EUCROF
European CRO Federation

eClinical forum
A NETWORK POWERED BY PEERS

Data Formats Used in Clinical Trials

A white paper written by a joint task force from
the European CRO Federation and the eClinical Forum

Version PR1

April 2023

Document History

Version	Author	Date	Changes
PR1	Joint EUCROF and eCF Task Force	2023-02-23	Initial Issue

Table of Contents

About EUCROF

The European Contract Research Organisation Federation (EUCROF) consists of members from most European countries and partner members from nearby countries with the aim of promoting clinical research of high quality in Europe in general and in the European Union in particular. EUCROFs objectives include supporting discussions with European bodies (EMA/EU Commission), promoting discussions on selected topics with representatives of the pharmaceutical industry to enhance business relations and identify common concerns, and endeavouring to develop transcontinental relationships with other associations e.g., with ACRO (Association of Clinical Research Organisations) in the USA and JCROA (Japanese Clinical Research Organization Association) in Japan. For further information visit the website at www.eucrof.eu.

About the eClinical Forum

The eClinical Forum (eCF) is a global, technology independent group representing members of industries engaged in clinical research. The eClinical Forum's mission is to serve these industries by focusing on those systems, processes and roles relevant to electronic capture, handling, and submission of clinical trial data. The eClinical Forum has sought out opportunities to promote electronic Clinical Trials since its inception in 2000. The cross-industry forum has a broad view of research with members - Sponsors, Contract Research Organizations (CROs), Technology vendors (both clinical research and healthcare), Academia, and Investigators - and with invited outreach opportunities with global Regulatory representatives. For further information visit the website at www.eclinicalforum.org.

Disclaimer and License

The information presented in these works draws upon the combined current understanding and knowledge of EUCROF and the eClinical Forum on this topic and is provided as an aid to understanding the environment for electronic clinical research. The opinions of the author(s), EUCROF and the eClinical Forum do not necessarily reflect the position of individual companies. Users should assess the content and opinions in the light of their own knowledge, needs and experience as well as interpretation of relevant guidance and regulations.

For additional Disclaimer and License information, see Appendix 3.

1 Executive Summary

There are disparate views in our current climate on how trial data and documentation should be retained. This follows regulatory guidance that expects trial data to be maintained in a way that allows evaluation whilst also accepting that such formats are not necessarily practical over the full retention lifetime. This white paper explains the options, risks, and benefits of using a flexible multiformat approach to ensure trial data can continue to be accessed in the future.

Consideration is given to the use of proprietary formats vs open-source formats, and the possibility of concurrently maintaining trial data in different formats to aid with retrieval in the future. This white paper follows the publication of the Archiving Position Paper. It takes a focused view into the regulatory environment and best practices to ensure that the usage of common format files can be preserved and those data that were generated in live phases of life science research are retained and accessible for future regulatory interactions. The aim is to ensure that archived data can be retrieved and utilized in the best way, without the loss of provenance, format, metadata, context, or meaning.

2 Introduction

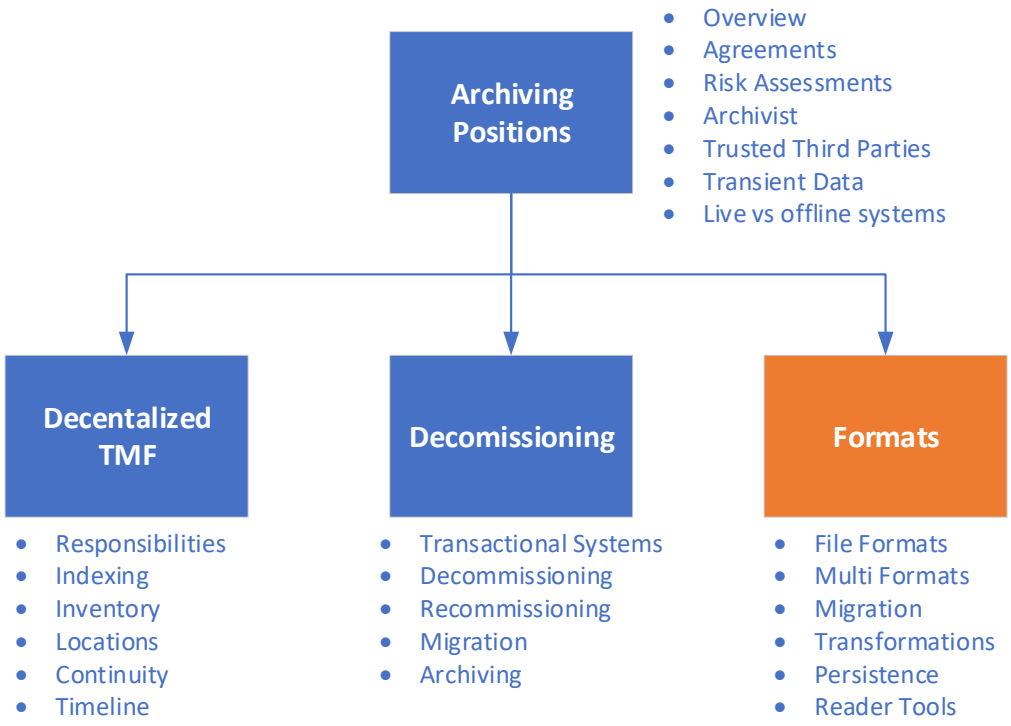
Currently there are multiple standards in use for the preservation of data, and the format of said information (please refer to appendix 1 for an indicative list). Electronic documentation in use today comes in proprietary (e.g. .docx) and open source (Open Office XML) formats that comply to different standards. One such format is ISO/IEC 29500-1:2016 that describes the processing language for Office Open XML formats and Markup Reference Languages. There are other formats such as PDF/A (ISO 19005-4:2020) that is utilized for the long-term archiving and storage of document formats.

Having a plethora of options for long term storage and reconciliation of the documentation styles leads to a problem, which to choose and when to apply the choice. If the choice is not optional at the time of archiving, decisions would have to be made on how to change course and take another option in the future. This is based on the technology that reads the files, and fundamentally, manpower based in the processes and procedures defined by which information preservation is conducted.

This white paper aims to provide thoughts and strategies for document preservation and the attributes of a system by which long term preservation can be achieved. There is a risk-based approach to the “uncertainty of future events, tools and formats” used to be able to read and gain information from underlying metadata in these scenarios. There is no magic bullet that can be applied to all file types, and a certain amount of data transformation loss needs to be accepted if electronic files are maintained past the expiry of their format. It is wise to review national archiving efforts (Appendix 2) and their preferred formats to ensure that the latest recommendations are being followed.

The ability to potentially transform mass amounts of data files and return them to a format that may not be the native format but where the content, format and metadata are preserved, is the goal of today’s technology solutions.

This document is a data/records/documents focus to the position paper ‘*Trial Master File Archiving and the Decommissioning of Computerised Systems Used in Clinical Trials*’ written by a joint task force from the European CRO Federation and the eClinical Forum and published on 2021-02-24. It is intended to provide hands-on practical guidance on the archive formats of clinical trial data and records created by describing examples of risks in practice with proposed recommendation(s).



3 Sponsor responsibilities

There are constant challenges for sponsor organizations regarding the default period to retain information. EU Regulation 536/2014 article 58 states a minimum trial master file retention period of 25 years. During that time, there will be significant improvements and changes in hardware and software technologies that create, modify, store, and retrieve data. These changes may make the current formats obsolete, or successive updates to the proprietary format may change the file rendition and make it impossible to reconcile the differences compared to past versions of the document.

The Sponsor is also responsible for storing (and archiving) disparate formats that only relate to proprietary software such as SAS data files. Native data files also exist that could pertain to any application creating a data file whose format has not been defined. The application using this format could range from pharmacovigilance databases to inhouse database applications made from commonly available tools, such as Microsoft Access. In this instance the name of the file's extension will not determine the application that will open the file and there would need to be additional information as to the provenance of those data held within the file.

Current common static formats being used for archiving, such as PDF/A (ISO 19005-4), are an applicable approach for storage of information for long term preservation. There are thoughts that this format will exist for many years and that software readers for the format will be available in the future. This will avoid the need to change the original file structure as the reader tool will just "read" with its core functionality and display the file type. Adobe's Acrobat Reader, among others, uses this level of provenance, being a freeware application to render all versions of Portable Document Format (PDF) files.

ICH GCP E6 R2 defines retention from the Clinical Investigator and the Sponsor as:

Section 4.9.5 - *Essential documents should be retained until at least 2 years after the last approval of a marketing application in an ICH region and until there are no pending or contemplated marketing applications in an ICH region or at least 2 years have elapsed since the formal discontinuation of clinical development of the investigational product. These documents should be retained for a longer period however if required by the applicable regulatory requirements or by an agreement with the sponsor. It is the responsibility of the sponsor to inform the investigator/institution as to when these documents no longer need to be retained (see 5.5.12).*

Section 5.5.3 h - *Ensure the integrity of the data including any data that describe the context, content, and structure. This is particularly important when making changes to the computerized systems, such as software upgrades or migration of data.*

Section 5.5.11 - *The sponsor specific essential documents should be retained until at least 2 years after the last approval of a marketing application in an ICH region and until there are no pending or contemplated marketing applications in an ICH region or at least 2 years have elapsed since the formal discontinuation of clinical development of the investigational product. These documents should be retained for a longer period however if required by the applicable regulatory requirement(s) or if needed by the sponsor.*

In further guidance from EMA 'Guideline on computerized systems and electronic data in clinical trials', states that:

Section 4.1 - *Data governance should address data ownership and responsibility throughout the life cycle, and consider the design, operation and monitoring of processes/systems to comply with the principles of data integrity including control over intentional and unintentional changes to data.*

Section 6.10 - *It should be verified that the file remains accessible and depending on the media used for storage and available software through the retention period. This could imply migration of data for example*

Suitable archiving systems should be in place to safeguard the data integrity for the periods established by the regulatory requirements including those in any of the regions where the data may be used for regulatory submissions, and not just those of the country where the data are generated.

Source documents and data should always be available when needed allowing authorized individuals to meet their regulatory obligations.

Data should be maintained in a secure manner and should only be transferred between different locations in a validated process. Data should be archived in a read only state.

Taking the regulations and guidance into account Sponsors have a high degree of responsibility as to the disposition of those data and the regulators maintain that they may want to have those data available in host systems or be able to manipulate those data as if they were live, if they are not held in host systems.

This poses challenges for access to the data and ensuring that they are in systems that are 'malleable enough' to allow for manipulation even if a reader tool or secondary system does not contain the full functionality of the original system.

4 Reader tools Vs fully functional applications

Once an application has been replaced in the market, the only real option for reading those data produced in past cases, would be a reader tool application. The company could consider keeping defunct versions of software alive for this purpose, but this brings complexities that most industries would prefer to avoid, the biggest of which would be the security risks of those software and the lack of security updates on connected systems allowing the potential for unwanted excursions into their systems.

If aged software is kept alive, there is also the matter of the hardware needed to run the software over an extended period up to and beyond the 25-year limit set by the EU Clinical Trial Regulation. This is an intrinsic factor influencing the monetary and headcount cost, and reproducibility of those data if the native applications are kept alive. Product support is typically not offered past the end of service date. Continued support may be available from the provider for a cost or would have to be handled in house by the company itself.

A piece of software that is designed to read the aged file format would be a preferred option, however, this also comes with the cost of producing, validating and updating when needed. This would also require the buy in from global regulatory authorities as per EMA (Guideline on the content, management, and archiving of the clinical trial master file, Ch 6) as this software would not be the original software and only used to read the data. Potentially manipulation of those data would be limited or restricted.

The reader tool has the advantage of being totally adjunct to the file type and the original software so could be stand alone, ensuring that those essential files remain in the same format with the same data arrangement, and the same way of organising the metadata, therefore retaining the originality of the file itself, which could be an advantage during regulatory review and security.

5 Multiple format approaches

The use of multiple data formats when archiving is an alternative approach to document and data preservation that spreads the risk amongst several file formats, with the resulting risk minimization that not all formats will become obsolete within the archival period.

This approach greatly increases the chances that the files archived and stored for long durations will be able to be read if needed at a future date.

There should be a strategy for the use of multiple formats. This strategy would ideally define a number of essential formats that should be supported when archiving all trials, together with the encouragement to include additional file formats supported by the systems used for any given trial. The essential formats should include those static and dynamic formats currently requested by the regulatory authorities but need not be limited to those formats, indeed it would be advantageous from a risk minimization point of view to include additional file formats.

6 Risk assessments

6.1 The challenge of formats

The risk assessment determines how the company will deal with and assess the mechanisms by which the files will be archived for a long period of time. One of the most important risks to assess will be the formats of those data, and the choice of single format, multiformat or changing the format from a proprietary to a generic format.

Currently, there is no guidance on how or when these decisions should be taken or advice on the most prudent choice in this area. Below are aspects that should be considered when choosing formats for archiving:

6.2 Single format

To choose a single format, perhaps because of the inability to transform or change the format of the file to any other for the readability or the provenance of raw data files, presents a significant risk. The choice of a single format should not be taken lightly as a number of constraints have to be put in place to ensure that the files have the greatest possibility of being read in the future. This includes ensuring validated tools for viewing of the file format are in place and that these tools will be upgraded on successive operating platforms. There must also be support for the format and the viewing tools moving forward.

If support for the chosen single format should diminish in the future the organisation is then left having to choose between:

1. Sourcing (and maintaining) a validated viewing tool for the original format
2. Changing the format
3. Migrating to a new format

These different options are described elsewhere in this white paper.

6.3 Multiple formats

One of the safest options is to archive the data in multiple formats, including the metadata. The formats archived should contain those readable by commonly used applications, such as XML. Opting for plugins for commonly used browsers with backward viewing capabilities could be easier than choosing a bespoke, format-dependent viewing tool that would require updating on a periodic basis and whose long-term compatibility with COTS Operational Systems cannot be guaranteed.

If multiple formats are archived these should include both static and dynamic formats and contain all metadata. The static data sets could be used to prove that the data in the dynamic format files have not been altered in any way. The archive should contain an index and be arranged in a way that is easily understandable to the reader. It should be apparent what is contained within the files and what format they are in. It should also be apparent if they are in a static or dynamic format.

The choice of using multiple formats comes down to the cost for producing those formats and the required storage area. The cost of producing multiple formats can be minimized by simply using those formats offered by the computerized systems being used in the trial, usually at no additional cost. With regards to storage area requirements, most organisations already archive using the PDF/A format, which is currently the file format that requires the largest amount of storage area. Indeed, a multitude of other file formats can be added to such an archive without greatly increasing the overall storage area requirements, so this is not seen as a limitation.

The use of multiple file formats is a risk mitigation in itself. When an individual format becomes obsolete you rely on the ones you have left in the archive.

If a large number of the formats used in the archive should become obsolete during the retention period of the archive, then you may end up in the same position as if the archive only contained a single format, with the same possible mitigations as described above. But this is seen as unlikely, and less so for each additional format added to the archive.

6.4 Change of format

Changing the format of the file should not be taken lightly as there are several consequences that need to be taken into consideration, the first being if the format of the file needs to be changed more than once because of ageing. The second consideration would be the retention of the original file, even if it cannot be read anymore. Is there any value saving it? Could it be used in the scenario where the transformed file has been questioned? Would this require a reader tool to be developed or planned to be developed in this instance? The third is loss of provenance due to the difficulty in ensuring that the associated metadata such as electronic signature(s) remain authenticatable.

There could be scenarios where certain parts of the file format would require updating, but not the full specification of the file would be changed, such as a different way of ordering the metadata in the file, without affecting the data stored in the file. This may be a very rare occurrence, but consideration would need to be taken whether this is a new file type or an augmented file type and how this would be documented for regulatory scrutiny in future years.

When a format is changed, the risk-based decision to what format must be taken carefully, e.g., would data be transformed to CDISC SDTM format in an XML frame, knowing that the viewing capabilities of the transformed file would not exactly be a replica of what was seen with the original rendition. While it has to be ensured, that format changes don't modify or delete any data required for the reconstruction of trial (related) events, the risk assessment should clearly assess any unavoidable modifications of metadata or the dynamic characteristic/context of the data and justify, why these modifications are acceptable. There would need to be guidance from Regulators on the acceptability of this proposal and whether they would treat this data as per the original source data rather than having to request that the original source data be available.

Format changes can be a resource intensive process. This implies the need for ongoing headcount from the company. This should also be codified into approved processes with appropriate oversight to ensure that the relevant stakeholders are aware of the situation with their most precious asset.

6.5 Data Persistence

Some data may have been generated using methods and equipment that in the current market have been deemed no longer fit for purpose or have been decommissioned and the retention of those data needs to be carefully assessed and if retained indicate its ancestry in the active dataset for the product.

Data persistence is achieved by moving completed trial files and datasets from an active state to offline locations. This housekeeping effort is done for practical data management reasons. It is understood that this information, whilst required as evidence of the original analysis, could be useful for future analysis methods which could extract more value from the old data. The choice of formats for this datum is important to ensure that it is useable throughout its lifecycle.

This needs to be considered when producing a risk assessment for those file types, and the eventual effect of those decisions on the retained data sets should also be considered alongside the file types that those data are stored in.

6.6 Assessing the Risk

Risk can be defined as three dimensional and as per common practices considering the impact, likelihood and rate of detection of the risk. As defined in ISO 31000, risk is the effect of uncertainty on events. In this case time is the greatest uncertainty. The likelihood of changes to the formats of those data over time needs to be carefully evaluated on a file type basis and documented from a company perspective. As many proprietary file types are used, these are the ones at the highest end of the risk scale. This is due to the uncertainty of future development of the file type and how those file types would be accessed when the program/equipment is mothballed.

Risk should be documented and evaluated using the same parameters to ensure consistency of approach. This paper will not go into detail of the many ways that risk can be evaluated. The standpoint of the authors is that this should be performed by one or more qualified persons, who have knowledge of both the clinical/technical subject matter and deep understanding of how risk analysis works including the negative effects of a poor risk analysis on outcomes.

Consideration needs to be taken into how long the files need to be kept in an active, semi-active or dormant state as to the file format and the risk of transforming the files instead of using a reader tool from a third party that can be updated instead of updating the file itself.

7 Migration

A big decision is when to migrate data into a new format under the guidance of the user group looking after the data provenance usage of defunct applications.

Migration is a huge effort and is not to be taken lightly as this will affect the company's entire data set and will also set an industry wide trend on transforming data that is being retained for long periods of time, instead of producing reader tools to access those data.

Migration needs to be clearly defined and the path of migrating those file types needs to be traceable, so it is clear of any datapoint mappings. The issue also arises whether the original files are retained alongside the transformed files, this is something that would have to be decided, and their risk appetite for future challenges of any transformed data.

When migrating data to new formats, the interaction of those data with other potential datasets needs to be evaluated. This interplay between the original and target formats would require addition to the table concerning the change of format and any effect that would occur to those data during that change and the residual effect on other data sets. This includes whether a combined analysis of those data could still be performed or whether they would have to be displayed in different ways.

8 Conclusion

There are regulatory expectations that the provenance of clinical trial data evidence be maintained for the retention period. This needs to be done in a realistic and efficient manner. It is the recommendation of this white paper that the clinical trial data and metadata should be maintained in multiple formats as opposed to maintaining the original system which is not seen as a viable alternative. This mitigates the risks of individual formats being rendered obsolete by ensuring that the data and metadata are still available in other formats.

Future scenarios cannot be foretold so a broad ranging approach is prudent. It is worth noting that any change of archiving formats after the end of the study will be resource intensive, with no guarantee that the regulator will regard the resultant information as authentic.

It is essential for the company to conduct a risk-based analysis of all file types in use, taking into consideration the future reliability of the software tools that produced those files. Ensure that if those files are proprietary, and are at risk of becoming unsupported, then a remediation action is taken to ensure their survival.

Overall, the company should ensure that an active inventory is taken of its file formats and that sufficient effort is put into the protection of those files that could be needed in the future.

9 About the Authors

This white paper was authored by the EUCROF and eClinical Forum Joint Task Force on Archiving and Decommissioning.

You can contact the task force via the EUCROF and eClinical Forum websites (www.eucrof.eu and www.eclinicalforum.org) if you would like more information or if you have any comments on the contents of this white paper.

Although the Task Force was initiated as a joint effort by EUCROF and the eClinical Forum, team members representing the following organisations have also participated in the authoring of this white paper:

- ECRIN - https://ecrin.org/
- The ePRO Consortium - https://c-path.org/programs/eproc/
- Medicines for Europe - https://www.medicinesforeurope.com/
- RQA - https://www.therqa.com/

We would also like to thank the many organisations and individuals who reviewed and commented on this white paper before we released it.

10 Appendix 1: Open formats

This appendix provides an overview and starting point from the domain user perspective and describes open formats, which are pertinent to the clinical research arena.

In addition, this appendix will give some comments on formats, which are declared as open, but may have some limitations.

Without further explanation, a detailed list of other, actual open formats, which are not listed before can be found here

https://en.wikipedia.org/wiki/List_of_open_formats

This Wikipedia article describes further open formats which could be for further interest for the archival of files for clinical trials and not listed before. The article is an entry point with further links to describe formats for text, graphic, audio and video formats, but also technical formats for compression and encryption. This article should periodically be revisited as new open formats may be adopted on an ongoing basis.

10.1 Core Technological Standards

ASCII, ISO 8859 and Unicode

A technical description of ASCII, ANSI and UTF-8 is described here in detail:

https://en.wikipedia.org/wiki/ASCII

https://en.wikipedia.org/wiki/ISO/IEC_8859

https://en.wikipedia.org/wiki/Unicode

https://en.wikipedia.org/wiki/UTF-8

From an archival perspective, the above character encoding standards represent the “lowest common unit” or “lowest common character definition” for text formats. These standards are documented and free from licensing obligations as they represent the foundational basis of computer character description. Whilst originally based on the English language these standards also include amongst others Chinese, Japanese and Korean characters.

With this definition, every word or every number is just a collection of one or multiple character encodings.

For further reference in this document, we use the term ‘ASCII file’ meaning that this is a simple text file based on the definition above.

Every so called ‘text editor’ (example Notepad on Windows, Vi on Unix systems) should be able to create, edit and save ‘ASCII files’.

CSV and structured files

Comma Separated Values (CSV) is a collection of ASCII characters with a “separator” to separate data structures. The first version used the ‘comma’, where today CSV files can have any characters as a separator. Often a “|” or a semicolon “;” is used as well. Usually, one line contains one record, however, there are versions of CSV, where the first line names the column of a table.

Therefore, CSV is just a structured “ASCII file”, based on the ‘free’ ASCII definition.

https://en.wikipedia.org/wiki/Comma-separated_values

Other domains, such as genomics also use structured ASCII characters. For example, FASTA is a text-based format for describing nucleotide sequences and amino acids.

https://en.wikipedia.org/wiki/FASTA_format

10.2 Markup Language,

Extended Markup Language (XML) is based on Structured General Markup Language (SGML), which itself has been descended from IBM’s General Markup Language (GML) in 1960’s. These are based on ASCII itself and add structural definitions.

With these types of file types, structural validation of the data file is possible. For XML, the “.xsd” format (XML Schema Description) exists as the metadata description, which describes the structure of an XML file and, is written in XML itself.

As another example, HTML file formats are also a markup language.

https://en.wikipedia.org/wiki/Standard_Generalized_Markup_Language

https://en.wikipedia.org/wiki/XML

https://en.wikipedia.org/wiki/Markup_language#HTML

10.3 Clinical Data Interchange Standards Consortium (CDISC) Standards

CDISC provides foundational standards to support clinical research processes from end to end. These standards focus detail various models, domains and specifications for data representation and include:

CDASH - Clinical Data Acquisition Standards Harmonization

SDTM – Study data Tabulation Model

AdaM – Analysis Dataset Model

ELDF – ESdat Electronic Lab Data Format

ODM – CDISC Operational Data Model

SEND - Standard for Exchange of Nonclinical Data

Besides such foundational standards CDISC has also defined technical data interchange standards like ODM (Operational Data Model), Define-XML and Dataset-XML. Most of them are based on XML and therefore candidates to be used the archival of clinical data.

The CDISC ODM format specifically offers an archival attribute to support the archival of clinical data. An archival ODM file provides a full transactional history of all inserts, updates and deletes of clinical data points up to a selected time point.

CDISC standards cover a wide range of clinical data/meta-data albeit limitations do exist when describing layout / form definitions and business rule implementation. ODM does incorporate the ability to address vendor specific data requirements in the form of extensions to the standard. The extensions only permit data to be handled in particular ways according to the features in proprietary electronic trial systems. Extensions offer no way to display or handle data outside of proprietary software.

For further reference see www.cdisc.org.

10.4 DICOM

Digital Imaging and Communications in Medicine (DICOM) is the standard for the communication of medical images and associated meta-data. Typically, DICOM permits integration between scanning machines, workstations and network devices.

https://en.wikipedia.org/wiki/DICOM (ISO 12052:2017)

10.5 Document Standards

Word processing software from various vendors now provide the ability to store documents using an open standard such as PDF, Libre/Open Office and Microsoft Office formats. Open standards enable free or licensed software to access documents which reduces the dependency on specific applications and can provide backwards compatibility to older applications.

If possible, PDF should embed the used fonts to avoid font substitution issues.

PDF/A is the ISO standardized version of PDF.

https://en.wikipedia.org/wiki/OpenDocument (ODF) (ISO/IEC 26300-1:2015)

https://en.wikipedia.org/wiki/Office_Open_XML (ISO/IEC 29500-1:2016)

https://en.wikipedia.org/wiki/PDF (ISO 32000-2:2020)

https://en.wikipedia.org/wiki/PDF/A (ISO 19005-4:2020)

10.6 Email Standards

Email provides a convenient method to exchange clinical trial correspondence. It is historically used to show intent. The volume of messages also adds complexity to the long-term retention of all correspondence. Email should not be used for significant financial, commitment, or protocol decisions. This information should be saved as a further document in the trial information management records. Email messages themselves may be saved in non-proprietary format such as ASCII, however loss of functionality may result. EML is a file extension for a file in the MIME RFC 822 format which can include ASCII, hyperlinks and attachments. If the email body is written in HTML, then HTML text should be archived.

https://www.w3.org/Protocols/rfc822/

10.7 Image Standards

In addition to the DICOM standards for data originating from medical scanners, smaller images (such as medical photographs) may be stored in a range of open standard formats. These fall into three types of vector, compressed and uncompressed. Examples of file formats include:

Vector	Compressed	Uncompressed
SVG	JPEG	BMP
	GIF	
	TIFF	

https://en.wikipedia.org/wiki/List_of_open_formats#Imaging

10.8 Video Standards

Video files may be stored in a variety of formats using compression technology to limit file size. Video files typically utilize a container within which is contained both video data and an associated audio track. File name extensions determine which programs may open the file and which coder – decoder (codec) will be used to access the audio and video data streams.

https://en.wikipedia.org/wiki/List_of_open_formats#Video

10.9 Audio Standards

Audio files may be uncompressed, compressed with no loss of fidelity or compressed with some fidelity loss. Some formats are proprietary and may be linked to digital rights management. Examples of open standard formats include:

Uncompressed	Lossless Compressed	Lossy Compressed
WAV*	FLAC	MP3
AIFF*		AAC

*Both formats are documented, but not listed as free and open.

https://en.wikipedia.org/wiki/List_of_open_formats#Audio

10.10 Biostatistics standards

Within biostatistics file types are generally of two types:

1. Executable (statistical) programs
2. Statistical analysis results

Technically a “program” is written in ASCII characters and stored as an ASCII file. This means, the source code of any program can be seen as a normal ASCII file. Computer compilers (ex. C Language) or interpreters (ex. Java, Ruby) will use these files and translate them into bytecode, which can be executed on the computer.

Therefore, all programs used for calculations and data transformations should be archived as ASCII files, if possible. This reserves the option for re-compilation if needed.

Statistical programs usually written the same way and then executed in a proprietary environment (e.g., SAS) or in a non-proprietary environment (e.g., R open-source). ASCII is a preferred format for several regulators.

The results of statistical analysis may be output to formats specifically intended for use by enterprise statistical software.

SAS has published the SAS Transport format, which can be used for archival. Regulatory bodies accept the XPT format as part of the drug submission and marketing process.

SAS Version 5 Transport format

https://www.loc.gov/preservation/digital/formats/fdd/fdd000466.shtml

http://support.sas.com/techsup/technote/ts140.pdf

SAS Version 8 and 9 Transport format

https://www.loc.gov/preservation/digital/formats/fdd/fdd000467.shtml

http://support.sas.com/techsup/technote/ts140_2.pdf

Other formats for storage of analytical result can be any open format. Usually, the idea is to have it in a structured and table like format like in CSV or CDISC Dataset-XML.

11 Appendix 2: Global Digital Preservation efforts

This appendix provides a listing of global archiving organization and agency efforts to archive government and organization information. The links may change overtime but a keyword search should find the updated link.

- Digital Preservation Coalition - https://www.dpconline.org/
- International Archival Organisations https://www2.archivists.org/assoc-orgs/i_a_o
- International Council on Archives - https://www.ica.org/en/digital-records-and-archives-management
- EU - https://ec.europa.eu/cefdigital
- USA - https://www.archives.gov/preservation/electronic-records
- Canada Archive Rules http://www.cdncouncilarchives.ca/archdesrules.html
- China National Archive Administration https://www.saac.gov.cn/
- UK - https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/guidance/
- Scotland - https://www.nrscotland.gov.uk/record-keeping/electronic-records-management
- Brazil - https://www.gov.br/arquivonacional
- Australia - https://www.naa.gov.au/information-management/storing-and-preserving-information/preserving-information/born-digital-file-format-standards
- Switzerland - https://www.bar.admin.ch/bar/en/home/archiving/digital-documents.html
- ETSI – TS103-643 V1.1.1 Techniques for assurance of digital material used in legal proceedings https://www.etsi.org/deliver/etsi_ts/103600_103699/103643/01.01.01_60/ts_103643v010101p.pdf
- Sedona Principles - https://thesedonaconference.org/sites/default/files/publications/The%20Sedona%20Principles%20Third%20Edition.19TSCJ1.pdf

12 Appendix 3 Disclaimer and License for the Fair Use of these Materials

This work is the property of the eClinical Forum and is released under a Creative Commons license for non-commercial use. Under the terms of the license, you are free:

- to Share: to copy, distribute and transmit the work
- to Remix: to adapt the work --- Under the following conditions:
- Attribution: You must attribute the work to EUCROF and the eClinical Forum (but not in any way that suggests that EUCROF or the eClinical Forum endorses you or your use of the work).
- Immutable: Wording of the Positions must remain exactly as published
- Noncommercial: You may not use this work for commercial purposes without express license agreement with EUCROF or the eClinical Forum.
- Share Alike: If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

With the understanding that:

- Waiver: Any of the above conditions can be waived if you get permission from EUCROF and the eClinical Forum.
- Public Domain: Where the work or any of its elements is in the public domain under applicable law, that status is in no way affected by the license.
- Other Rights: In no way are any of the following rights affected by the license:
  - Your fair dealing or fair use rights, or other applicable copyright exceptions and limitations.
  - The author's moral rights.
  - Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

Notice: For any reuse or distribution, you must make clear to others the license terms of this work.