Arkivum

Archiving and Preserving Clinical Trial Data: Reflections on the Joint EUCROF/eCF Task Force whitepapers

June 2024

Matthew Addis

The European CRO Federation (EUCROF) in partnership with the eClinical Forum (eCF) have recently authored a series of inter-related whitepapers that cover the challenging topic of how to archive and preserve Clinical Trial Data (CTD). These whitepapers build on a previously published report on Trial Master File Archiving and the Decommissioning of Computerised Systems Used in Clinical Trials (2021).

Before looking at the whitepapers, a note about CTD. CTD refers to all information (however recorded and wherever held) pertaining to a given clinical trial that is needed to "permit and contribute to the evaluation of the conduct of a trial and the reliability of the results produced". This includes data, metadata, audit trails, documentation, and other forms of information – including the documents that make up the eTMF (electronic Trial Master File) from the perspectives of clinical investigators (and their delegated parties) and the trial sponsor (and the sponsor's delegated parties).

ICH E6 (R3) reveals a shift in emphasis from 'essential documents' (terminology in R2) to 'essential records' in R3's Appendix C. This includes "data and relevant metadata (including documentation of data corrections) in the data acquisition tools", along with records such as protocols, validation of systems, and qualification of suppliers.

Similarly, the 2023 EMA Guideline on computerised systems and electronic data in clinical trials covers data, metadata and audit trails, including 'source data' such as "hospital records, clinical and office charts, laboratory notes. Other examples are emails, spreadsheets, audio and/or video files, images, and tables in databases". CTD includes all of the above. The retention stage of the data lifecycle is no exception.

This blog post reviews the three new EUCROF/eCF whitepapers in the context of digital preservation good practice.

Outside the regulated life sciences community, organisations such as national libraries and archives now have 30+ years of experience of ensuring their digital data and records remain accessible and usable. Do the recommendations of the EUCROF/eCF whitepapers stack up? Will they likely work in practice given the experience of others who have wrestled with similar challenges for several decades? Can internationally developed digital preservation good practice provide additional support for those following the EUCROF/eCF recommendations?

A spoiler alert: the answer is a resounding 'yes'.

The three more recent EUCROF/eCF whitepapers are:

- 'The Decommissioning of Computerised Systems Used in Clinical Trials' (2023)
- 'Nature of a Distributed Trial Master File - Practical Aspects' (2023)
- 'Data Formats Used in Clinical Trials' (2023)

All these topics come into play when thinking about retention of CTD (including eTMF) for 25 years or more. No software system lasts forever, which means decommissioning and migrations will be a fact of life during the period of retention. Today's world of distributed and decentralised trials only serves to multiply the number of such systems and in turn multiplies the risks when trying to successfully retain data. The use of appropriate data formats is key to both migration success and long-term retention, which makes understanding and selection of data formats doubly important. And last, but not least, a good retention and archiving strategy that includes the use of digital preservation is essential for ensuring CTD remains accessible and usable over very long timescales. This not only to address regulatory compliance such as CTD retention in a way that meets the ALCOA+ principles of Data Integrity, but also to enable potential secondary use and value to be gained from reuse of CTD in the future, for example by ensuring CTD remains Findable, Accessible, Interoperable and Reusable (FAIR).

But how well do these reports align with the recommendations of the digital preservation community? Will the recommendations work in the real world? Can digital preservation good practice be used to take a proportionate and risk-based approach that can be justified and stood behind, for example when questioned by inspectors or the wider business? The short answer is that the reports provide a cohesive and solid foundation for successful retention and archiving of CTD.

The EUCROF/eCF reports do not make much use common digital preservation terminology, or explicitly reference digital preservation good practice. However, many of the suggestions are exactly what people from the digital preservation community would recommend and expect. The rest of this post explores the reports in more detail, looks at how digital preservation supports the recommendations made, and suggests some further good practice on long term digital preservation that can help.

Challenges of decommissioning, migration and archiving.

The EUCROF/eCF reports provide a good overview of the challenges organisations face when conducting a clinical trial, including how to minimise these challenges through good practice and preparation in earlier stages of the lifecycle, for example when systems and formats are selected. The digital world is not like traditional archiving paper in boxes. Today, we have multiple digital systems and vendors used in clinical trials. CTD is increasingly distributed and decentralised. Leaving clinical data in live systems for long periods of time, even if 'locked', is not a viable archiving strategy. Live systems will inevitably come to their end of life, or a customer may choose to switch to another vendor. And even if live systems did last forever, they don't support digital preservation (more on that below). Over time, the direction of travel of live systems will naturally evolve towards new approaches for new clinical trials and away from supporting old trials and old data. The 'IT world' does not have a good track record of long-term backwards compatibility and support for 'legacy' formats and systems – life sciences is no exception. Yet migrating data from 'live systems' into an archiving environment is often seen as problematic, time-consuming and risky. That said, and as the EUCROF/eCF report on decommissioning clinical systems describes, this is nowhere near as challenging as trying to archive CTD in live systems for 25 years, especially in a way that keeps all aspects of Data Integrity intact. As also described, migrating archived data back into a system that is 'recommissioned' so data can be made accessible again, such as for an inspection, is not an easy task either. Those responsible for retention and archiving of CTD are left between a rock and a hard place. The EUCROF/eCF reports do well in helping readers understand and navigate this tricky terrain.

Migrating distributed CTD that are held by multiple systems (EDC, eCOA, ePRO, CTMS, TMF and more) and from across multiple organisations (Sites, CROs, Sponsors, Third Party Vendors) into a consolidated set of dedicated archives is a sensible long-term strategy. Note that I've said *archives* plural – clinical data will naturally remain distributed even after archiving, for example split between Sites, Sponsors and other parties. But even in that distributed context, substantial consolidation is possible, although it is unlikely to happen overnight, for example at the moment a trial is closed. An inventory needs to be maintained of CTD, where it is located, who is responsible for it, how risks are monitored and managed, and how, why, and when it will be migrated. This is a point well made by the report on distributed Trial Master Files. Knowing what you have, where it is, and whether it is at risk is a cornerstone of digital preservation. The Digital Preservation Coalition's Rapid Assessment Model (DPC RAM) provides a way to assess the digital preservation maturity level of this approach. The approach of archiving data in distributed live systems, even with detailed inventories and risk management, would still only be somewhere around the Basic level (Levels in DPC RAM are on a scale of 0-4 with 4 being Optimised and 2 being Basic) – and that's only for the short term.

Migration of data into a dedicated archiving environment greatly reduces the complexity and risks of archiving CTD compared to using a set of live systems. This includes long-term safe storage of clinical data, but critically also includes maintaining that data in a readable form and with the means to view and interact with that data to support both inspections and reuse. Although expressed in different terms, the principles for ALCOA+ Data Integrity principles and FAIR data are but two sides of the same coin and both can be achieved through digital preservation. On the DPC RAM digital maturity scale, this would be a Managed or Optimised approach (levels 3, 4). Over 25 years, this makes a very material difference to the likely success of CTD remaining conformant to the ALCOA+ and FAIR principles. Digital preservation good practice can be aligned directly with the objectives of long-term GxP Data Integrity and provides a solid foundation for long-term risk-based CTD retention. There's more information on how digital preservation good practice supports the ALCOA+ principles in an eBook we produced in 2023 and a webinar we ran in 2024 on risk-based Data Integrity. Both are freely available from the Arkivum website.

Data, software and systems.

The EUCROF/eCF report on Data Formats Used In Clinical Trials emphasises the importance of using open specifications and open standards for data formats. This has long been recognised as a key part of digital preservation, especially for protecting against technical obsolescence and supporting data portability. The file format sustainability criteria from the Library of Congress and the format risk assessment done by NARA are good examples. For example, the criteria for file formats to be considered low risk are met very well by CDISC data standards such as ODM, CDASH, SDTM, ADaM and the eTMF Reference Model (including the Exchange Mechanism Standard), which has recently been added to the CDISC fold. Standardised data formats are often supported by multiple software systems, which helps reduce lock-in and migration issues, and in the case of CDISC standards they are also required by regulators such as the FDA so are in widespread use. The landscape of clinical data standards is widening, such as new formats for defining study protocols, and again this aligns with regulator expectations such as ICH M11. There is active work in this area by Transcelerate and CDISC, for example USDM and DDF, and wider work such as CTTI recommendations and HL7 FHIR. This will all further bolster the aspects of a clinical trial that can be defined using open standards.

This is all good news when archiving CTD. Standardised and open data formats tend to have longer lifetimes, they tend to have good software support without lock-in to just one vendor, they are more likely to have migration pathways when data needs to be moved to a newer format, and it is generally a lot easier to do Data Integrity checks, for example format conformance. The move in the clinical data domain towards machine readable data formats is beneficial because it helps maintain the dynamic nature of data as required by the regulators. It also moves us away from the paper-based mindset of 'printing to PDF' and into a modern world of digital data and records keeping. The more aspects of CTD that can be described using well-defined open formats and specifications, the better!

In the digital preservation world, using a canonical set of open file formats for long-term preservation along with a risk-based approach to technology watch and format migration has been used successfully for a wide range of data types (documents, data, images, audio, video, email and more). This approach can be seen in DPC RAM, for example as part of Content Preservation at the Managed and Optimised levels. Moreover, it's been proven to be effective in the real world and hence should be the preservation method of choice for CTD where possible.

That's not to say that other strategies such as capturing and archiving original software applications along with original data formats, for example using Virtual Machines, should be discounted. This may be the only viable approach in the short term, especially for proprietary data formats where there is no migration path, for example when trying to maintain the dynamic nature of

instrument data in lab environments. But as the EUCROF reports point out, it is wrong to believe that this is the only approach for dynamic data because of a mistaken expectation that the regulators will expect to see the original data in the original system even when that data is archived. This is not the case. The use of data formats that can support dynamic access and easy inspection, even if this is done using a separate 'viewer' or system, should be perfectly acceptable for archived data – provided of course that exports, migrations and the applications used are validated and suppliers are qualified.

Finally, after all this talk about data, it is worth remembering that CTD included records, documentation and audit trails. These need to be retained and remain accessible and usable. For example, retaining audit trails and records includes those for the live systems where data was originally collected and processed, for migrations of data between systems, including into an archive system, and for everything that happens to the data afterwards during retention such as data checks, data access and the application of preservation plans. The lack of audit trails and good record keeping is a recipe for inspection findings!

Digital Preservation good practice.

I've mentioned digital preservation good practice several times and I've used DPC RAM as an exemplar. Digital preservation is "the series of managed activities necessary to ensure continued access to digital materials for as long as necessary" (DPC). This aligns well with the objectives of CTD retention where principles such as ALCOA+ data integrity needs to be maintained for 25 years or more and data needs to remain FAIR so it can be reused. Digital preservation needs people with appropriate skills and experience, it needs resources and infrastructure, it needs systems and tools, it needs appropriate policies and procedures, and it needs sustainable funding. All these aspects are covered by models such as DPC RAM and in resources available from the digital preservation community.

For example, DPC resources of direct relevance to the EUCROF/eCF reports and the approaches they recommend are: the DPC RAM self-assessment framework; specific technical guidelines on things such as preserving different types of data (websites, emails, databases, documents, images, audio and video – all of which are increasingly featuring in clinical trials); and how to create digital asset registers.

However, the DPC provides a much broader set of guidelines, for example: setting preservation policy; creating business cases for archiving and preservation including risks, costs and benefits; procuring preservation systems; and tools for assessing staff skills and gaps. This is where perhaps the EUCROF/eCF reports are limited in their scope. Digital preservation is very much an active and ongoing process which requires people, resources and sustainable investment. The DPC guides can be used by organisations to help with all these dimensions. For example, the DPC guides can help organisations with making the business case for investing into archiving that delivers both compliance (e.g. 25 year ALCOA+) and value (e.g. FAIR reuse). It cannot be overstated that digital preservation is not just a technical issue. The risks are not only about formats, systems, suppliers, contracts, governance and oversight. Successful digital preservation needs to take a holistic approach that considers whether an organisation has the right people, the right skills, the proper governance, and the right level of investment. This is where digital preservation good practice can add to the good work already done in the EUCROF/eCF reports.

Conclusions

The EUCROF/eCF whitepapers provide excellent and detailed recommendations for what to do at the end of a clinical trial, especially when decommissioning systems, planning for archiving, and selecting appropriate data formats. The recommendations fit well with established digital preservation good practice and should set organisations in good stead when setting sail for clinical trial data archiving and retention. Following good practice for digital preservation ensures that journey not only starts successfully but remains on course and navigates the rough and uncharted seas that will inevitably be encountered upon the way. The voyage is long one when ensuring that clinical trial data remains accessible and usable for 25 years or more! Consolidating data into dedicated archiving and digital preservation environments that are seaworthy and sailed by people with the right digital preservation skills and experience makes that journey safer, smoother and ensures valuable clinical trial cargo will remain intact and ready for use at any time.

Acknowledgements

Arkivum would like to thank Tony Hewer (Founder & Director, Cepheus Consultancy Limited) for his invaluable contributions to this blog post. These include: taking the time to discuss the EUCROF/eCF reports with Arkivum; suggestions for the content of the blog; and review and improvements to the draft versions.

About Us

Arkivum is the only GxP validated digital preservation solution to guarantee long-term data integrity (ALCOA+). Our truly vendor independent solution supports clinical trial sponsors, CROs and sites in meeting clinical and all GcP archiving requirements. This is achieved by providing a centralised and easily manageable repository for commercially valuable assets (e.g. eTMF/EDC/ePRO).

Contact Us

hello@arkivum.com
Quadrant House, 20 Broad Street Mall, Reading, RG1 7QE