Appendix 1

Best Archiving Practice Specifications

This document has been published under the auspices of the EU Telematics Implementation Group - electronic submissions (TIGes)

Please note that this document has been published with the aim that agencies and applicants can gain practical experience of archiving eCTD and NeeS electronic Submissions. This document provides agencies with recommendations on archiving submissions, and introduces applicants to preparing quality archivable eSubmissions.
The TIGes considers that through this process authorities and applicants will gain valuable experience of what is required to ensure sustainability of electronic Submissions in the future.

Version 1.0

June 2013

Document Control

Change Record

Version	Author(s)	Comments
0.3	Arian Rajh, Pieter Vankeerberghen, Jaana Pohjonen	
0.9	Arian Rajh, Pieter Vankeerberghen, Jaana Pohjonen	
0.9F	Juha-Pekka Nenonen, Jaana Pohjonen Pieter Vankeerberghen Arian Rajh	
1.0	TIGes	Endorsed at the meeting on 30 May 2013

Table of Contents

1. Introduction

This Appendix to the Best Archiving Practice defines detailed specifications for archiving of eSubmissions. It contains a description of a recommended digital archival information system, archival information packages and the content of these packages. Content description consists of guidelines for files and fonts. Archival information package and content specification are recommended to be applied by all agencies. The solution to implement an archival information system varies by agency but the main principles described here should be applicable in most cases.

This Appendix can be updated separately from the Best Archival Practice document as file format versions and other such details will change over time.

2. Digital archival information system

2.1. Functional model

Digital archiving in agencies should be based on the OAIS reference model, a simple communication model of information packages being transferred from producer to archive and to user. Archiving of eSubmissions should be based on applicant - digital archive - assessors communication model, according to which assessors can use the eSubmissions before or in parallel with archiving.

According to the OAIS reference model, on which the digital archive concept described in this guidance leans on, long-term preservation task of information packages is provided by fulfilling various functions such as **ingest**, **archival storage**, **data management**, **system administration**, **preservation planning** and **access** to submitted and archived packages.

The **ingest function** ensures acceptance of submitted information packages in specified and agreed format. This means that format-specific specifications must be met so digital archival information system can accept documentation. In eSubmission scenario eCTD and NeeS guidelines, validation criteria, and eCTD specifications establish a unique and final form of submitted information package and represent agreement on the delivery of content. The ingest function enables the system to ingest the submission information package on hard optical media or through the gateway. This package is to be validated according to agreed criteria.

It is advisable that archival storage is implemented as separate subsystem that cannot be accessed by users without records management related access rights. **Archival storage** should not be considered as equal to file system or document management storage for active documents. It is a record management part of the system (software and hardware). This kind of storage contains carefully designed and created archival information packages as archival copy of documentary holding of the agency.

The **Data management function** should be responsible for updating and managing metadata of archived content. It maintains the integrity of the database.

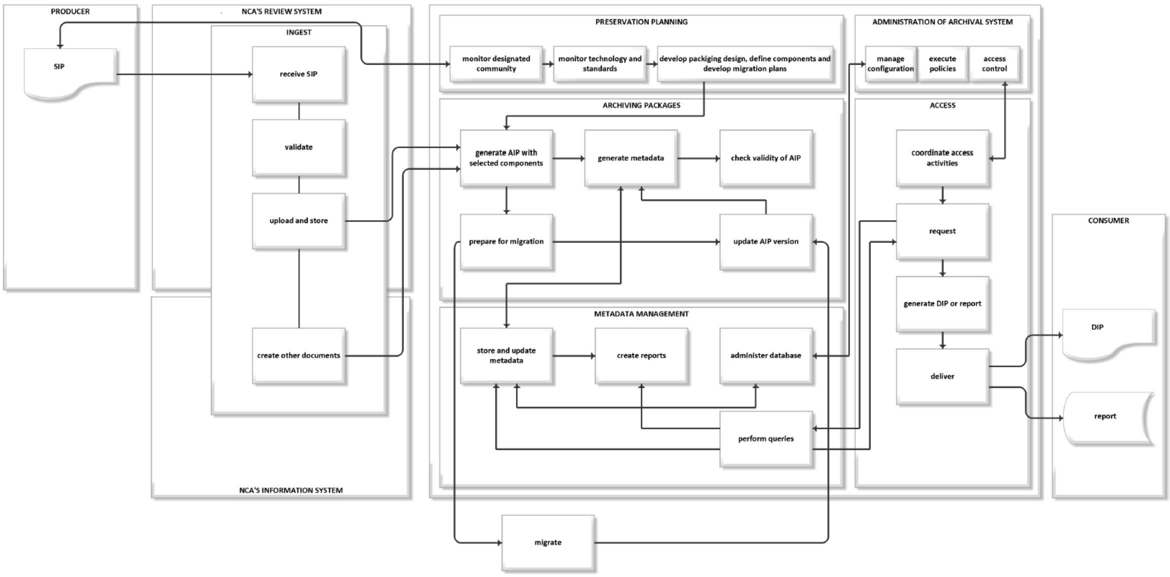
The **administration function** is the central function of OAIS-based system. It communicates with all other functions, monitors the system's behaviour and the environment, manages users' rights and restrictions, provides control and protects the system. In OAIS archive the administration function is also responsible for checking the compatibility of submitted information packages according to agreed specifications or other agreement on the delivery of packages. Mechanism of monitoring the compatibility of submitted information packages with specifications should be considered as an integral part of OAIS-based system; regardless of whether it is a part of the digital archival information system itself or is derived from a third application (EURS is Yours, docuBridge, etc.).

The **preservation planning function** is the most specific part of the OAIS-based system while it enables proactive monitoring of technological and other changes in the environment of the system, other systems with which it communicates, and changes in the standards or tools for creating eSubmissions. Preservation strategies like migration and others may be proposed on that basis. The preservation planning function has other mechanisms as well. It helps in user needs analysis and in monitoring technology in order to extend the lifespan of digital archival information system and replacing outdated modules and applications. Monitoring of standards is a mechanism tasked to monitor relevant standards related to the submitted materials. The mechanism of designing and planning archival information packages has important role in the long-term preservation of the objects in digital archival information systems.

Access function is associated with user needs like search and retrieval facilities as well as user rights and restrictions.

Functional model (Model 1) represents possible model of digital archival information systems for agencies. This model is derived from OAIS ISO standard. Agencies are able to use this architecture, its parts, or to derive its own architectures with implemented functions of Open archival information systems. Although more agencies possess document management system for manipulation of eSubmissions and other digital documents, for archiving purposes it is advisable to upgrade DMS with recordkeeping functions and archival hardware. Good example of combining document and records management functionalities are today's enterprise content management systems (ECMS). Majority of ECMS systems aspire to meet requirements stated in archival standards (OAIS, MoReq etc.).

Model 1: Reference functional model of the digital archival information system



3. Archival information packages

3.1. Information model and components

Basic entity of the OAIS reference model is **information object**. Information in digital archival information system is defined as a combination of data and their representation. Information object is made of **data object** and **representation information** that enables (computer) interpretation of data object.

Data objects could be physical objects or **digital objects**. Without the representation information, data object cannot realise its interpretation potential. Representation information enables interpretation of data object by assigning semantics to the series of bits. **Types of representation information** are **structure information**, **semantic information** and **representational networks**. **Types of information objects** are **content information** (that occurs when representation information interprets content data object), **preservation description information**, **packaging information** and **descriptive information**.

Preservation description information is a set of metadata necessary for understanding content information over long period of time and it is divided into reference information, context information, provenance information and fixity information. Reference information consists of identifiers and references to content information.

Context information is information about content-environment relationship. Provenance information is specific contextual information about the origin of the content and changes of the content since its creation. Fixity information (e.g. checksums) is necessary for establishing and retaining integrity of the content.

Packaging information is a set of metadata provided by the mechanisms that logically or actually attaches components into information package. Descriptive information is additional metadata for current and future usage of the content information.

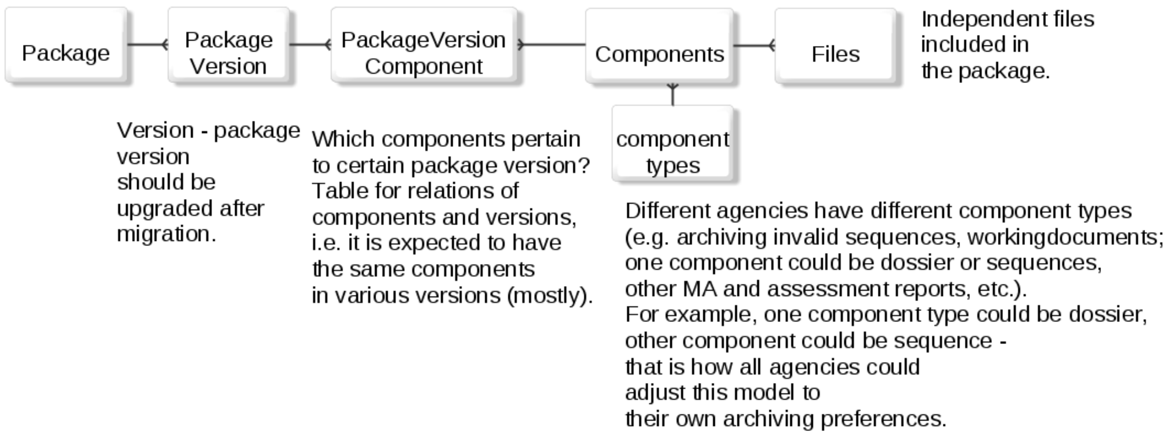
Content information is the basic type of information object and it is primary target of long-term preservation. Content information is to be prepared for long-term preservation in the form of the **information package** that contains content information, preservation description information, and packaging information (and in addition it may contain descriptive information). Information packages are divided into **submission information packages**

(SIP), **archival information packages (AIP)** and **dissemination information packages (DIP)**. The eSubmission provided by the applicant and received by the agency is a submission information package. Although this Best Archival Practice guidance offers recommendations to applicants on preparing quality eSubmissions, focus of this document is providing recommendations for archiving eSubmissions as archival information packages in the agencies.

Archival information package should be seen as archival copy of the eSubmission separately stored according to digital recordkeeping practices. Advice on eSubmissions is contained under part 3 that describes content. Recommendations on archival information packages are provided under parts 2 and 3, i.e. under parts related to archival packages and content.

Dissemination information packages can be any form of content provided to internal or external assessors after content has been archived. Design of the dissemination package is out of scope of this guidance.

Model 2: Archival information packages and components



Independent files included in the package.

Version - package version should be upgraded after migration.

Which components pertain to certain package version? Table for relations of components and versions, i.e. it is expected to have the same components in various versions (mostly).

Different agencies have different component types (e.g. archiving invalid sequences, workingdocuments; one component could be dossier or sequences, other MA and assessment reports, etc.). For example, one component type could be dossier, other component could be sequence - that is how all agencies could adjust this model to their own archiving preferences.

Archival information package is a logically connected set of PDF/A files converted form PDF files of eSubmission and created by the agency upon business process related documents. Recommendation is to archive flattened files bounded into package by using identifiers. Files are logically sampled into components. Component can be files from eSubmission and files created by the agency. Agency can also prescribe other types of components. Packages and components should be versioned by DMS/RMS/ECMS software. It is important to add version metadata after implementation of preservation-related procedures such are conversions and migrations. Integrity of each version should be protected after versioning and authenticity should be guaranteed by the digital archival information system.

One example could be a package as a set of PDF/A files without parent hierarchy of directories and with XML index files that contain identifiers and other metadata related to all files in the package**.**

3.2. Reference metadata

Metadata is added to secure long term preservation of eSubmissons and other documents and to enable future usage of archival information packages. Metadata can be produced during the business processes and/or added when transferred to long-term digital archival information system. This transfer could occur in any moment of time, for example, archival copy of submission information package could be created after validation of eSubmission and files created by the agency could be added later in process. Another approach is to create archival information package and add eSubmission and agency's files after approval of medicinal products. Each metadata category has one or more specified elements that are needed to confirm possibilities for preservation and usability after the documentation has lost its status as active subject in business process.

Here is presented the minimum of metadata needed for each archival information package. Every organization may have more elements of metadata that must be used according to the country legislation, agency's policy or other specific standards and requirements.

Minimal set of metadata required for long-term preservation of archival information packages is divided into four areas which relate to first four areas of General International Standard Archival Description – ISAD(G). ISAD(G) contains several areas of archival description. However, areas of (5) allied materials (5) and description control (7) could be added into digital archival information systems on level higher than the level of archival information packages (allied materials metadata can be added, for example if archival information package is a result of digitization of paper submission, but this case is out of scope of this guidance).

DMS/RMS/ECMS systems or similar functionalities are used for versioning, protecting integrity of packages and components, and for restraining access rights. Version numbers and integrity metadata should be added by the DMS/RMS/ECMS or similar functionalities to metadata of the archival information package. Content and structure related metadata, as well as conditions and access-of-use metadata can be fully managed by the DMS/RMS/ECMS, but these metadata should be added to the package when package is transferred into another digital archive or information system.

1. Identification related metadata
   - integrity - authenticity check metadata and authenticity check date
   - package version – version of archival information package
   - component version – version of component in the package
   - package id - package identifier
   - component id - component identifier
   - component type – type of component
   - component name – name of the component
   - file id - file identifier
   - filename
   - document name
   - file format and version

2. Context related metadata
   - entity of origin (Agency's name)
   - country
   - role – role of the Agency in the procedure
   - procedure number – metadata containing procedure number, number of the authorisation, number taken from the classification system etc.
   - procedure type
   - procedure start date
   - authorisation date
   - applicant
   - medicinal product name

3. Content and structure related metadata
   - retention period
   - retention period end date

4. Conditions and access of use
   - restriction security period (period during which the documentation is classified confidential)

- restriction security period end date (date when classification confidential ends or permanent)
- restriction security reason (explains why restriction to access documentation exists i.e. particular legislation or implementation of internal policy)

6. Notes area
   - notes – notes related to procedures with archival information package

3.3. Authenticity and protection of archival packages

Each modification of the content should be planned, traceable and documented. On the level of archival information packages this means that the version of component and thus the version of package is created. Each version should be protected with mechanisms that can be tested even in system-independent manner. Integrity related data should be easy to archive and maintain, i.e. hash algorithms data etc.

4. Content

There are two uses for the documents submitted in the sequences and dossiers: the first is the eSubmission format or the submission information package and the second format is the long-term archive format or archival information package. Both are closely related. For dissemination information packages for their internal or external assessor agencies can prepare dataset according to the case or provide access to one of the packages.

4.1. Content of eSubmissions

The recommendation to applicants is to use PDF 1.4 (best practice for non-archival file) from the new ICH M2 recommendations (April 2011) and the ICH eCTD 3.2.2 specification, including the usage of ISO 32000-1:2008 specification without its extensions. This means that PDF versions 1.4 to 1.7 are accepted for submission information package. The PDFs should ideally be text searchable (please refer to specific guidance documents where parts are mentioned which need to be text searchable). Security settings are not allowed at all, except for bibliographic references and PDF forms which are downloaded from agencies' sites. These PDF forms carry certain security settings and the applicant is not allowed to change them.

Alternative recommendation is to have archival PDF/A (ISO 19005-1, ISO 19005-2, ISO 19005-3) already contained in submission information package. PDF/A files should be text searchable as referred in specific guidance documents. PDF/A files (PDF/A-1, PDF/A-2) should be created with B level of conformity. PDF/A files can also include hyperlinks with valid targets, inherit zoom and fast web view functionalities. Additional value of using the archival file format (PDF/A) is in creating eSubmissions that are more sustainable. If the applicant decides to use PDF/A, PDF/A files have to be used across entire sequence of eSubmission.

The recommended standard is ISO 32000-1 without extensions and with the following restrictions. PDF files must not contain:

- Javascript
- dynamic content which can include audio, video or special effects and animations, attachments
- 3D content

4.2. Recommendations regarding PDF files in eSubmissions

- Preferably generated from electronic source
- PDF version must be at least 1.4 to at most 1.7. A higher version than PDF 1.7 fails to meet the eCTD and NeeS best practice criteria. The reading of the PDF version is specified in the ISO standard.
- PDF version can be archival PDF/A derived from PDF 1.4 to 1.7 (PDF/A 1 or 2 of preferably B level of compliance, e.g. PDF/A-1B derived from PDF 1.4)
- The paths in the cross-document hyperlinks use forward slashes (as in the ISO specs to allow reading on non-windows equipment)
- For all PDF files use inherit zoom for cross document hyperlinks and bookmarks
- For all PDF files use fast web view or linearised (please note that this increases the document size)
- When bookmarks are present in the PDF files, the settings in the PDF file should allow that the file opens with the bookmark pane open.
- Initial page settings and magnification level are set to default
- The maximum size of a file is 100 MB.

4.3. Fonts used in eSubmissions

Applicant can use base fonts which are the fonts mostly used and recommended by ICH in eCTD specification 3.2.2.

The ISO 32000-1 standard in section 9.6.2.2 addresses the 14 standard fonts which have to be present in an ISO conforming reader. These 14 fonts origin from the original postscript fonts (postscript is the predecessor of PDF). Standard Type 1 Fonts (Standard 14 Fonts) from the ISO 32000-1 specification are as follows: Times-Roman, Helvetica, Courier, Symbol, Times-Bold, Helvetica-Bold, Courier-Bold, ZapfDingbats, Times-Italic, Helvetica-Oblique, Courier-Oblique, Times-BoldItalic, Helvetica-BoldOblique, Courier-BoldOblique. These fonts, or their font metrics and suitable substitution fonts, shall be available to the conforming reader. This means that e.g. Helvetica is substituted by Arial.

Preliminary examination of submissions shows that e.g. for listings the condensed or ‘narrow’ font style for Arial was used. As ‘Arial Narrow’ is a variant from Arial, this leads very likely to problems with the conversion to PDF/A-1 when eSubmission contains PDF files 1.4-1.7 (see further) and the narrow font is missing locally.

The 14 standard fonts are supposed to always be present in a conformant PDF reader and are therefore included in the ICH eCTD v3.2.2 specifications. These 14 base fonts are to be used and embedded. The Adobe Reader covers the license of the 14 base fonts.

Alternative recommendation is to use non-copyrighted fonts and to embed them in PDF files (it is necessary to do full embedding, and not just subset embedding). This ensures long-term sustainability of files in eSubmissions.

Font embedding

PDF format in general is designed such that the content is viewed and printed as the author meant it to be read, over time. When the author uses a text processor using certain fonts, references to those fonts end up in the PDF. When a reader does not have exactly these fonts, the fonts are substituted by those which resemble those referred to in the PDFs.

When one uses a Microsoft text processor on a Microsoft operating system to prepare a document with the Arial font which is then converted to PDF, the Adobe PDF converter replaces the Microsoft font with the Monotype font. This can be seen with file –properties – fonts. Fortunately, the Arial of Monotype has the same appearance as the Microsoft font.

When one copies text from the PDF into e.g. an assessment report in a Microsoft text processor on a Microsoft operating system, the font is again substituted with the Microsoft font. In these cases, substitution always occurs, but to assure the PDF can be archived as the writer intended, it is safe to embed all fonts used.

As part of the ISO 32000-1 standard, each PDF reader is supposed to include the base fonts mentioned above. Therefore, ICH M2 recommends including only these fonts. The implied risk is that the PDF document does not appear on screen or on paper identical how the author intended. This continuous substitution which is hidden from the reader undermines document authenticity which could be prevented by embedding fonts in inside the PDF. The advantage is that the PDF files remain readable over time and over different systems.

There are three disadvantages or items to consider: (i) silent font substitution, (ii) copyright issues and (iii) increase of PDF size. The adobe reader must be configured in a manner that the embedded font is used when PDF contains embedded fonts. The reader can be wrongly configured – it can always perform substitutions, even when fonts are embedded.

To minimise copyright related risks in the future, we recommend using non-copyrighted fonts that are metrically the same as the fonts that are used by most text processors (i.e. Arial, Times New Roman, Courier).

The usage of fonts in Europe is different from the US. In the US many fonts are copyrighted but it is allowed that these are embedded, whereas in Europe it depends on the license which goes with the font. When one reads the Adobe Professional license, font embedding of the Adobe fonts is permitted, provided the reader of the PDF has a valid license to read these PDFs with the fonts embedded. The license specifies which functionality (reading, editing, printing, etc) is allowed. Extraction of the font files from these PDFs to further use is not covered by the license.

Each font file has a flag to indicate whether the font can be embedded, but this is rather a technical flag and the Monotype website clearly states that the user of the font has to consult the license which goes with the font and not to rely on the flag on the font file.

Recommendation for the eSubmission format regarding fonts - the author of the PDF is advised to only include the 14 standard fonts (as specified in ICH eCTD 3.2.2) in the PDFs or to embed non-copyrighted fonts that are metrically the same as standard fonts. The author should ensure that embedded fonts do not imply a license problem for the reader of the PDF and the simplest way to that is to use non-copyrighted fonts.

The third disadvantage concerns the increase in file size with embedding of the fonts. Experience has shown that this increase is moderate when standard fonts are embedded.

4.4. Recommendations regarding PDF files in archival information packages

The PDF format is already designed so that it is backwards compatible, regardless of the version number. This means that the old PDF files will look the same as when created on more recent viewers. For long-term archiving, however, this is not enough, and the PDF/A format as specified in the ISO 19005-1:2005, ISO 19005-2:2011 (and ISO 19005-3:2012, under development) standards was developed. This is now the industry standard for archiving of electronic records. PDF/A-1 is based upon PDF v.1.4.

There are 2 sub classes in the PDF/A format.

- Level B: PDF/A level B specifies that all fonts must be embedded, security is not allowed, audio, video and java script neither and executable file launches are also not allowed.

- Level A: Level A goes further than level B, where now the PDF must be tagged. Tagging here means that content extraction of the PDFs is facilitated. Other features of level A conformity are: structure tree, language specifications, and mapping to Unicode. Level A is more user-friendly for visually impaired users.

PDF/A2, which is also an ISO standard (ISO 19005-2:2011) is based upon ISO 32000-1:2008. PDF/A2 is rather new and currently not yet widely used.

The recommendation for agencies is to use any version of PDF/A with level B of conformity for archiving purposes.

In general, the PDF/A format specifies that:

- Audio and video content are forbidden
- JavaScript and executable file launches are forbidden
- All fonts must be embedded as well as legally embeddable for unlimited, universal rendering. This also applies to the so-called PostScript standard fonts such as Times or Helvetica. It is also important that no subsets of the fonts are embedded, but the complete font file
- Colour spaces specified in a device-independent manner
- Encryption is forbidden
- Use of standards-based metadata is mandatory
- External content references are forbidden. This means that the PDF may not refer to external content to display the PDF, such as external images. Please note that cross document hyperlinks remain fully possible
- LZW and JPEG2000 image compressions are forbidden in PDF/A-1, but JPEG 2000 compression is allowed in PDF/A-2

- Transparent objects and layers (Optional Content Groups) are forbidden in PDF/A-1, but they are supported in PDF/A-2
- Provisions for digital signatures in accordance with the PAdES (PDF Advanced Electronic Signatures) standard are supported in PDF/A-2
- Embedded files are forbidden in PDF/A-1, but PDF/A-2 offers the possibility to embed PDF/A files, allowing archiving of sets of documents in a single file

In PDF, security settings are achieved by encryption. As encryption is not allowed, there cannot be any security settings in a PDF/A file.

Most recent Adobe PDF Readers allow to view PDF/A as PDF/A's, where the hyperlinks are not functioning while the hyperlinks are present in the PDF/A file. Therefore, the reader has also an option as specified in the ISO standard for PDF/A to read a PDF/A as a classical PDF. In the latter case, the hyperlinks do work. For more reference the user is advised to read the PDF/A standard (ISO 19005-1) section 6.6.3.

Archival requirement is to use PDF/A (B level) for archival packages of eSubmissions.

4.5. Fonts in archival information packages

Fonts set some constraints on the authoring process. Font embedding (all fonts used, without the subset) leads to PDF/A files which tend to be larger than the source PDF files. The license for the fonts to be embedded must allow reading and archiving without the requirement that the reader must verify that he or she has the proper licenses for the fonts.

Conclusion - the ISO 19005-1 format or newer ISO 19005-2:2011 is recommended for archiving and applicants are advised to anticipate this with embedding of fonts which do not pose copyright issues.

5. Preservation responsibilities

Both the European Medicines Agency after Centralised Procedure and Referent Member State in Mutual Recognition Procedure and Decentralised Procedure procedures should keep submissions as archival information packages according to the life cycle of the medicinal product (at least), then it should notify involved agencies and Concerned Member States about the final destination of packages (destruction/other), and, finally, it should offer possibility of transfer of packages to their digital archival information systems or other systems. Interoperability should be ensured by using common minimal set of metadata and standardised PDF files.