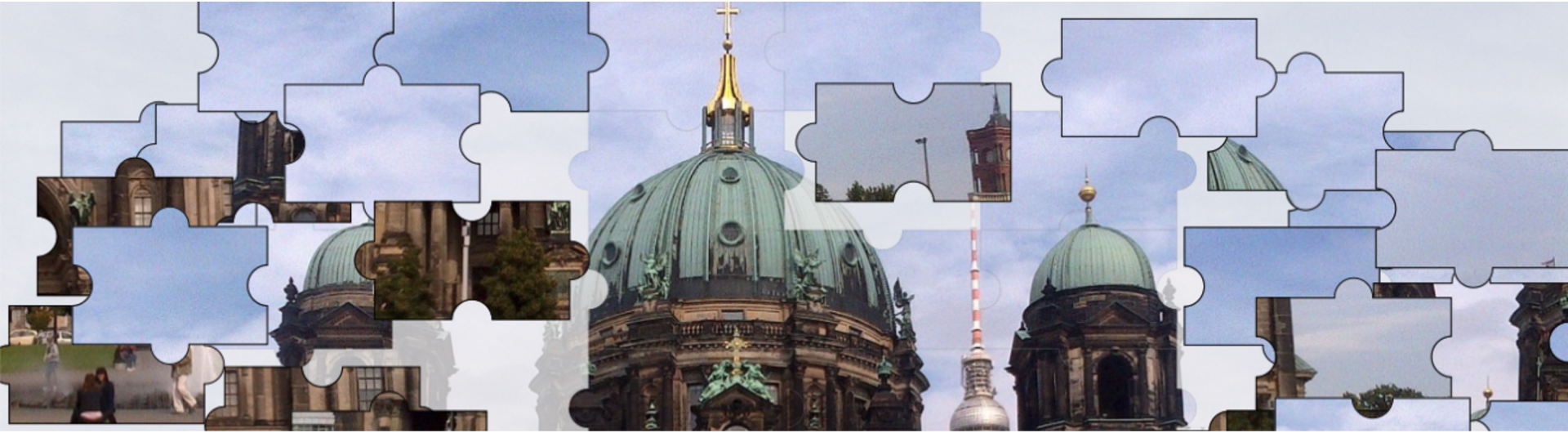


PDF Days Europe 2025

PDF 2030

Putting the pieces together

Kevin L. De Vorsey | CEO | ThinkBox.DIGITAL

Please Allow me to (Re) Introduce Myself

Kevin L. De Vorsey – CEO - ThinkBox.DIGITAL

I have been involved in museum curation, digital preservation, and archiving since 1992:

- National Museum of the American Indian – Smithsonian Institution
- Gallery Systems
- American Museum of Natural History
- The National Library of New Zealand
- U.S. National Archives and Records Administration

I’ve been a member of ISO TC171 SC2 since 2011 and am currently Convener for WG5.

ThinkBox.Digital

We are a collective of digital preservation and electronic records management experts.

Other members include:

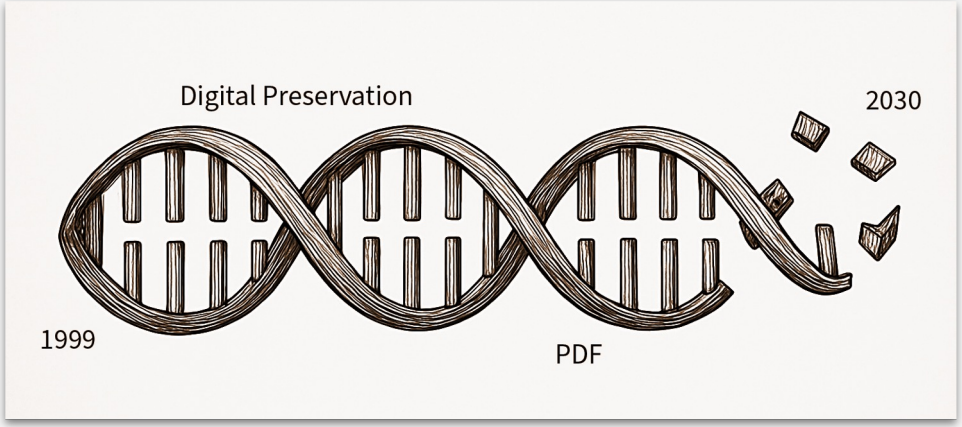
- Lisa Haralampus
- Greg Lepore

Our goal is to help vendors improve their products and organizations to improve their workflows to mitigate risk and to ensure that their information will be available for as long as it is needed.

Digital Preservation and PDF/A Timelines

Topics we will discuss:

1. How digital preservation has evolved in theory and practice.
2. The preservation model that was in effect when PDF/A was developed.
3. Exciting changes in the PDF landscape (PDF/R, new raster filters, Rich Media, C2PA).
4. How we can make PDF/A an important target format for digital preservation moving forward.



Key Dates in PDF

1993
PDF 1.0

2001
PDF 1.4

2005
PDF/A-1
ISO 19005-1

2006
PDF 1.7

2008
ISO 32000:1

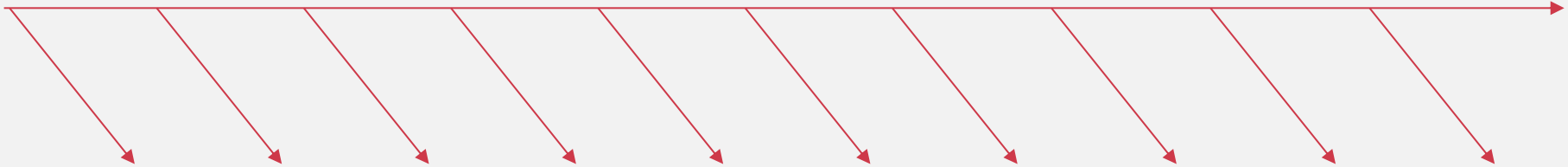
2011
PDF A-2
ISO 19005-2

2012
PDF/A-3
ISO 19005-3

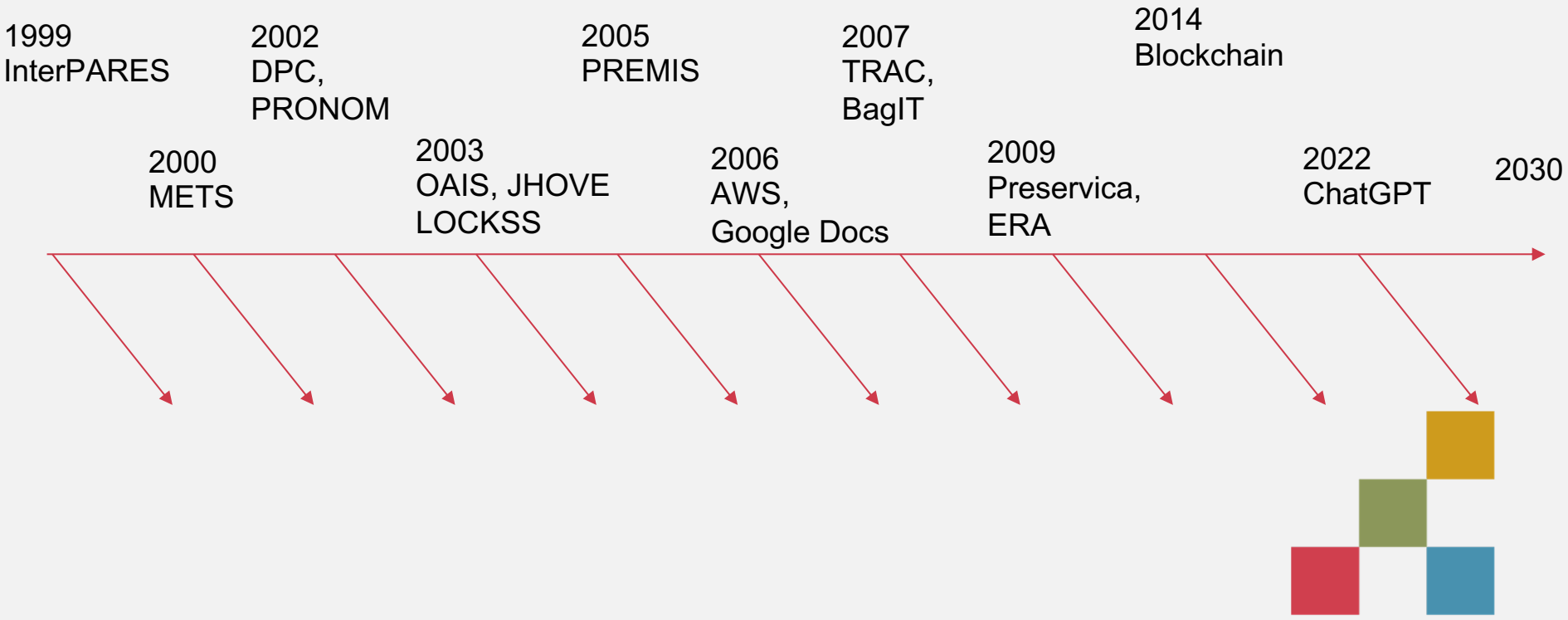
2017
PDF 2.0
ISO 32000:2

2020
PDF/A-4
ISO 19005-4

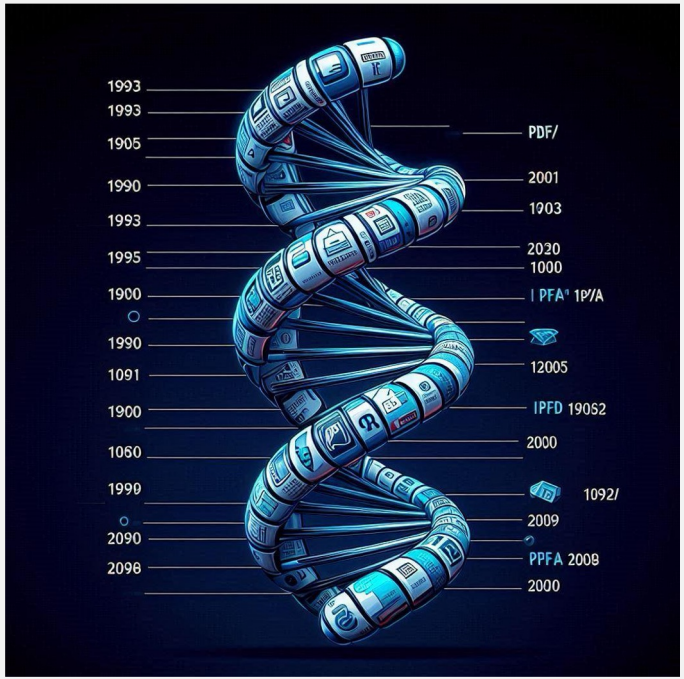
2030

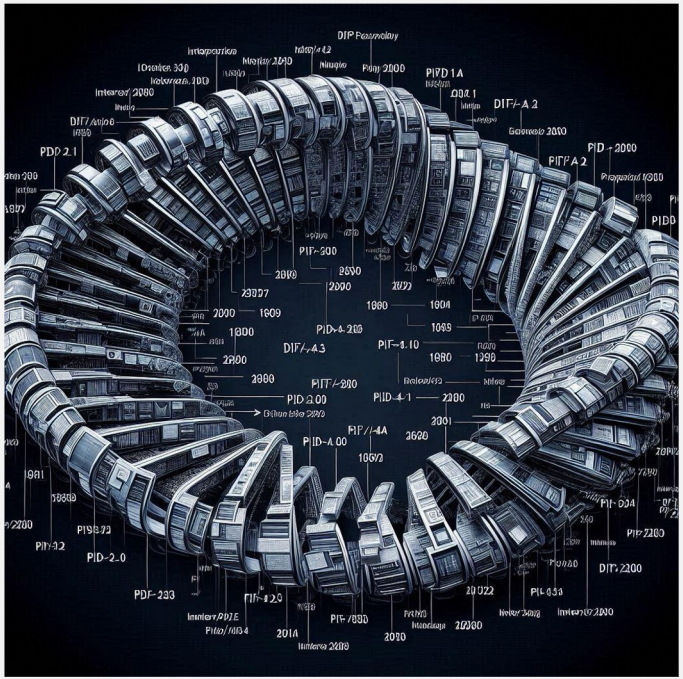


Key Dates in Digital Preservation



AI: Here to Help?





I had hoped the AI could save me some time with this presentation but when I asked Copilot in PowerPoint to create a double helix combining the previous timelines I got things like these.

Digital Preservation: 2005

Static Preservation

Digital preservation as a discipline began in the late 1990s.

Early practices included normalization and bit preservation.
Content was migrated (or normalized) to a small number of “trusted” formats.
NAA developed the XENA utility for “converting digital objects into open formats for preservation”.
NARA issued Transfer Guidance identifying 13 formats that could be used to transfer permanent electronic records. Normalization was the responsibility of the creating agency.

Offline Storage media, chosen for capacity and “durability”.
DLT tapes, CD & DVD media.
Media was refreshed. Typically every 5 years.

This approach is more similar to backing things up than digital preservation.
Poor ability to identify file formats or monitor for corruption.
Conversions between formats introduce numerous risks.

PDF/A-1 (2005)

PDF/A was developed during the era of normalization and bit preservation. It still accomplishes the original goals very well.

PDF/A:

- Ensures long-term accessibility and integrity of electronic documents.
- Preserves static visual appearance.
- Provides a self-contained, device independent, self-documenting data package.
- Restrictions: Prohibits features believed to compromise long-term archiving.
- Limited codec and media type support requires conversion to a supported filter (a motion GIF must be converted to a supported static raster filter).
- Fonts must be embedded.
- No external dependencies are allowed that could affect visual appearance.

2005 in Summary

The State of PDF and Digital Preservation

- Repositories attempted to limit the number of formats they preserved through normalization or policy enforcement.
- Preference for static content.
- Minimal external dependencies. Where possible external dependencies were eliminated often resulting in incomplete records.
- Preservation through bit copying.
- No continuous hash monitoring.
- Bit verification only conducted during refreshment.

Digital Preservation Today

Active Preservation

Repositories today provide:

- Cloud-based storage.
- Tiered storage in different geographic regions.
- Automated monitoring for alteration or corruption and refreshment.
- Domain specific metadata standards and controlled vocabularies for preservation (METS and PREMIS) describing each item's content, context, structure, and preservation history.
- The ability to capture and preserve content in native formats (even formats like Macromedia Flash, WordStar and Lotus 1-2-3).
- Management of preservation and access versions of of content (intellectual entities and representations).
- The ability to capture and maintain related content such as hyperlinks and attachments.

Formats are Changing

- Content is less likely to be created, used, and stored in binary formats such as TIFF and WordPerfect.
- Cloud-based content in Google Workspace, MS Office 365, email and social media platforms can be amorphous and difficult to capture.
- While format identification and verification tools (Apache Tika Vera PDF, Droid, TRID) have improved, examining attachments to identify personally identifiable information remains an issue.

PDF/A's Pros and Cons

Pros: PDF/A supports storing many types of data and provides lots of options for capturing metadata.

Los Alamos National Laboratory

Cons: It provides limited filters for raster data requiring normalization of unsupported codecs.

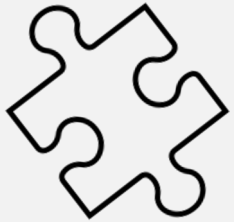
It can resemble a cardboard box.
If you have lots of boxes without labels it can be difficult to identify, store and treat them appropriately.

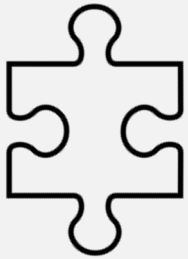
PDF/A 2030

Implementation and Adoption are Key

Today's implementations reflect legacy ideas with regards to digital preservation. Implementation policies and uniform adoption are the keys moving forward. Recent examples of work defining how the standards should be implemented include:

- EA PDF Specification for email.
- NARA's PDF Portfolio Guidance.
- C2PA (G+LAM)
- Doc RM





Why am I so excited?

PDF/A is well positioned to step in as a digital preservation target format:

- Wide native support across hardware and software ecosystems.
- Ubiquity
- International oversight and development.
- Declarations
- PDF/R
- Rich Media
- 3D PDF
- EA PDF
- Third party tools (Apache Tika, Vera PDF)
- PDF Association Whitepapers
- TC 171 SC2 + PDF Association LWGs and TWGs = continual improvement.
- GitHub open to the public to register issues.
- C2PA

The PREMIS Community when they saw C2PA Action Assertions

Sony Pictures

PREMIS Events

Most PDFs begin their lives as the result of a preservation event such as digitization, conversion, printing, or export. Unfortunately, information describing the type of event, the source of the content, and the agents involved has rarely been consistently captured.

Event

2.1 eventIdentifier (M, NR)
- 2.1.1 eventIdentifierType (M, NR)
- 2.1.2 eventIdentifierValue (M, NR)

2.2 eventType (M, NR)

2.3 eventDateTime (M, NR)

2.4 eventDetailInformation (O, R)
- 2.4.1 eventDetail (O, NR)
- 2.4.2 eventDetailExtension (O, R)

2.5 eventOutcomeInformation (O, R)
- 2.5.1 eventOutcome (O, NR)
- 2.5.2 eventOutcomeDetail (O, R)
  - 2.5.2.1 eventOutcomeDetailNote (O, NR)
  - 2.5.2.2 eventOutcomeDetailExtension (O, R)

2.6 linkingAgentIdentifier (O, R)
- 2.6.1 linkingAgentIdentifierType (M, NR)
- 2.6.2 linkingAgentIdentifierValue (M, NR)
- 2.6.3 linkingAgentRole (O, R)

2.7 linkingObjectIdentifier (O, R)
- 2.7.1 linkingObjectIdentifierType (M, NR)
- 2.7.2 linkingObjectIdentifierValue (M, NR)
- 2.7.3 linkingObjectRole (O, R)

C2PA Action Assertions

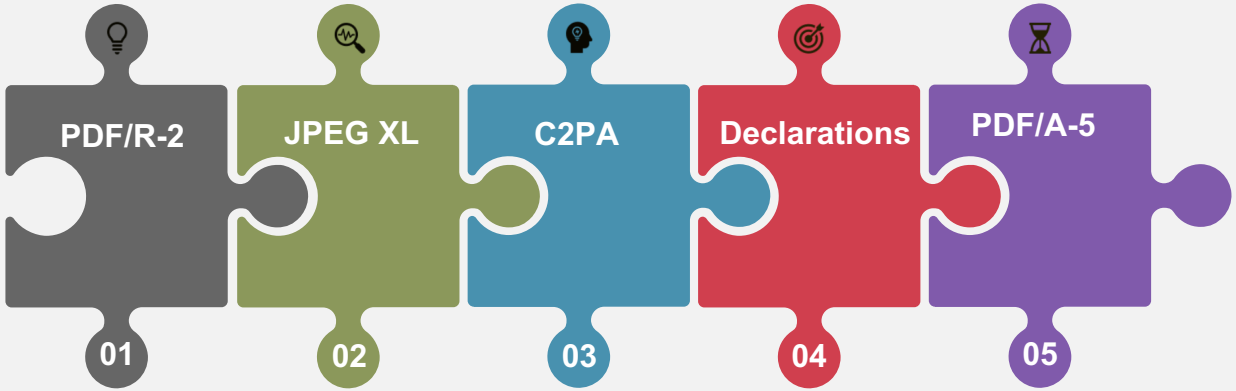
Coming to PDF Soon

C2PA incorporates the very similar concept of Actions. The Library of Congress is leading a group working to map metadata and help develop actions for specific media types.
Be sure to check out my G + LAM poster session.

c2pa.redacted	One or more assertions were redacted
c2pa.removed	A `componentOf` ingredient was removed.
c2pa.repackaged	A conversion of one packaging or container format to another. Content is repackaged without transcoding. This action is considered as a non-editorial transformation of the `parentOf` ingredient.
c2pa.resized	Changes to either content dimensions, its file size or both
c2pa.transcoded	A conversion of one encoding to another, including resolution scaling, bitrate adjustment and encoding format change. This action is considered as a non-editorial transformation of the `parentOf` ingredient.
c2pa.translated	Changes to the language of the content.
c2pa.trimmed	Removal of a temporal range of the content.
c2pa.unknown	Something happened, but the claim_generator cannot specify what.
c2pa.watermarked	An invisible watermark was inserted into the digital content for the purpose of creating a soft binding.

PDF/A 2030

Perfectly Preservable Packages



01 Scanner Support

PDF/R is created on a scanner.

02 Modern Codecs

Raster data is encoded using a modern codec such as JPEG XL

03 Authenticity

C2PA manifests describe the events that occurred during the creation of the PDF/A file.

04 Content Statements

Declare conformance with external standards such as HIPAA

05 Fit for purpose

A purpose built digital preservation package that is discoverable and whose contents are provably authentic.

PDF association

Questions?

Kevin L. De Vorsey
ThinkBox.DIGITAL
Kevin@thinkbox.digital

Thank you!