creates. Therefore, an Archival Information Package must also be compiled from the original version of the data resource and submitted to the Archival Storage for preservation. Standards for documentation may differ slightly for the preservation version AIP and the original version AIP.

The AIP that the Service Provider submits to the Archival Storage is, ideally, a self-describing and self-documenting package that includes both the data content and description content.⁵ The most important of the description sets included in an AIP is the Preservation Description Information (PDI). A detailed view of the composite objects in an AIP is shown on the Figure 2.⁶

⁴ 'AHDS Preservation Metadata Framework', 2002

⁵ for the OAIS definition of the AIP see CCSDS, 'Reference Model for an Open Archival Information System (OAIS)', CCSDS 650.0-B-1, 2002, pp. 4-33 to 4-45

⁶ cf. also 'AHDS Preservation Metadata Framework', 2002, pp. 43-48

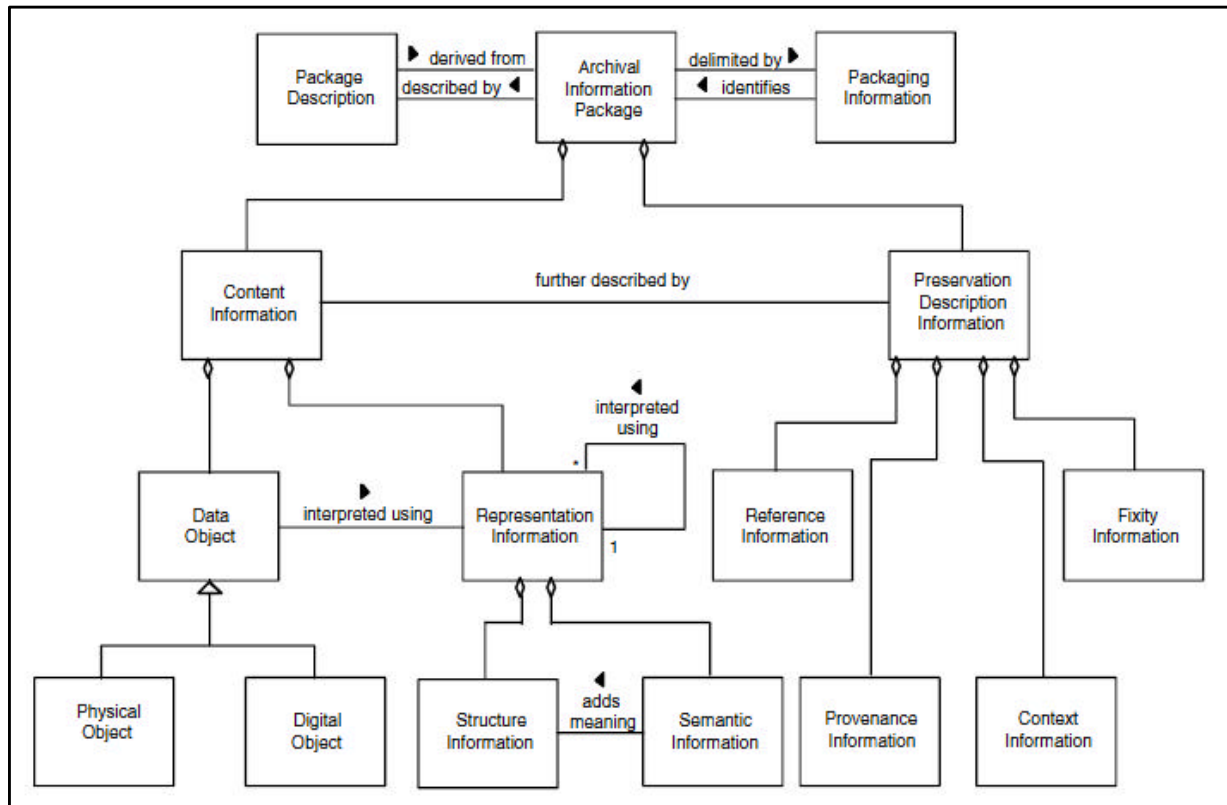


Figure 2. Detailed view of the OAIS Archival Information Package contents.

The mapping between the SIPs and AIPs may not always be necessarily one-to-one. The AHDS Service Provider may choose to split a deposit into several collections it preserves, or it may assemble several deposits into one collection. The possible permutations of the SIP-AIP mapping are the following:

- > **One SIP – one AIP**
This is the most common situation where the Service Provider submits for preservation the whole deposit it received, together with all relevant description.
- > **Many SIPs – one AIP**
The Service Provider may choose to collect a number of smaller deposits or deposits made over a period of time into one collection, describing and preserving them as a single collection. The intellectual control of such AIPs through metadata is crucial in maintaining the integrity of the collection.
- > **One SIP – many AIPs**
The Service Provider may choose to split a larger deposit into smaller collections or studies (or even define single files as AIPs, if they are very large). This will facilitate efficient preservation processing in the Archival Storage.
- > **Many SIPs – many AIPs**
The Service Provider may decide to re-arrange a number of deposits into differently composed collections (e.g., according to the data type, data collection sources, chronological periods in data, etc.). In this case, the AIPs submitted to the Archival Storage would differ from deposited SIPs.

> One SIP – no AIP

The AHDS Service Providers are accepting also “virtual” deposits, which means that the Service Provider maintains the catalogue information or provides a link to the data that is hosted somewhere else, but the Service Provider does not carry the responsibility for preserving the data resource. In this case, the depositor may submit new data or update a “virtual” deposit, and the Service Provider will process the deposit normally (skipping some stages that are not relevant), but will not submit an AIP for preservation.

The management of these combinations will be handled through preservation metadata in the collections management database.

Output

The preparation of the AIP from the preservation version results with:

No.	Goal
F.1	One or several AIPs ready to be submitted for preservation.

Recommendation

- In developing the centralised preservation facility, the AHDS should also develop a specification for its “internal” Archival Information Package, i.e., the submission that the Service Providers will be transferring to the AHDS preservation system.

4.7 Submit the Archival Information Package for preservation

The procedures involved in the AHDS Service Provider and Archival Storage interface have not been specified yet, but the submission of AIPs for preservation in the Archival Storage follows the same principles as the transfer of SIPs to the Service Provider. A submission session must be negotiated with the Archival Storage, the Archival Storage may define a processing queue for submissions and send a confirmation the successful transfer of AIPs to the Service Provider. Access to stored AIPs for the Service Providers will have to be negotiated with the Archival Storage.

Until the AHDS Archival Storage facility is fully functional, Service Providers will continue to make their own arrangements for ensuring the storage and long-term preservation of their AIPs. Commonly this involves making several copies of the preservation version of data resources, or copying the preservation version to a server area where it gets mirrored to other servers. The AHDS does not currently have a consistent policy on using off-line 'safety' copies of its collections. It is assumed that such a policy will be developed and implemented once the centralised preservation service for all Service Providers is set up.

Once the preservation version is safely stored in an archive, the Service Provider must update its metadata schemas to record the new location (if relevant) of the files and document the archiving procedures. After the AIP files have been safely transferred to the Archival Storage, any intermediate versions and copies of the data resource files made for validation or processing can be deleted. The Service Provider may choose to keep a copy of the preservation version AIPs on its own servers, but as a rule, this should not be necessary once the AHDS Archival Storage has been set up.

If the depositor submitted some documentation in the paper form and the Service Provider has a policy of preserving it, all hard-copy versions of documentation and other materials pertaining to the submission must be filed and stored in the paper-archive of the Service Provider.

Output

The transfer stage of the AIP from results with:

No.	Goal
G.1	One or several AIPs transferred to the Archival Storage.
G.2	Updated metadata records.
G.3	Deleted any intermediate copies of the data resource.
G.4	Filed the hard-copy materials associated with the data resource.

Recommendation

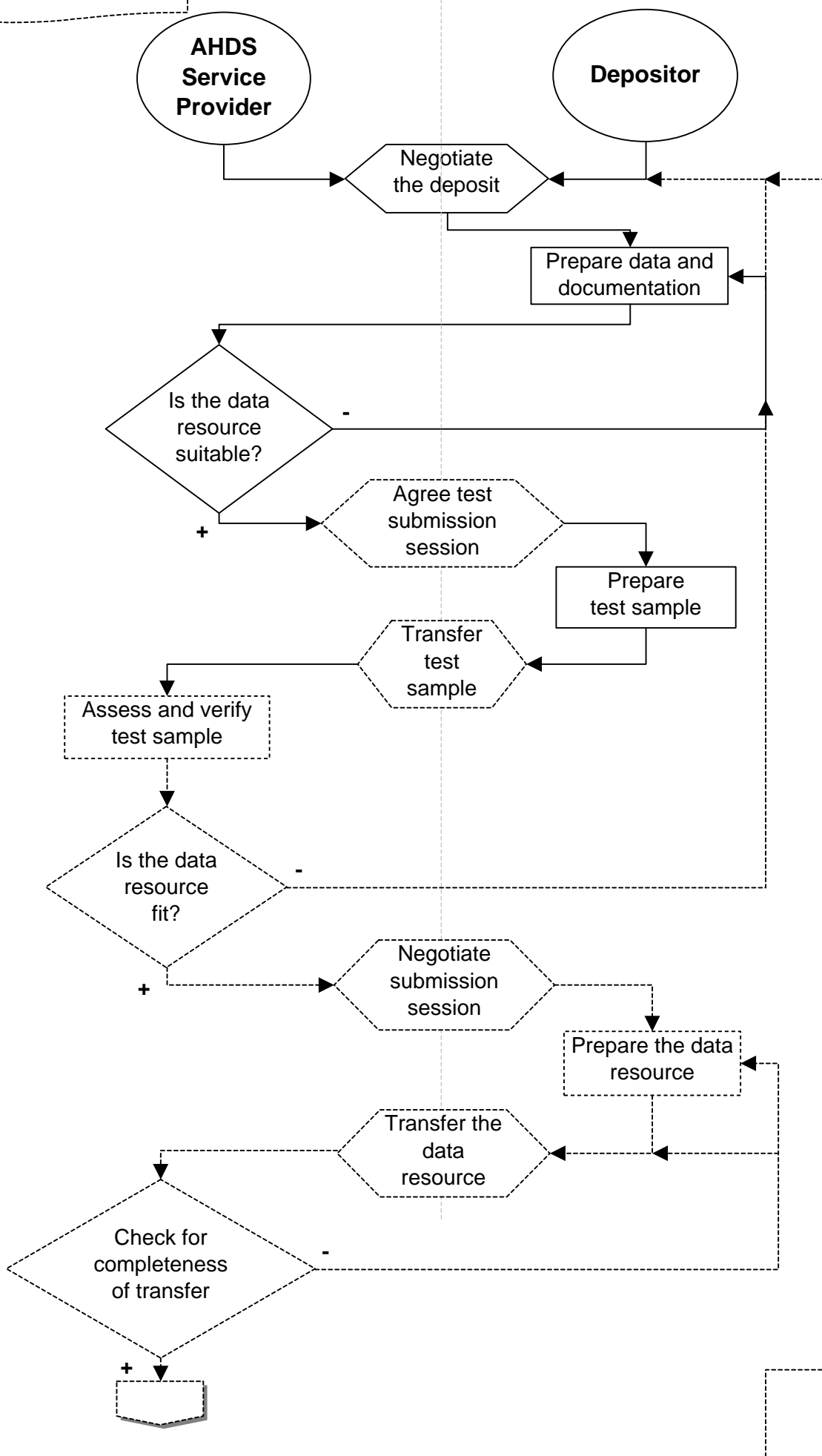
- The AHDS must agree on a set of preservation metadata and define clearly the responsibilities for updating it and rules for access to it for the networked 'architecture' of many Service Providers and one AHDS preservation system.

Appendix I

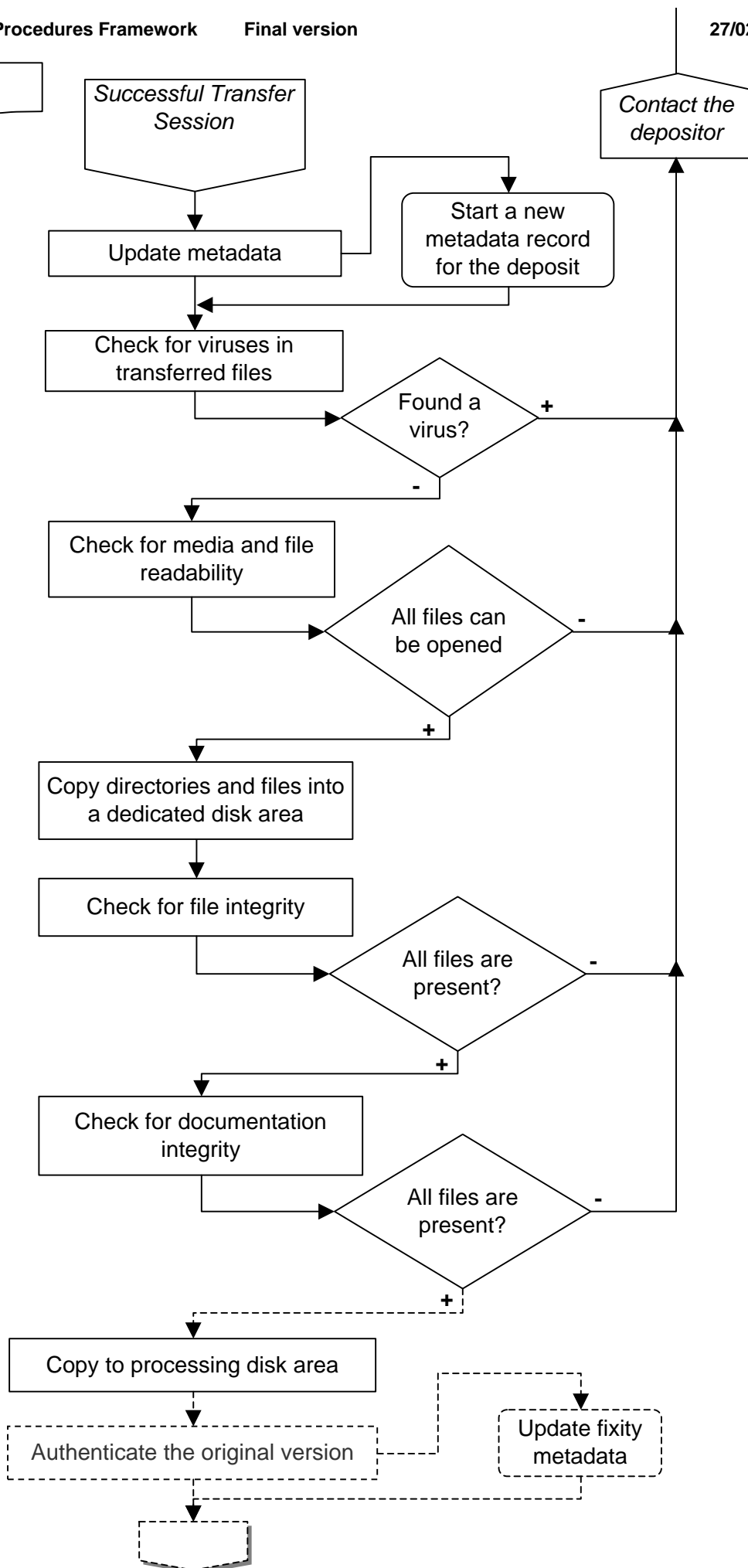
Flow chart of ingest procedures

See the attached file Flowchart.xls.

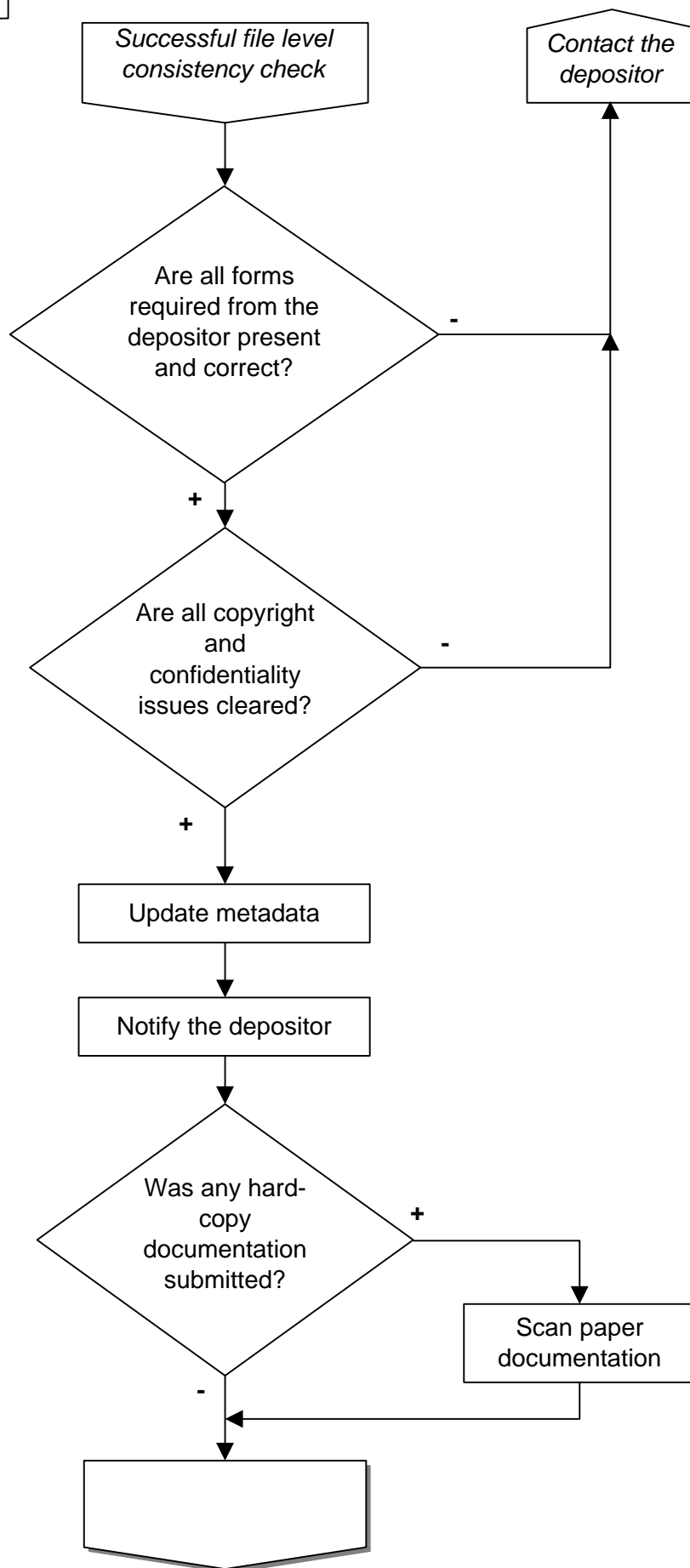
PRE-ACCESSION

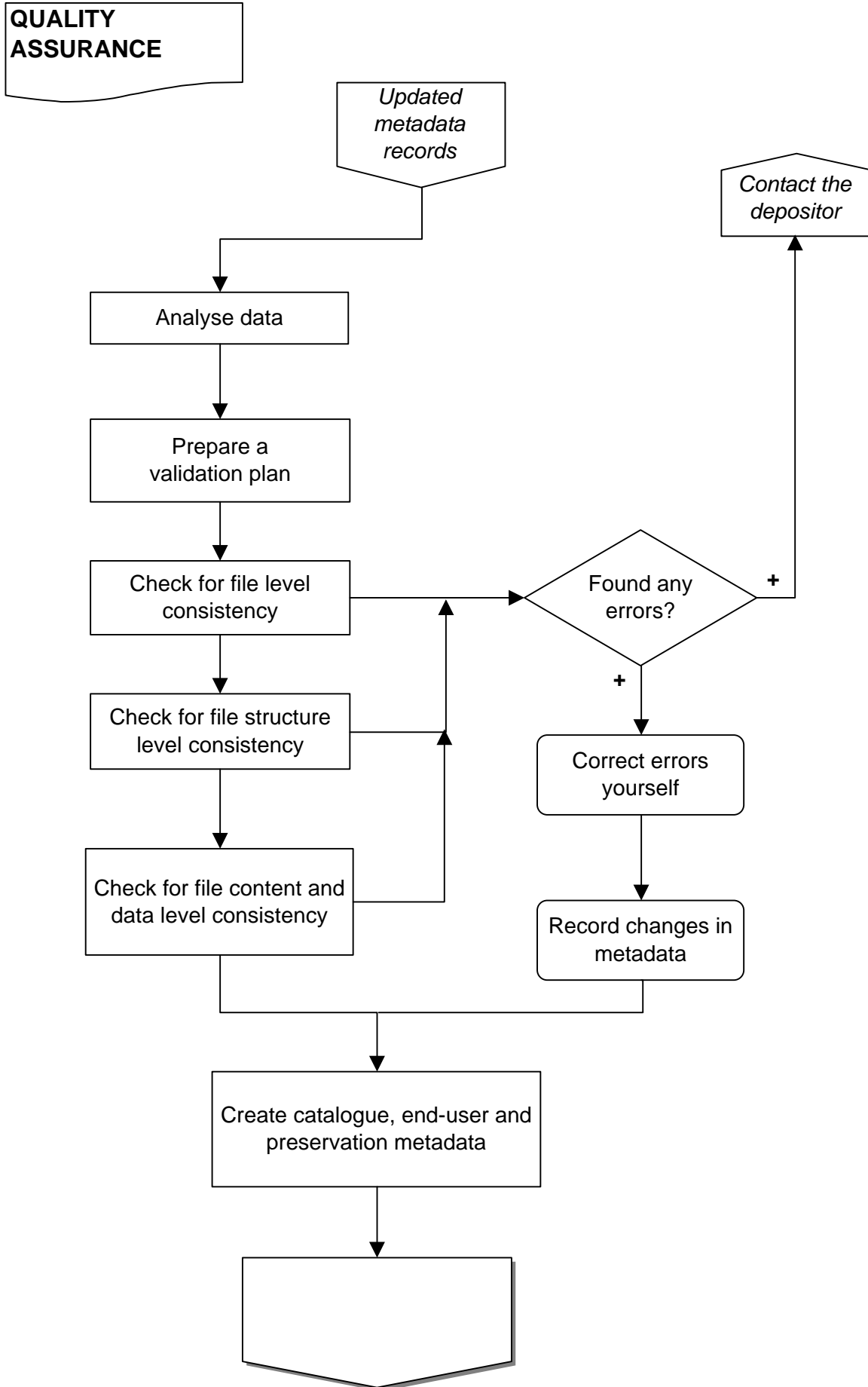


ACCESSION

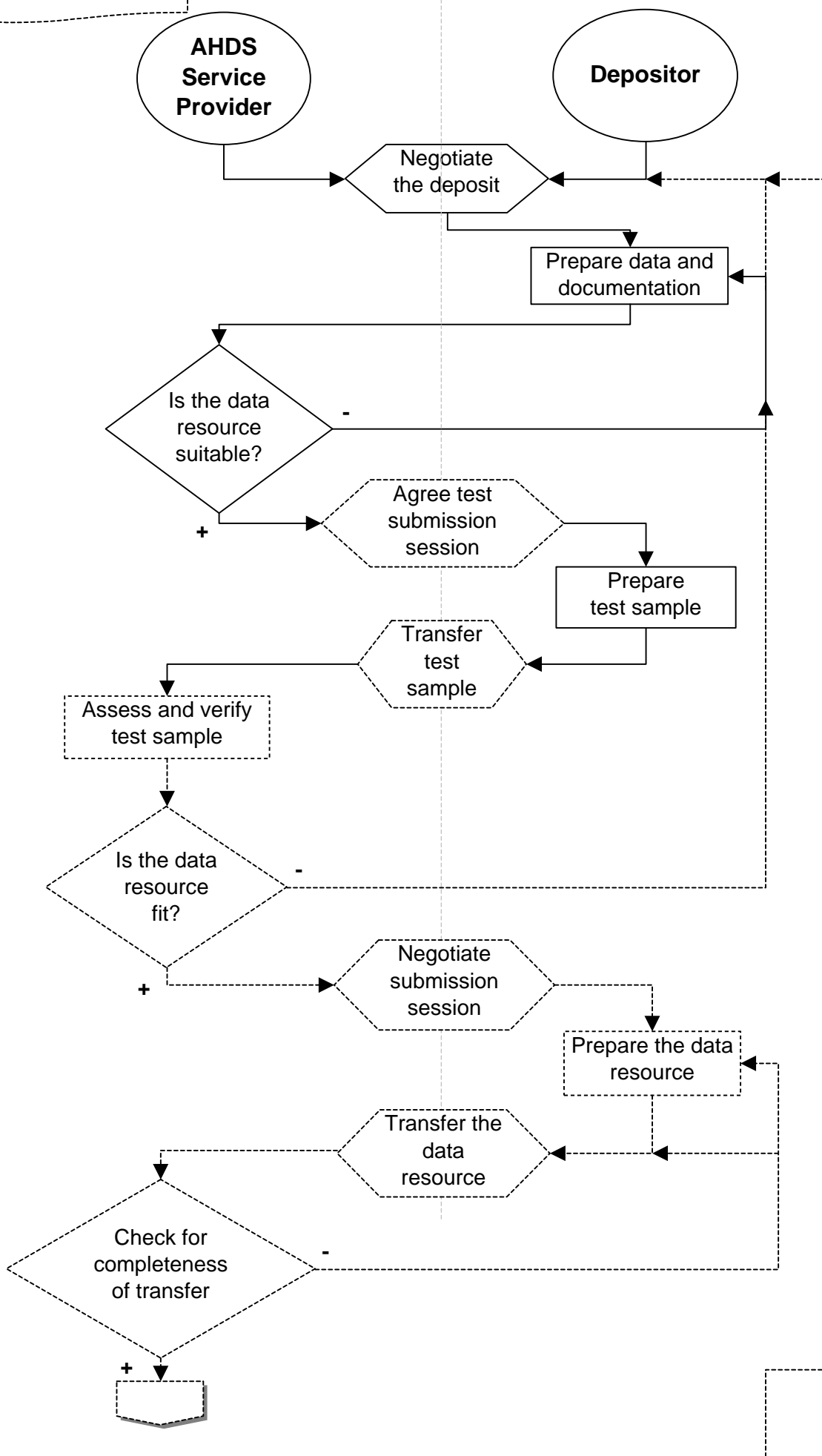


**METADATA
MANAGEMENT**

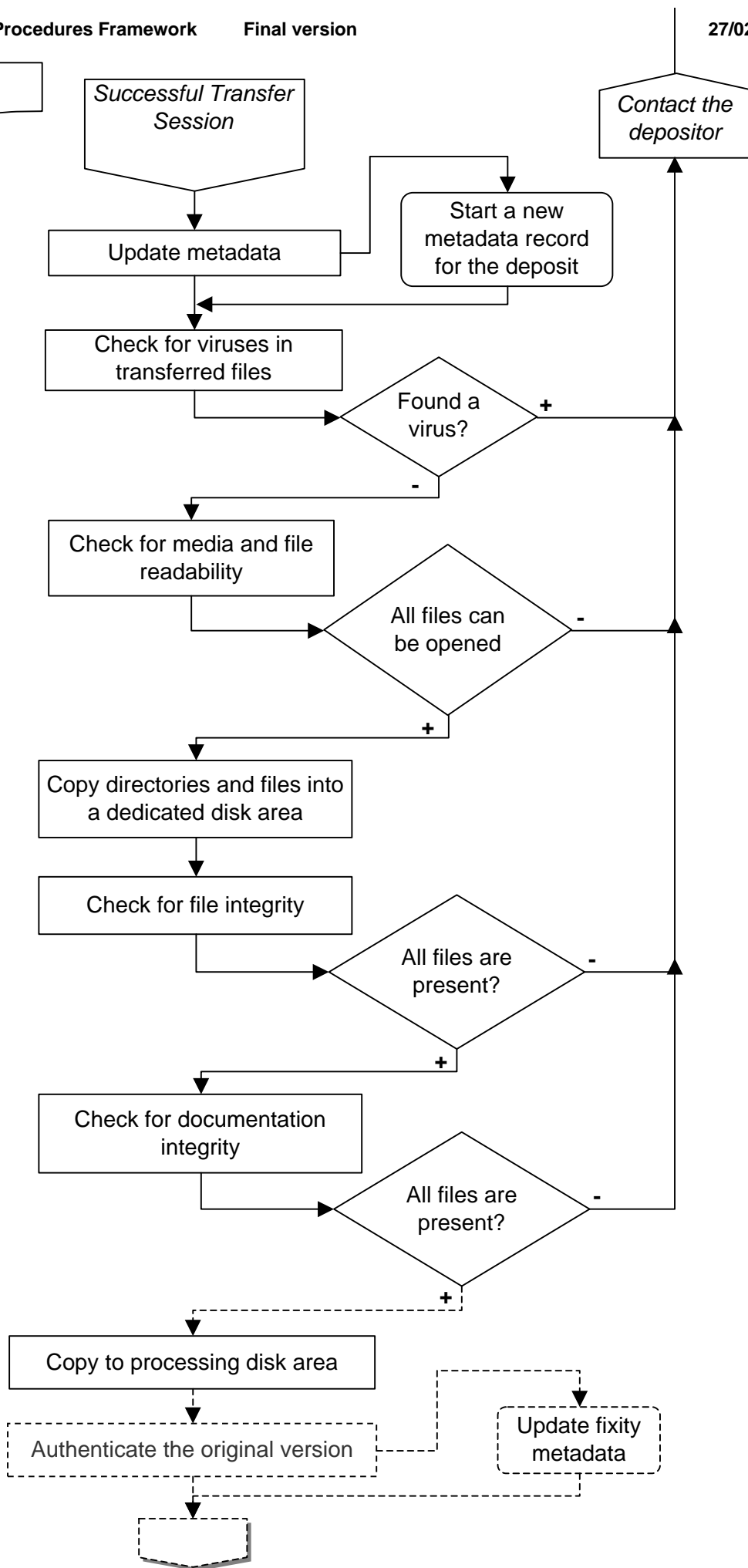




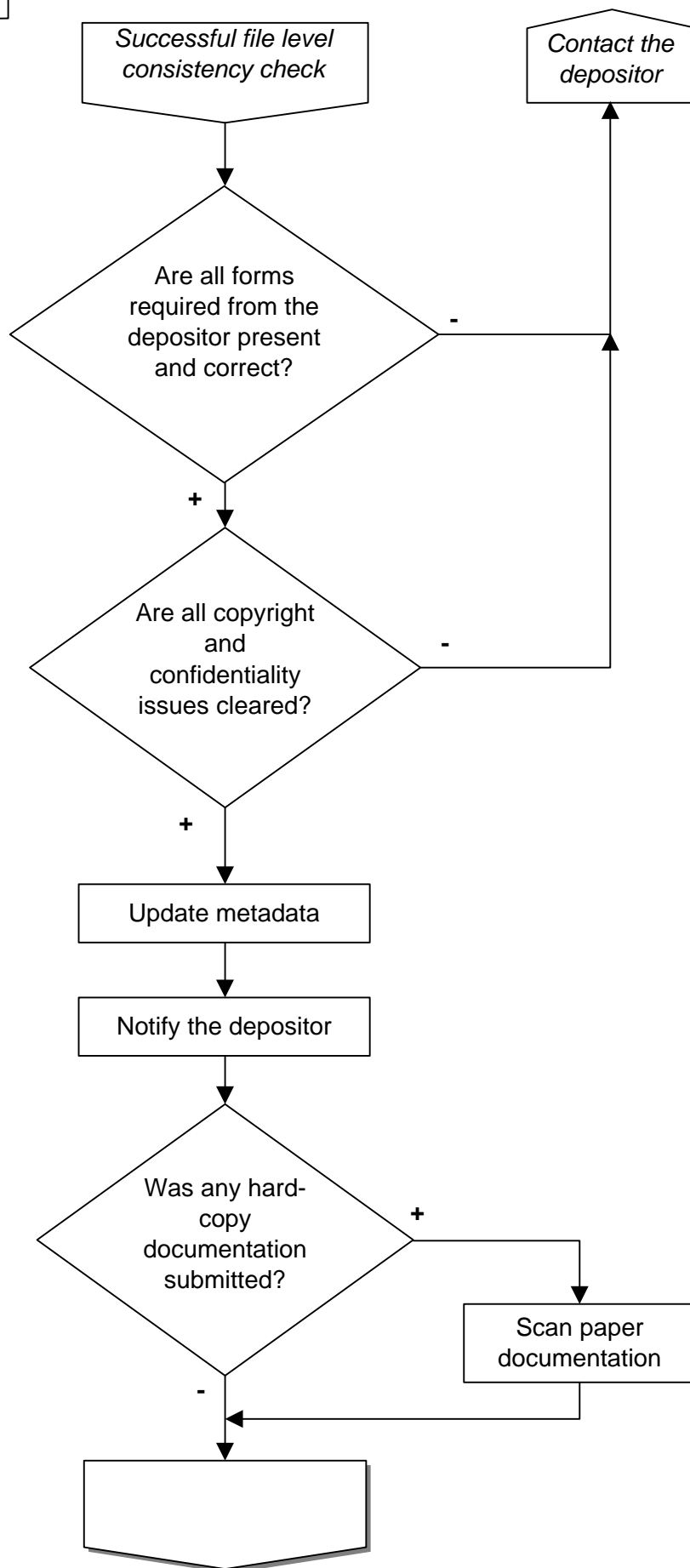
PRE-ACCESSION



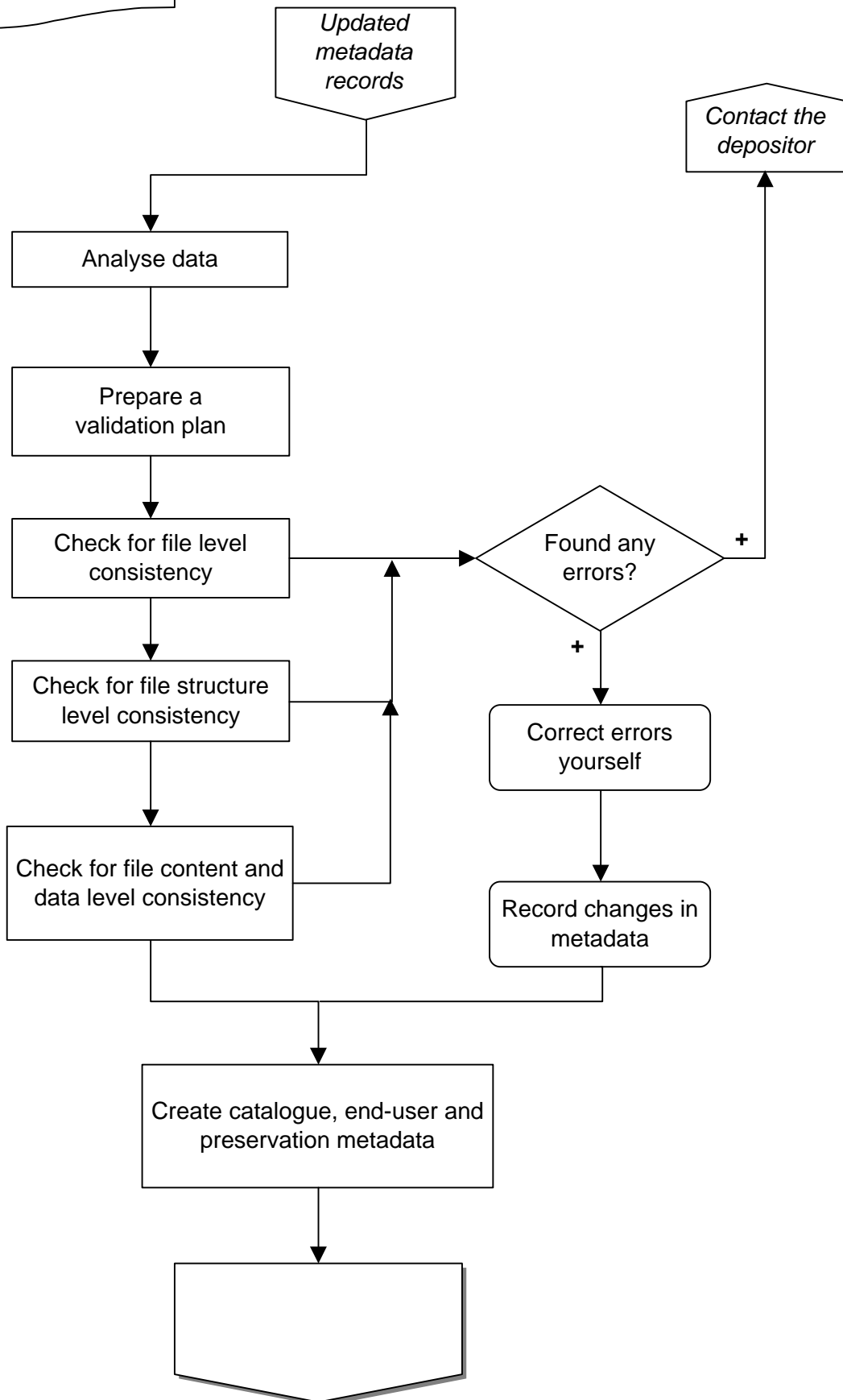
ACCESSION



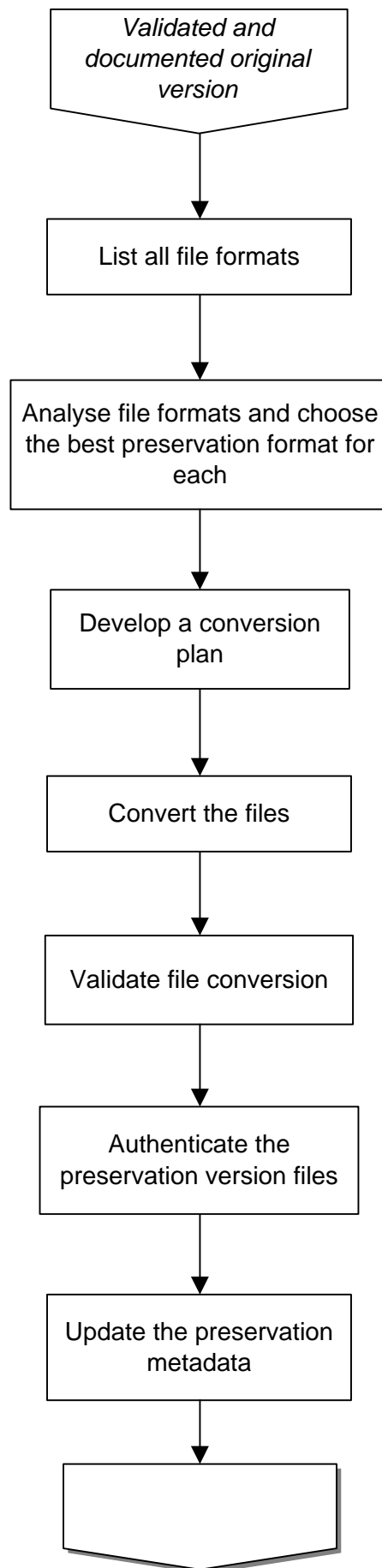
**METADATA
MANAGEMENT**



QUALITY ASSURANCE



**CREATE THE
PRESERVATION
VERSION**



**PREPARE THE AIP
AND SUBMIT TO THE
ARCHIVAL STORAGE**

