

The Importance of CDISC Standards When Retaining, Archiving and Preserving Clinical Trial Records and Data

Matthew Addis, Arkivum

14 May 2025





2025

CDISC + TMF
EUROPE INTERCHANGE

GENEVA

CONFERENCE & EXPO: 14-15 MAY | TRAININGS: 12, 13, 16 MAY

The Importance of CDISC Standards When Retaining, Archiving and Preserving Clinical Trial Records and Data

Matthew Addis, CTO, Arkivum



Meet the Speaker

Matthew Addis

Title: Chief Technology Officer

Organization: Arkivum

Matthew is CTO and co-founder of Arkivum, where he is responsible for technical strategy. Matthew is a subject matter expert on data archiving and digital preservation, including how this can be applied in regulated environments for GxP data. Matthew's expertise includes digital preservation strategies and techniques, long-term data retention using risk-based approaches to Data Integrity, implementation in SaaS solutions and architectures, and last, but definitely not least, achieving environmental sustainability for archives and repositories.



Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*
- *The author(s) have no real or apparent conflicts of interest to report.*



Agenda

1. Regulatory requirements for retention and archiving
2. The importance of Long Term Digital Preservation (LTDP)
3. Risk assessment and format sustainability
4. LTDP benefits of CDISC standards
5. Data management planning with LTDP in mind



Regulatory Requirements for Retention and Archiving



Records and Data

Retention Periods

**GxP Regulations and
Guidelines**

ALCOA+ Data Integrity

Risk Based Approach



Example: ICH E6 (R3) Essential Records

- Trial documentation
- Data and metadata from data acquisition tools
- Source data
- Datasets and statistical analysis
- Submissions and exchanges with regulators
- Supplier qualification
- Systems validation
- Shipping and storage of investigational products
- Manuals
- Maintenance and calibration records
- Staff training
- SOPs

Systems

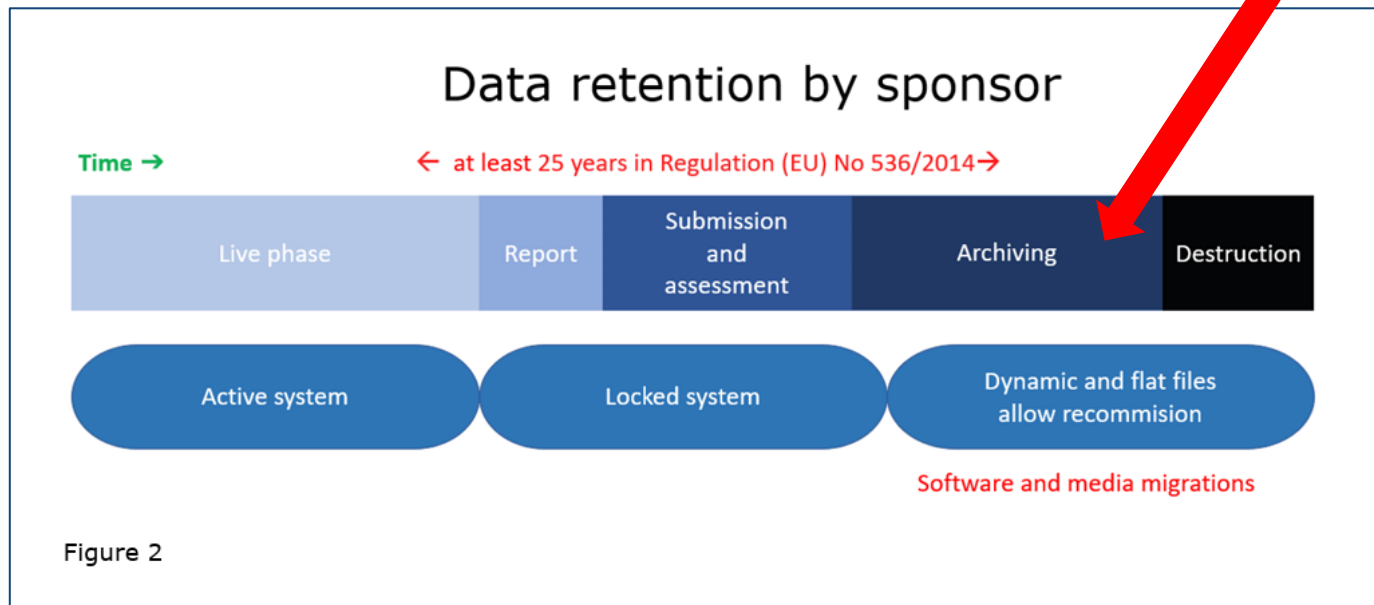
EDC, ePRO, eCOA, CTMS,
IRT, eTMF, RIMS, QMS, LMS,
LIMS, ELN, CDS, EMS, DMS,
IMS...

Data

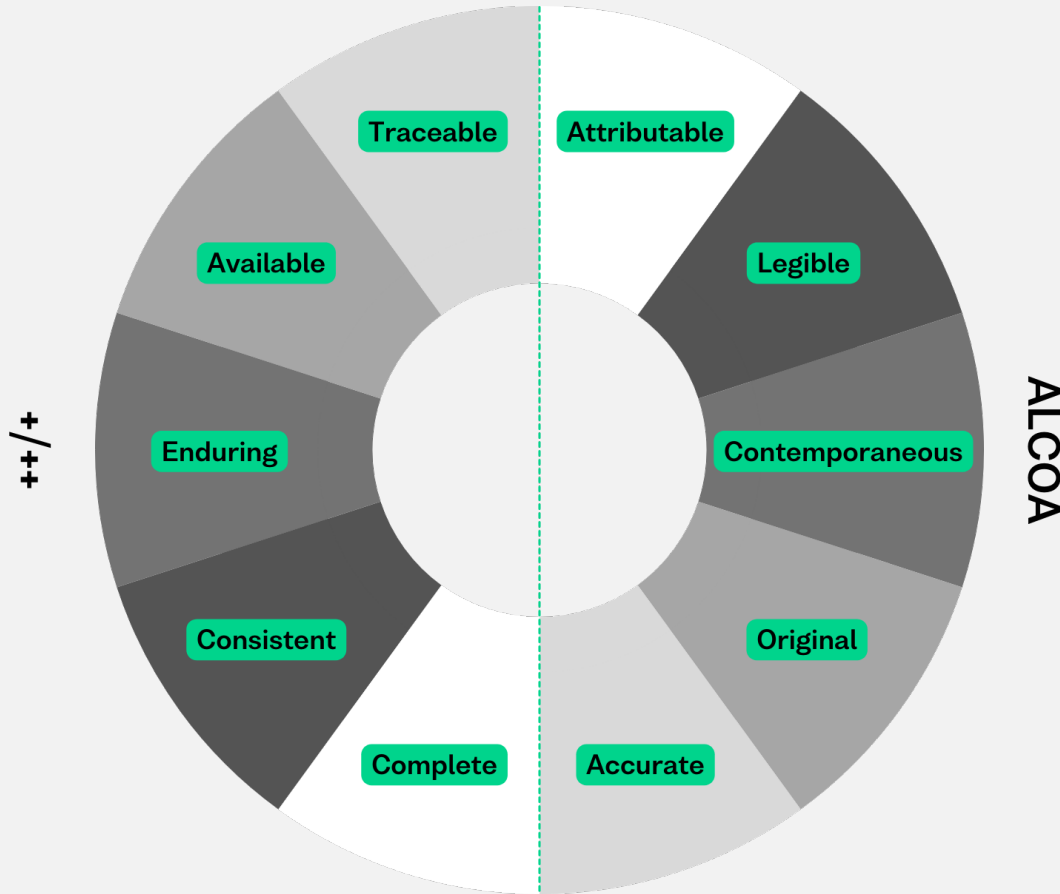
Documents, datasets,
metadata, audit trails, emails,
images, spreadsheets, raw data
...

Example: EU CTR Retention Period

25+ Years!



Data Integrity: ALCOA+ Principles



- ALCOA+ is pervasive in the GxP guidelines
 - MHRA
 - FDA
 - EMA
 - WHO
 - OECD
 - ICH
 - PIC/S
- **Data = files, metadata, audit trails, documentation**



ALCOA++ - A Way to Think About Risks

Legible & Traceable

- Will documents and data be readable after 25 years?
- Can the audit trail be used to recreate events from 25 years ago?
- Is there documentation so someone can still understand the data?
- Will you be locked into proprietary/ legacy formats and software?

Enduring

- Can data become corrupted or lost when it is being stored?
- Will data become spread across and locked into many EoL systems?
- Is the data immutable and can attempts to change it be detected?
- Are there controls over who can remove or delete data?

Attributable, Accurate & Contemporaneous

- Can timestamps be altered?
- Is the audit trail permanent?
- Will signatures always validate?
- Can any deletions or changes go unnoticed or unapproved?

Available (25 years)

- Can data be discovered easily (metadata)?
- Can data be retrieved quickly (ready access)?
- Is everything documented?
- Will data become spread across lots of legacy systems and be impossible to find?
- Does BCDR cover cyberattacks, vendors going bust, disasters in the cloud?
- Are there sufficient budget, staff and skills to sustain the archive?

Complete & Correct

- Can data integrity issues go undetected during transfers or archiving?
- Can you prove the entire TMF (structure, files, metadata, audit logs) is complete?
- Can you prove migrations (formats, systems, people) were successful?

Threats to Data Integrity

- Systems fail: data is corrupted or lost
- Backup failures
- Accidental or deliberate alteration of records
- End of Life systems: data cannot be migrated
- Proprietary formats, vendor lock in
- Suppliers go bust: no BCDR plan
- Cyber attacks and ransomware
- Formats not supported: data can't be accessed
- Audit trails expire or are deleted
- Data migrations are not validated
- Data is distributed and can't be found
- Systems are not validated for archiving
- EoL systems not secure: no recommissioning
- No evidence that data hasn't changed
- No one understands old data formats
- No one understands obsolete applications

Consequences of Data Integrity Failures

- Health and safety of study participants
- Failed inspections & CAPAs
- Rejection or delay to MAA
- Removal of drug from the market
- Financial penalties
- Quality issues with products
- Delayed sales or MNA
- Data not findable or usable
- Cost of doing repeat work
- Cost of doing additional work
- Loss of funding and revenues
- IPR can't be exploited
- Reputational damage
- Ethical issues

Managing Risks: Long Term Data Integrity

Causes of Data Integrity Failures

- Data is corrupted or lost
- Backup failures
- Accidental or deliberate alteration of records
- End of Life systems, data cannot be migration
- Proprietary formats, vendor lock in
- Suppliers go bust, no BCDR plan
- Cyber attacks and ransomware
- Formats not supported, data can't be accessed
- Audit trails expire or are deleted
- Data migrations are not validated
- Data is distributed and can't be found
- Systems are not validated for archiving
- EoL systems not secure, no recommissioning
- No evidence that data hasn't changed
- No one understands old data formats
- No understands obsolete applications

Consequences of Data Integrity Failures

- Health and safety of study participants and patients
- Failed inspections & CAPAs
- Rejection or delay to MAA
- Removal of drug from the market
- Financial Penalties
- Quality issues with products
- Delayed sales or MNA
- Data not findable or usable
- Cost of doing repeat work
- Cost of doing additional work
- Reputation damage
- Ethical issues
- Loss of funding and revenues
- IPR can't be exploited



Probability	Harm Severity			
	Minor	Marginal	Critical	Catastrophic
Certain	High	High	Very High	Very High
Likely	Medium	High	High	Very High
Possible	Low	Medium	High	Very High
Unlikely	Low	Medium	Medium	High
Rare	Low	Low	Medium	Medium
Eliminated	Eliminated			



The Importance of Long Term Digital Preservation

Digital Preservation

“the series of managed activities necessary to ensure continued access to digital materials for as long as necessary”

Digital Preservation Coalition

Systems for ingest, storage, preservation, management and access.

Technology

Organisation

Resources

Preservation strategy, processes, procedures

Funding, staff, skills and expertise

Data Archiving and Digital Preservation are Not the Same

Archiving

- Place where data is held for safe keeping
- Data is typically read-only
- Backed up
- Restricted access
- Kept 'as-is' with no changes or updates
- Sometimes held within a live system, e.g. after data is 'locked'
- Often treated as the digital equivalent of 'boxes of paper in a storage facility'
- Not a viable solution for data that needs to be readable and usable for 25 years!

vs.

Preservation

- Long-term safe storage with fixity checks
- Data Integrity checks and management (files, metadata, audit trails)
- Technology watch and management of technical obsolescence
- Preservation actions so content maintains its meaning and remains usable
- Metadata ensures content is documented, discoverable and usable
- Evidence of ongoing data integrity and application of digital preservation
- All the processes, techniques and systems for indefinite retention and use

Digital Preservation: Good Practice and Maturity Models

- Good practice: internationally recognised and tested
- Practical and specific things to do in the real world
- Created by organisations with decades of experience
- Start simple and work up
- Self-assessment tools and planning



Find out more:
<https://ndsa.org/publications/levels-of-digital-preservation/>



Find out more:
<https://www.dpconline.org/digipres/dpc-ram>



Find out more:
<https://www.coretrustseal.org/>

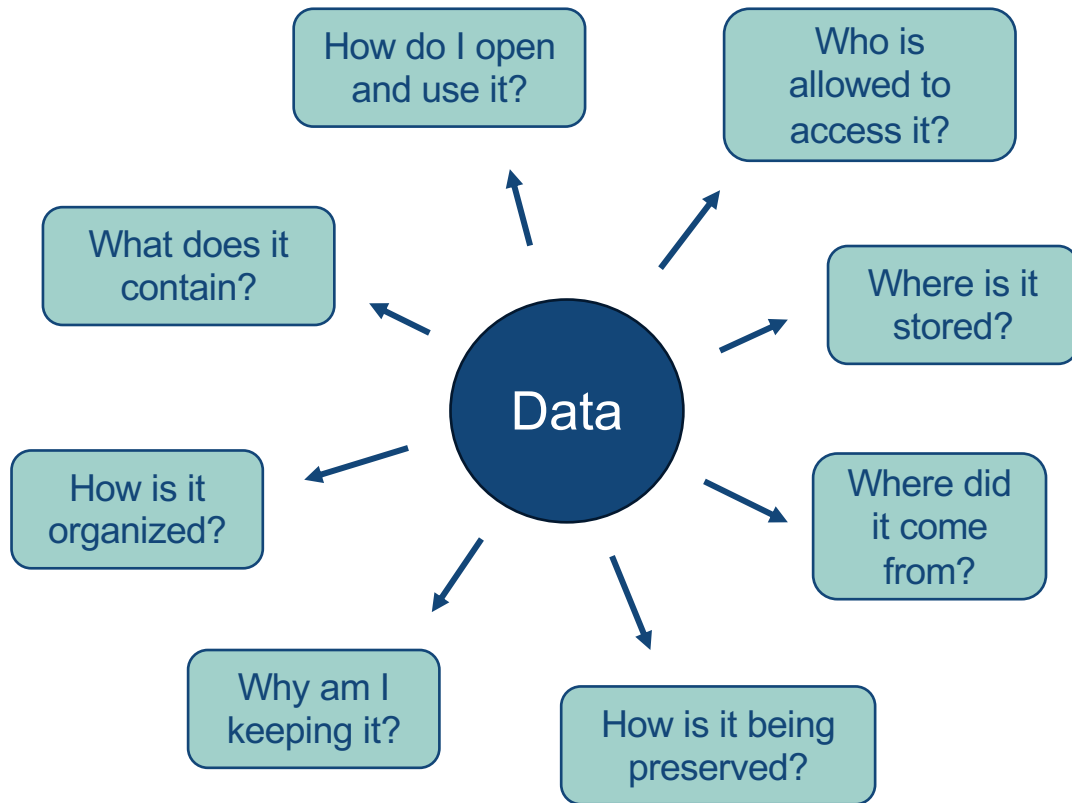
Digital Preservation: Ensuring Digital Content Remains Usable

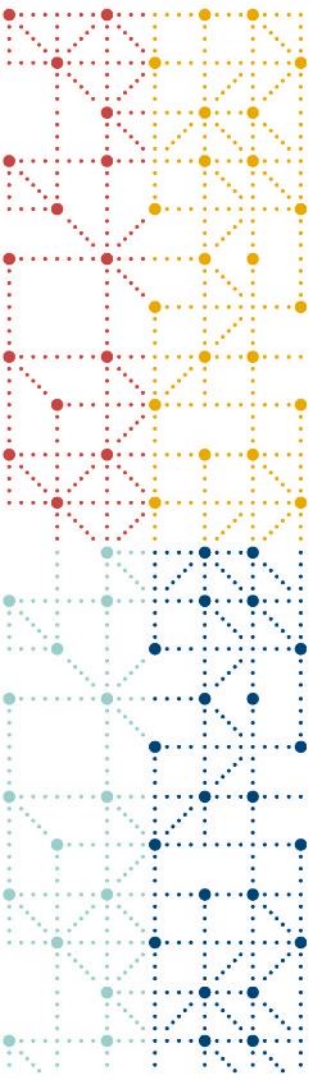
- Create inventories
- Document file formats
- Capture metadata
- Assess obsolescence risks
- Check format compliance
- Migrate formats
- Emulate software
- Verify access and usability

Functional Area	Level			
	Level 1 (Know your content)	Level 2 (Protect your content)	Level 3 (Monitor your content)	Level 4 (Sustain your content)
Storage	Have two complete copies in separate locations Document all storage media where content is stored Put content into stable storage	Have three complete copies with at least one copy in a separate geographic location Document storage and storage media indicating the resources and dependencies they require to function	Have at least one copy in a geographic location with a different disaster threat than the other copies Have at least one copy on a different storage media type Track the obsolescence of storage and media	Have at least three copies in geographic locations, each with a different disaster threat Maximize storage diversification to avoid single points of failure Have a plan and execute actions to address obsolescence of storage hardware, software, and media
Integrity	Verify integrity information if it has been provided with the content Generate integrity information if not provided with the content Virus check all content; isolate content for quarantine as needed	Verify integrity information when moving or copying content Use write-blockers when working with original media Back up integrity information and store copy in a separate location from the content	Verify integrity information of content at fixed intervals Document integrity information verification processes and outcomes Perform audit of integrity information on demand	Verify integrity information in response to specific events or activities Replace or repair corrupted content as necessary
Control	Determine the human and software agents that should be authorized to read, write, move, and delete content	Document the human and software agents authorized to read, write, move, and delete content and apply these	Maintain logs and identify the human and software agents that performed actions on content	Perform periodic review of actions/access logs
Metadata	Create inventory of content, also documenting current storage locations Backup inventory and store at least one copy separately from content	Store enough metadata to know what the content is (this might include some combination of administrative, technical, descriptive, preservation, and structural)	Determine what metadata standards to apply Find and fill gaps in your metadata to meet those standards	Record preservation actions associated with content and when those actions occur Implement metadata standards chosen
Content	Document file formats and other essential content characteristics including how and when these were identified	Verify file formats and other essential content characteristics Build relationships with content creators to encourage sustainable file choices	Monitor for obsolescence, and changes in technologies on which content is dependent	Perform migrations, normalizations, emulation, and similar activities that ensure content can be accessed

The Importance of Metadata

- Descriptive
- Structural
- Technical
- Administrative
 - Rights
 - Retentions
 - Preservation
 - Provenance
- Digital Asset Registers





Risk Assessment and Format Sustainability

Library of Congress: Sustainability Criteria

- Disclosure
- Adoption
- Transparency
- Self-Documentation
- External Dependencies
- Impact of Patents
- Technical Protection Mechanisms

Format Descriptions

Still Image

- [SVG_1_1](#)
- [TIFF_6](#)
- [All still image format descriptions](#)

Sound

- [WAVE](#)
- [MP3_FF](#)
- [All sound format descriptions](#)

Moving Image

- [MPEG-4_FF_2](#)
- [AVI](#)
- [All moving image format descriptions](#)

Textual

- [PDF/A_family](#)
- [DOCX/OOXML_2012](#)
- [All text format descriptions](#)

Web Archive

- [ARC_IA](#)
- [WARC](#)
- [All Web archive format descriptions](#)

Datasets

- [DBF](#)
- [HDF5](#)
- [All dataset format descriptions](#)

Geospatial

- [ESRI_shape](#)
- [GeoPackage_1_0](#)
- [All geospatial format descriptions](#)

Email and PIM

- [MBOX](#)
- [MSG](#)
- [All email and PIM format descriptions](#)

Design and 3D

- [STEP](#)
- [X3D](#)
- [All design and 3D format descriptions](#)

Accessibility

- [SRT](#)
- [WebVTT](#)
- [All accessibility-related format descriptions](#)

Aggregate

- [RAR](#)
- [ZIP](#)
- [All aggregate format descriptions](#)

Generic

- [ASF](#)
- [RIFF](#)
- [All generic format descriptions](#)

NARA: File Format Risk Assessment

- Disclosure
- Adoption
- Transparency
- Self-Documentation
- External Hardware Dependencies
- External Software Dependencies
- Impact of Patents
- Technical Protection Mechanisms


SOP for NARA Digital Preservation Framework

Table of Contents

Table of Contents	1
SOP Revision and Review History	2
SOP Purpose Statement and Scope	2
Authority for Creating the SOP	3
When does this SOP take effect?	3
Terms Used	3
NARA Acronyms and Terms	3
Non-NARA Acronyms and Terms	3
Infrastructure/Equipment	3
Computer Hardware, Software	3
Other Equipment and Supplies	4
Methodology	4
File Format Matrix	4
Risk	4
Prioritization	8
Preservation Action Plans: File Formats	9
File Format Identifiers Section	9
Links Section	9
Proposed Preservation Actions Section	10

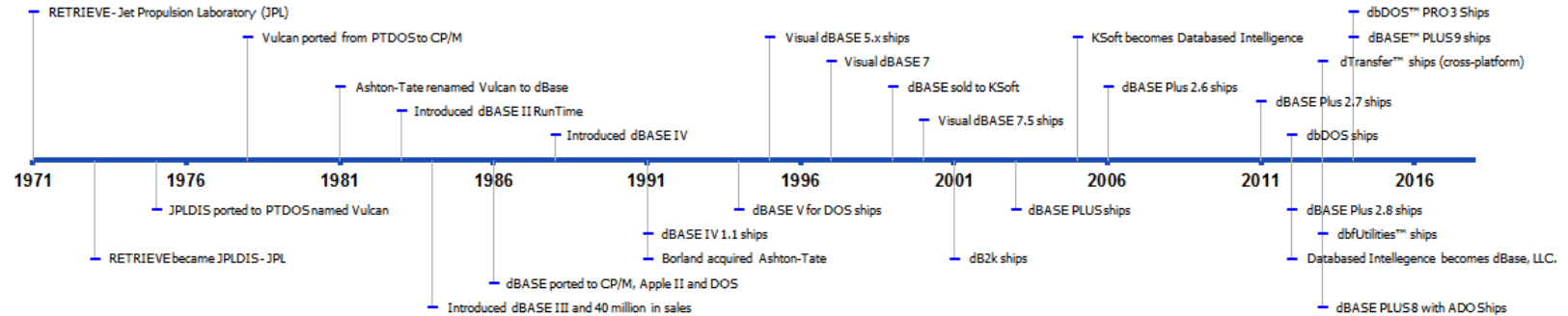
NARA Transfer Guidance: Preferred	NARA Transfer Guidance: Acceptable	Numeric Risk Rating	Risk Level	NARA Format ID	Format Name	File Extension(s)	Category/Plan(s)
		5.00	Moderate Risk	NF00269	Microsoft Excel for Macintosh 4.0	xls	Spreadsheets
	X	17.00	Moderate Risk	NF00270	Microsoft Excel for Macintosh 98	xls	Spreadsheets
	X	17.00	Moderate Risk	NF00271	Microsoft Excel for Macintosh v.X	xls	Spreadsheets
		14.00	Moderate Risk	NF00665	Microsoft Excel Macro-enabled	xlsm	Spreadsheets
	X	22.00	Moderate Risk	NF00272	Microsoft Excel Office Open XML	xlsx	Spreadsheets
		24.00	Low Risk	NF00770	Microsoft Excel Template	xlt	Spreadsheets
		3.00	Moderate Risk	NF00656	Microsoft Excel unspecified version	xls	Spreadsheets
		7.00	Moderate Risk	NF00273	Microsoft Excel Workspace	xlw	Spreadsheets
		-26.00	High Risk	NF00278	Microsoft Multiplan 4.0	mod	Spreadsheets
		27.00	Low Risk	NF00346	OpenDocument Formula	odf otf	Spreadsheets
X		30.00	Low Risk	NF00349	OpenDocument Spreadsheet 1.0	ods fods ots	Spreadsheets
X		30.00	Low Risk	NF00513	OpenDocument Spreadsheet 1.1	ods fods ots	Spreadsheets
X		30.00	Low Risk	NF00514	OpenDocument Spreadsheet 1.2	ods fods ots	Spreadsheets
X		30.00	Low Risk	NF00800	OpenDocument Spreadsheet 1.3	ods fods ots	Spreadsheets
		13.00	Moderate Risk	NF00527	Quattro Pro Spreadsheet 7-8	wb3	Spreadsheets
		9.00	Moderate Risk	NF00528	Quattro Pro Spreadsheet 9-12, X3, X4	qpw	Spreadsheets
		-15.00	Moderate Risk	NF00390	Quattro Pro Spreadsheet for DOS 1-4	wq1	Spreadsheets
		-15.00	Moderate Risk	NF00526	Quattro Pro Spreadsheet for DOS 5.x	wq2	Spreadsheets
		-5.00	Moderate Risk	NF00391	Quattro Pro Spreadsheet for Windows 1-5	wb1	Spreadsheets
		-5.00	Moderate Risk	NF00529	Quattro Pro Spreadsheet for Windows 6	wb2	Spreadsheets
X		26.00	Low Risk	NF00143	Comma Separated Values	csv	Structured Data
	X	4.00	Moderate Risk	NF00183	Extended Binary Coded Decimal Interchange Code (EBCDIC)	ebcdic	Structured Data
		19.00	Moderate Risk	NF00189	eXtensible Metadata Platform	xmp	Structured Data
		-25.00	High Risk	NF00721	HLM Multivariate Data Matrix Format	mdm	Structured Data
X		31.00	Low Risk	NF00218	JavaScript Object Notation (JSON)	json txt	Structured Data
		-25.00	High Risk	NF00582	Mathematica Computable Document Format	cdf	Structured Data
		34.00	Low Risk	NF00605	Resource Description Framework (RDF) XML Triple	rdf	Structured Data
		37.00	Low Risk	NF00410	Standard Generalized Markup Language (SGML)	sgm sgml	Structured Data
		7.00	Moderate Risk	NF00415	Structured Data eXchange Format	sdx	Structured Data
		35.00	Low Risk	NF00782	Synchronized Multimedia Integration Language	smi smil	Structured Data
		15.00	Moderate Risk	NF00418	Tab Separated Values	tab tsv	Structured Data

OPF: Accepted File Formats

 <div>Preferred = 2 Accepted = 1 Accepted, but undesired = 0 Unaccepted = -1 No value = undefined or outside of scope</div>	National & Federal Archives																	
	Country	Australia	Belgium	Canada	Canada	Denmark	Estonia	Finland	France	Italy	The Netherlands	New Zealand	Norway	Sweden	Switzerland	United Kingdom	United Kingdom	USA
Institution	National Archives of Australia	State Archives of Belgium	Library and Archives Canada	BANQ	Rigsarkivet	Rahvusarhiiv	Digital Preservation Service	Archives Nationales	AGiD	Nationaal Archief	Archives New Zealand	Arkivverket	Riksarkivet	Das Bundesarchiv	The National Archives	Imperial War Museum	National Archives	
Formats (categorised by content information type)	Total Score	Format guidelines	n/a	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	N/A	Format guidelines	
Tagged Image File Format (TIFF) ver 4, 5, & 6	10	-1	0			2	0	2		2		1	2	-1	-1	0	2	
Structured Data, Databases																		
DBase (DBF)	-4	-1	0	1		-1	0	-1			1			-1	-1			
Delimited text with SQL data definition statements	8	-1	0	-1		-1	1	-1		2	1	1		-1	-1			
Extended Binary Coded Decimal Interchange Code (EBCDIC)	-1	2	0	1		-1	0	-1			1			-1	-1			0
Microsoft Access (MDB, ACCDB)	-5	-1	1	-1		-1	0	-1	-1	1	1			-1	-1			
Microsoft Excel 97 (XLS) Binary Document ver 8.0	3	-1	0	1	1	-1	1	1	0	-1	1			-1	-1			1
Office Open XML (XLSX)	15	-1	0	1	1	-1	0	1	0	1		1	1	-1	-1			1
OpenDocument Spreadsheet (ODS) ver 1.0-1.3	10	-1	0	1	1	-1	0	2	0	1		1		-1	-1			2
OpenDocument Database (ODB) ver 1.0-1.3	-4	-1	2	-1		-1	0	-1	-1	-1	2	1		-1	-1			
Plain text (TXT) with ASCII Text ver 7 bit or Unicode Text ver UTF-8 encoding	18	-1	0	2	2	-1	0	2	1			1	2	2	-1			2
System-Independent Archiving of Relational Databases (SIARD) ver 2.0, 2.1 & 2.2	30	-1	1	-1		2	2	2	-1		2	1	2	-1	2			2
Structured Query Language (SQL)	6	-1	2	-1		-1	0	-1	-1	2	2	1	1	-1	-1			
Tab-delimited text (TAB) with ASCII or Unicode encoding	14	2	0	-1		-1	0	2				1	2	-1	-1			
Markup Languages																		
Comma Separated Value (CSV)	51	2	2	2	2	-1	2	2	2	2	2	1	2	2	-1		2	2
Extensible Markup Language (XML)	43	2	2	1	1	-1	0	2	1	1		1	2	2	2		2	2
JavaScript Object Notation (JSON)	8	2	0	-1		-1	0	-1		1		1	2	-1	-1		2	2
NoSQL Databases																		
Extensible Markup Language (XML) + schema (XSD)	17					-1	2	2		2		1	2	2				
Resource Description Framework (RDF)	5					-1	2	-1				1		-1				
Textual Documents																		
eXtensible Markup Language (XML)	14	-1	0	-1		-1	0	2		1		1	2	-1	-1	1	2	
JSON	5		0			-1		-1		1		1	2	-1			2	

Regular Evaluation: Risks Change Over Time

dBASE Timeline



Dbasellc, CC BY-SA 3.0, via Wikimedia Commons



LTDP Benefits of CDISC Standards

CDISC Standards in the Clinical Research Process

■ Foundational Standard ■ Therapeutic Area
■ Data Exchange ■ Controlled Terminology

Non-clinical

Clinical

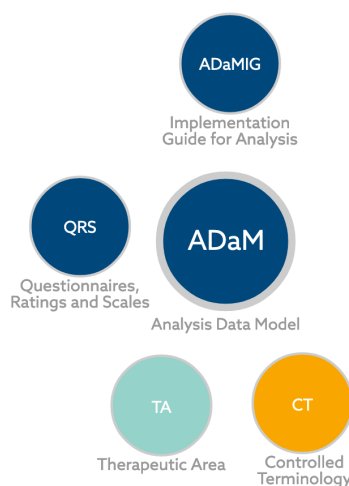
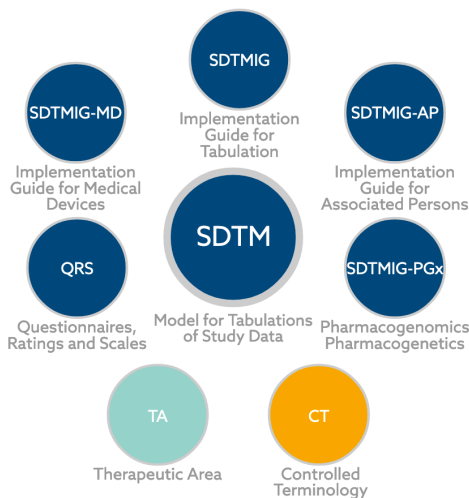
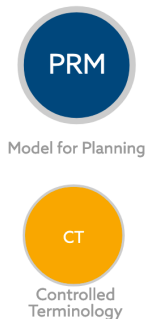
Organize

Plan

Collect

Organize

Analyze



Data Exchange



Therapeutic Areas

Controlled Terminology

Comparison with Sustainability Criteria

Sustainability Criteria	CDISC Standards
Disclosure	<ul style="list-style-type: none">✓ Open specifications✓ Good documentation
Adoption	<ul style="list-style-type: none">✓ Widely adopted✓ Support from multiple vendors and open-source projects
Transparency	<ul style="list-style-type: none">✓ Transparent: human and machine readable (XML, json etc.)
Self-documentation	<ul style="list-style-type: none">✓ Schemas, dictionaries and controlled vocabularies
External dependencies	<ul style="list-style-type: none">✓ Not tied to specific hardware or software
Impact of patents	<ul style="list-style-type: none">✓ Not encumbered by patents or other IP restrictions
Technical protection mechanisms	<ul style="list-style-type: none">✓ Content not locked or encrypted

Benefits of Open Standards and Specifications

- Longer lasting → fewer format migrations, less risk
- Less vendor lock-in
- More choice of suppliers and systems
- Simpler migrations
- Easier Data Integrity checks
- Retain more functionality, especially for dynamic data
- Readily available and open documentation



Allotrope Data Format (ADF)

APIs (Java & .NET class libraries)

Data Description
Semantic Graph Model

Data Cubes
Universal Data Container

Data Package
Virtual File System

HDF5
Platform Independent File Format

Descriptive metadata about

- Method, instrument, sample, process, result, etc.
- Provenance, audit trail
- Data Cube, Data Package

Analytical data represented by one- or multidimensional arrays of homogeneous data structures.

Data represented by arbitrary formats, incl. native instrument formats, images, pdf, video, etc.

Specifically designed to store and organize large amounts of scientific data.

IDMP

Identification of Medicinal Products
Data elements and structures
for the unique identification and exchange



PISTOIA ALLIANCE



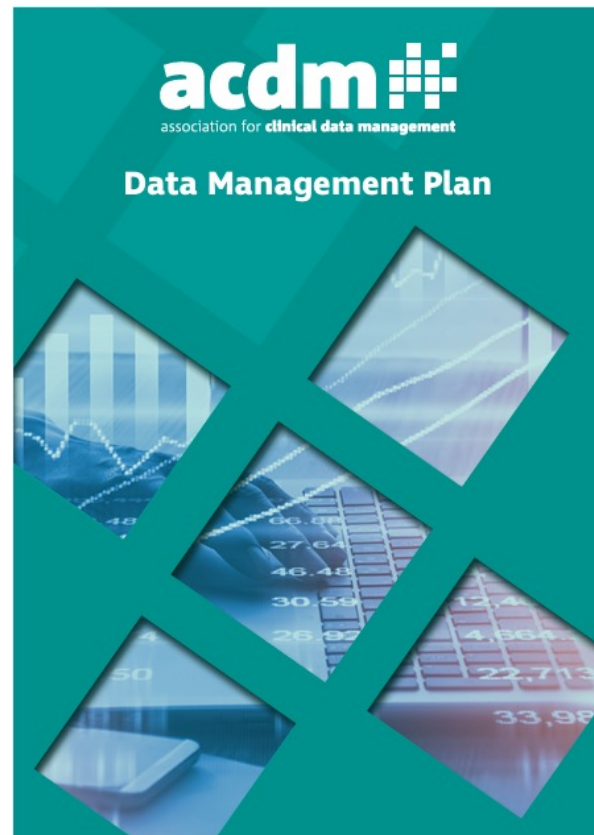
IDMP – O



Data Management Planning with LTDP in Mind

Data Management Plans

- Risk Assessment
- Selection and Appraisal
- Transfers and Migrations
- Archiving and LTDP
- Registries and Inventories
- Retention and Disposition
- Business Continuity and Disaster Recovery

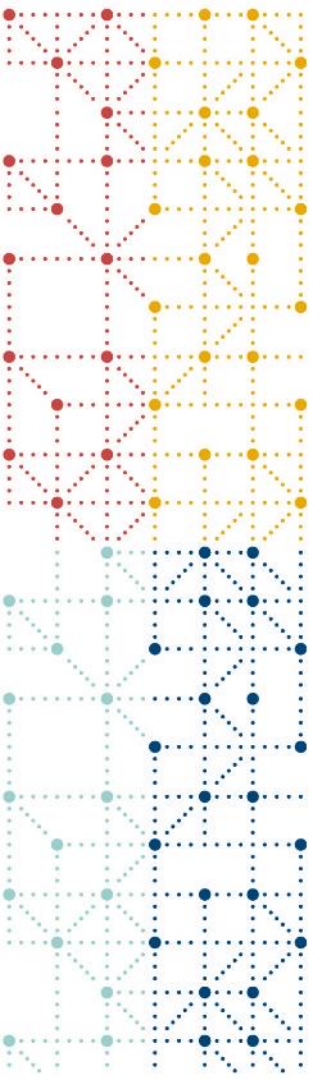


ACDM Template for DMPs
<https://acdmglobal.org/dmp/>



DMP: Data and File Formats

- Make records of the data / file formats being used
 - Who, what, when, why, where (e.g. as part of a data flow diagram)
 - Capture specifications, software applications, documentation, licenses etc.
- Test exporting and migrating data out of source systems
 - Define data integrity checks and validation
- Define an SOP for long-term risk-based data management
 - Acceptable formats for long-term retention
 - Process for assessing and managing risks
- Perform initial and ongoing format risk assessments
- Migrate / create versions using open specifications / standards where possible
 - CDISC, HL7 FHIR, DICOM etc.
- Align DMP with long-term Data Integrity based on ALCOA+
 - Files, data, records, metadata, audit trails, documentation



Summary



Collecting data in live systems



Data transfers between systems



New software releases, validation



Inspections, data checks, oversight



Hardware and Software migrations



Staff changes, recruitment, training



Obsolescence, data format migrations



Merger, Acquisition, Sale, Disposition



Migration of Archive Solution and Provider



25 years



Summary

- Think about the long-term risks to Data Integrity
- Take advantage of the good practice developed by the LTDP community
- Understand and document your data formats
- Take a risk-based approach to ensuring data remains readable and usable
- Make use of open standards and specifications such as CDISC
- Include long-term thinking in your Data Management Plan



Thank You!

