Benefits and Challenges of PDF Migration

August 12, 2020

Presented by Jessica Tieman, Digital Preservation Librarian, GPO

Anna Oates, Scholarly Communication and Discovery Services Librarian, Federal Reserve Bank of St. Louis, MO

What is Migration?

Illustration on digital preservation by Jørgen Stamp for https://digitalbevaring.dk, 2010

NDSA Levels of Digital Preservation (2013)

File Formats, Level Four:
Migration, emulation and similar activities as needed

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

When to Migrate?

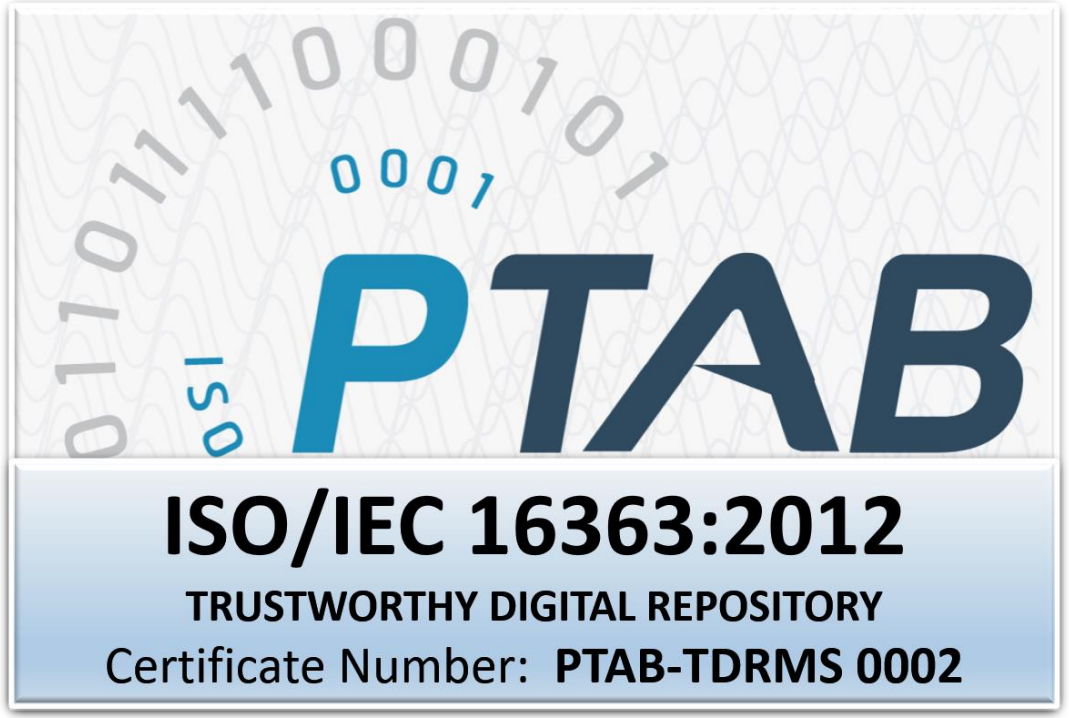
File Format Monitoring

- Library of Congress "Sustainability of Digital File Formats" https://www.loc.gov/preservation/digital/formats/index.html
- NARA "Digital Preservation Risk Assessment and Preservation Planning" https://www.archives.gov/preservation/electronic-records/digital-preservation-risk and "Digital Preservation Framework" https://github.com/usnationalarchives/digital-preservation
- Digital Preservation Coalition Watch Reports https://www.dpconline.org/technology-watch-reports/

GPO Digital Preservation & govinfo

2018 GPO becomes first digital repository in the US to become ISO 16363 certified

https://www.fdlp.gov/preservation/trusted-digital-repository-iso-16363-2012-audit-and-certification



Recommendation from auditors:

- Continue to work with Producers to improve standardization of the SIPs submitted.
- Current preservation system works well for current holdings, but GPO may need to prepare for new publishing paradigms

Why PDF Migration?

govinfo Browse About Developers Features Help Feedback

Discover U.S. Government Information

Search | Advanced

What are you searching for?

A to Z
Browse documents by alphabetical order

Category
Browse documents in specific collections

Date
Browse documents within a timeframe or date range

Recent

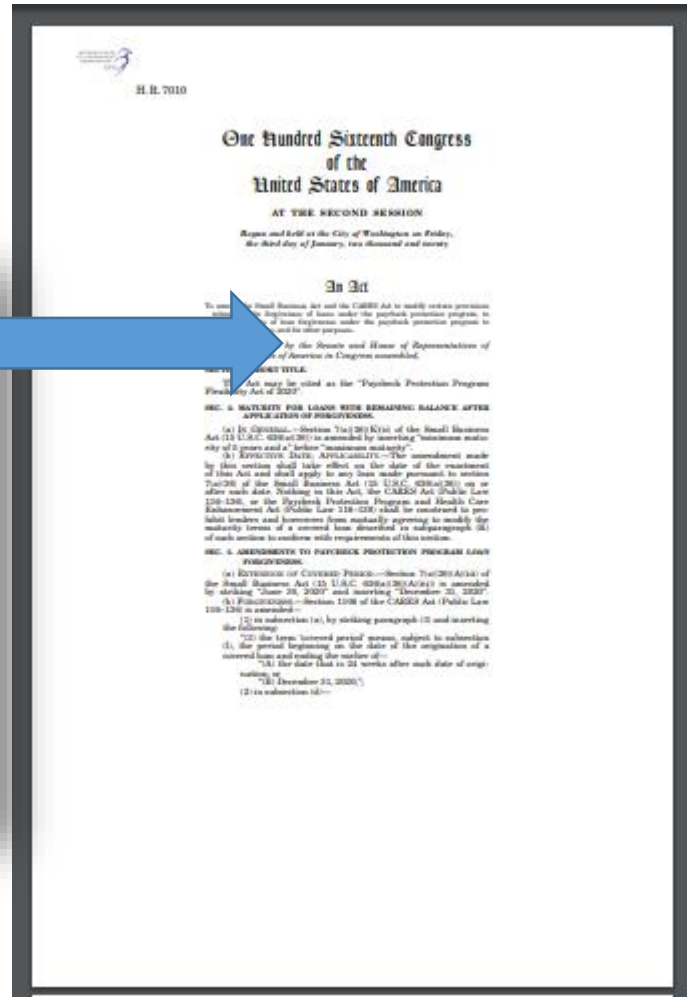
Congressional Record Daily Digest
Federal Register Table of Contents
House Calendar
Presidential Documents
Congressional Bills
All Documents

Popular

Budget of the U.S. Government

Trending

INVEST in America Act, H. Rept. 116-437
National Defense Authorization Act for FY 2021, S. Rept. 116-236
National Defense Authorization Act for FY 2021, S. 4049
George Floyd Justice in Policing Act of 2020, H.R. 7120
George Floyd Justice in Policing Act of 2020, H. Rept. 116-434, Part 1
JUSTICE Act, S. 3985
COVID-19 Related Documents



H. R. 7010

One Hundred Sixteenth Congress
of the
United States of America

AT THE SECOND SESSION

An Act

Digital Preservation Coalition 'Bit List' of Digitally Endangered Species (2019)

- Crowd-sourced, juried by over 27 prominent national libraries and archive institutions
- Creates a list of file formats and data types which are most "at-risk" or of concern for long-term preservation
- PDF/A deemed "vulnerable"
- All other PDF File formats deemed "endangered"

VULNERABLE

Digital materials are listed as Vulnerable when the technical challenges to preservation are modest but responsibility for care is poorly understood, or where the responsible agencies are not meeting preservation needs. This classification includes Lower Risk materials in the presence of aggravating conditions.

ENDANGERED

Digital materials are listed Endangered when they face material technical challenges to preservation or responsibility for care is poorly understood, or where the responsible agencies are poorly equipped to meet preservation needs. This classification includes Vulnerable materials in the presence of aggravating conditions.

https://www.dpconline.org/our-work/bit-list

We want PDF/A, right?

Institutions currently recommending PDF/A for preservation purposes:

- NARA https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#textualdata
- Library of Congress https://www.loc.gov/preservation/resources/rfs/textmus.html#digital
- FADGI (Federal Agency Digital Guidelines Initiative) http://www.digitizationguidelines.gov/guidelines/digitize-technical.html
- CENDI https://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf
- GPO https://www.fdlp.gov/preservation/preservation-at-gpo

What makes PDF/A more appropriate for long-term preservation?

PDF is currently in version 1.7, ISO 32000-1:2008

- Components of a PDF as a file format: line art, images, text, metadata, embedded objects or text, color schemas, object types which reference one another, font encoding and dictionaries, and occasionally interactive components such as JavaScript

PDF/A was introduced in 2008, the current standard is ISO 19005:2011, with accompanying PDF/A-2 19005-2 and PDF/A-3 19005-3 in 2012

- PDF/A serves a profile of syntax restrictions on features within a PDF (components) intended to ensure predictable visual representation of the document on all rendering software
- Examples of features which have restrictions: non-embedded fonts, JavaScript, audio and video content, LZW compression, non-embedded color spaces, encryption.

Are there risks to PDF/A Migration?

British National Library PDF/A Assessment (March 2019):

"wholesale migration of a PDF collection to PDF/A is unwise"

Institutions must truly understand all *significant properties* before migration of data

Is Migration really the "solution" ?

Are there ways to characterize our data without the risks of data loss we might face during migration ?

>Implementing PDF/A validation software first, is a potential solution

Is a migration pilot the only way to determine the feasibility of wholescale migration ?

>A better characterization of our content can better inform, and make more specific, what sort of "trigger events" would dictate migrating or not migrating

Is something "at risk" simply because it isn't PDF/A ?

>A "quality" PDF which is usable, able to be rendered, and accessible is always better than a PDF/A which is corrupted, has data loss, or is missing significant features intended for use by its creator or user community

References

British Library Digital Preservation Team. "PDF Format Preservation Assessment: Part 2: PDF/A Profile." 30 June 2019. 24 July 2020. https://wiki.dpconline.org/images/2/22/PDFA_Assessment_v1.0.pdf.

CENDI Digital Preservation Task Group. "Formats for Digital Preservation: A Review of Alternatives and Solutions." 1 March 2007. 24 July 2020. https://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf.

Digital Preservation Coalition. "The BitList 2019: The Global List of Digitally Endangered Species." Vers. 2nd Edition. November 2019. 24 July 2020. https://www.dpconline.org/docs/miscellaneous/advocacy/wdpd/2156-bitlist2019-report/file.

Johnston, Leslie. "PDFS: When a Standard isn't Standard in your Collections." 18 November 2018. 25 July 2020. https://www.dpconline.org/blog/idpd/pdfs-when-a-standard-isn-t-standard-in-your-collections.

Klindt, Marco. "PDF Considered Harmful for Digital Preservation." 2017. *iPRES Proceedings.* 24 July 2020. https://ipres2017.jp/wp-content/uploads/15.pdf.

Library of Congress. "PDF is Here to Stay: Archiving with the Portable Document Format." 10 March 2020. 24 July 2020. https://blogs.loc.gov/thesignal/2020/03/pdf-is-here-to-stay.

—. "PDF/A Family, PDF for Long-term Preservation." 20 March 2019. *Sustainability of Digital Formats: Planning for Library of Congress Collections.* 24 July 2020. https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml.

—. "Recommended Formats Statement." 2016. 24 July 2020. https://www.loc.gov/preservation/resources/rfs/textmus.html#digital.

National Archives and Records Administration. September 2019. *Records Management Regulations, Policy, and Guidance: Appendix A: Tables of File Formats.* 24 July 2020. https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#textualdata.

National Library of Medicine . "PDF File Migration to PDF/A: Technical Considerations ." 2007. *Proceedings of IS&T Archiving Conference.* 25 July 2020. https://lhncbc.nlm.nih.gov/system/files/pub2007020.pdf.

Oates, Anna I., et al. "Navigating the PDF/A Standard: A Case Study of Theses in Oxford's Institutional Repository." 2018. 24 July 2020. http://hdl.handle.net/2142/100236.

PDF/A Continued

About the format, cases of risk, institutional perspectives

PDF/A Specification

ISO 19005

Electronic document file format for long-term preservation

International Organization of Standardization specifications for long-term preservation of PDF 1.4.

Version

- PDF/A-1, based on PDF 1.4, formalized in ISO 19005-1:2004
- PDF/A-2, based on PDF 1.7 formalized in ISO 19005-2:2011
- PDF/A-3, based on PDF 1.7, formalized in ISO 19005-3.2012
- PDF/A-4, based on PDF 2.0, formalized in ISO 19005-4 [*under development*]

Conformance

- Conformance Level A (accessible)
- Conformance Level B (basic)
- Conformance Level U (Unicode)
- PDF/UA (Universal Accessibility)

> Because of the potential complications for the long-term preservation of PDF/A-3 files…recommends that tools that create PDF/A-compliant documents be engineered to identify (through the pdfaid:part element) files that have no embedded files, or whose embedded files are all in PDF/A format, as compliant with PDF/A-2 rather than PDF/A-3 (p. 10)

Arms, C., Chalfant, D., DeVorsey, K., Dietrich, C., Fleishhauer, C., Lazorchak, B., Morrissey, C. & Murray K.. (2014). *The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions: An NDSA Report*. http://www.digitalpreservation.gov/documents/NDSA_PDF_A3_report_final022014.pdf

Oettler, A. (2013). PDF/A in a Nutshell. PDF Association. http://www.pdfa.org/resource/pdfa-in-a-nutshell-2.0

Cases of Loss

τὸς ἀγαθὸς ἔστερξεν Ἄρης, ἐφίλησε δ' ἔπαινος, |

καὶ γήραι νεότης οὐ παρέδωχ' ὑβρίσαι· |

ὦγ καὶ Γ[λ]αυκιάδης δήιος ἀπὸ πατρίδος ἔργων |

ἦλθ' ἐπ[ὶ] πάνδεκτον Φερσεφόνης θάλ<α>μον (= *IG* II² 10998, *GVI* 1637).

ἔργων

Image from Microsoft Word source file

τὸς ἀγαθὸς ἔστερξεν Ἄρης, ἐφίλησε δ' ἔπαινος, |

καὶ γήραι νεότης οὐ παρέδωχ' ὑβρίσαι· |

ὦγ καὶ Γ[λ]αυκιάδης δήιος ἀπὸ πατρίδος ἔργων |

ἦλθ' ἐπ[ὶ] πάνδεκτον Φερσεφόνης θάλ<α>μον (= *IG* II² 10998, *GVI* 1637).

ἔργων

Image from file created as PDF and conformed to PDF/A-2a with Adobe Acrobat DC

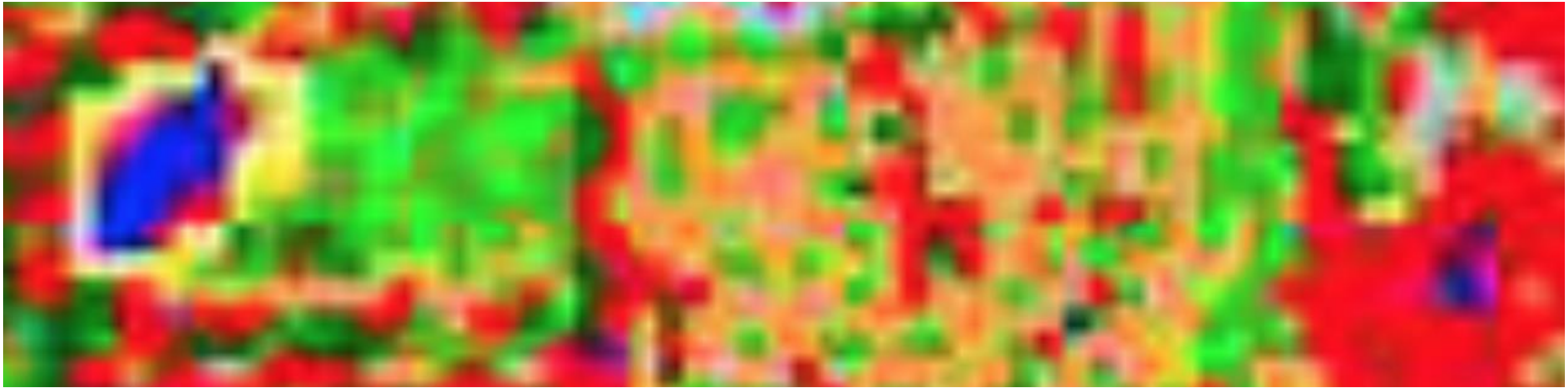


Image of embedded hyperspectral image in source PDF (version 1.5) file, where Interpolate key = “true.

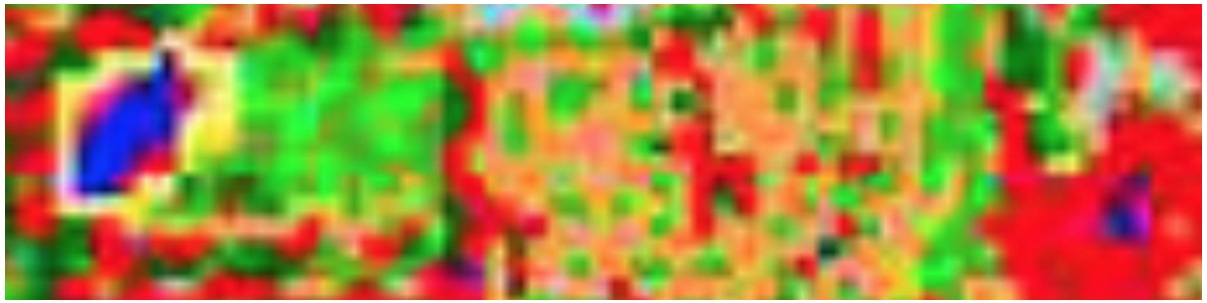


Image of embedded hyperspectral image in source PDF (version 1.5) file, where Interpolate key = “true.

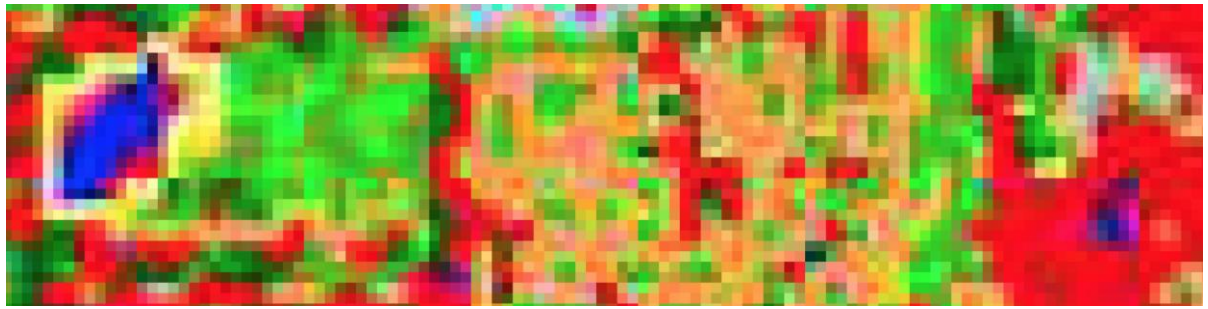


Image of embedded hyperspectral image in PDF/A-2a migrated with callas pdfaPilot.

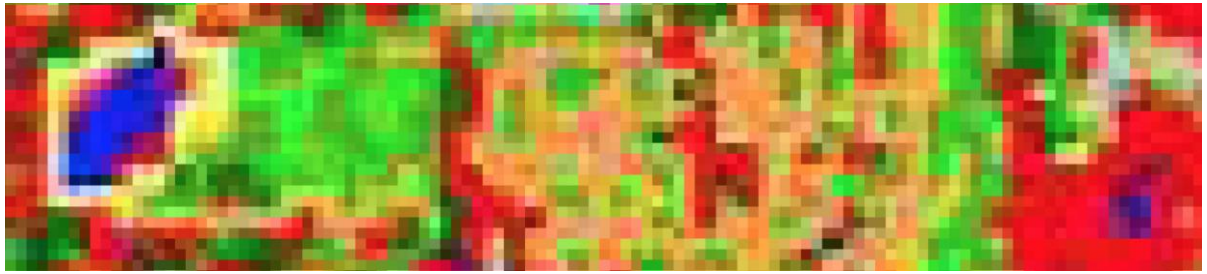


Image of embedded hyperspectral image in PDF/A-2a migrated with Adobe Acrobat DC.

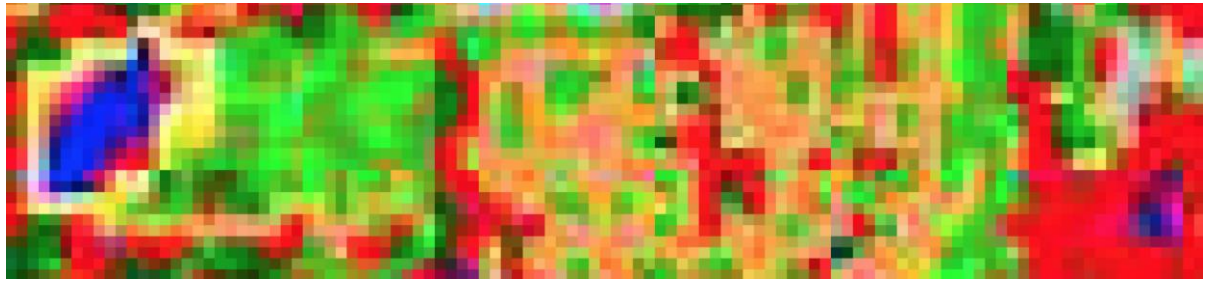


Image of embedded hyperspectral image in in PDF/A-2a migrated with PDFTron PDF/A Manager.

Archaeology Data Service on PDF/A-3

> By allowing the association of original data streams, creators can extend the potential for preservation and reuse of information in both the short and long term. Of course this new specification will require the development of new strategies to preserve these associated data streams.

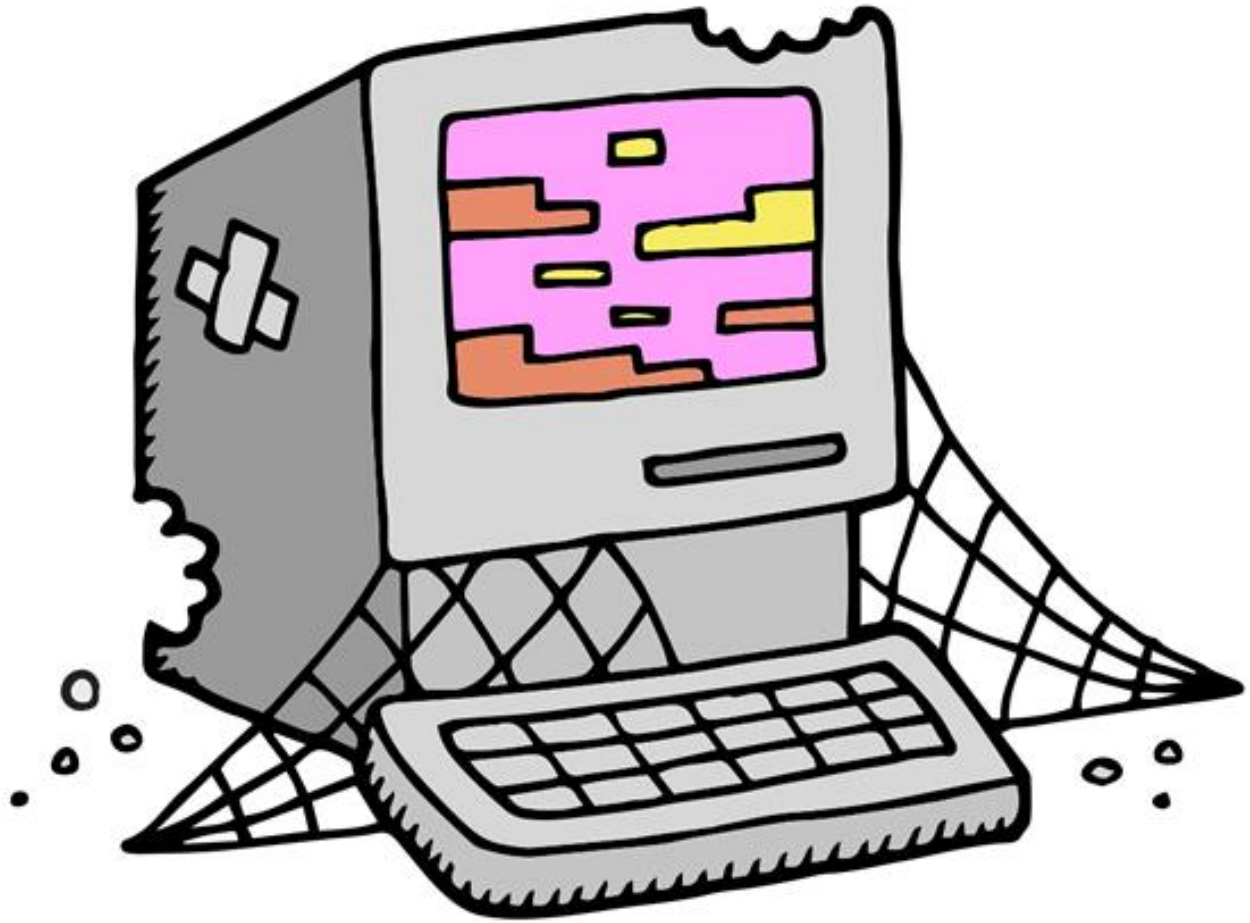
> What is more problematic, however, is that much of this appended content will lack the appropriate metadata that can provide important contextual information about complex data streams, assist in the assessment of the significant properties, and aid in the development of digital preservation strategies.

Moore, R., & Evans, T. (2013). Preserving the grey literature explosion: PDF/A and the digital archive. *Information Standards Quarterly 25*(3), 26. https://doi.org/10.3789/isqv25no3.2013.04

Institutional Perspectives

Depository Libraries

What you get is what you have



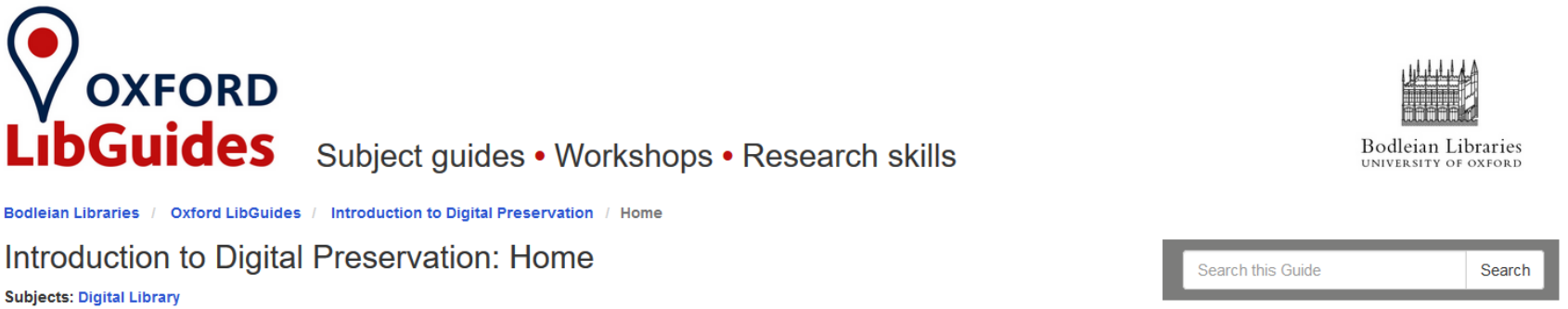
SPRUCE Digital Preservation Illustrations https://wiki.dpconline.org/, CC-BY-NC 3.0

Research Institutions

Instructional preservation with creation

- Virginia Tech, Developing your ETD http://etd.vt.edu/etddev.html
  - Virginia Tech requires that students deposit theses and dissertations as PDF/A-1b
- Introduction to Digital Preservation https://libguides.bodleian.ox.ac.uk/digitalpreservation/home

Converting to PDF/A Format. Carleton University. https://gradstudents.carleton.ca/thesis-requirements/pdfa-formatting/ . *Note: Carleton requires students deposit their theses as PDF/A.*



Introduction to Digital Preservation. University of Oxford. https://libguides.bodleian.ox.ac.uk/digitalpreservation

Format Considerations

- Significant properties by resource type and research domain
- Prohibitive document features
- Automated risk assessment

Thank you

Resources

- Arms, C., Chalfant, D., DeVorsey, K., Dietrich, C., Fleishhauer, C., Lazorchak, B., Morrissey, C. & Murray K. (2014). The benefits and risks of the PDF/A-3 file format for archival institutions: An NDSA report. http://www.digitalpreservation.gov/documents/NDSA_PDF_A3_report_final022014.pdf
- Mason, S., & Halvarsson, E. (2018). Digital preservation at Oxford and Cambridge training programme pilot. University of Oxford. https://doi.org/10.5287/bodleian:yrPm5qnaR
- May, P., Pennock, M., & Russo, D. A. (2019). The integrated preservation suite: Scaled and automated preservation planning for highly diverse digital collections. iPRES 2019: 16th Annual Conference on Digital Preservation Proceedings. https://doi.org/10.17605/OSF.IO/BCV2E
- Moore, R., & Evans, T. (2013). Preserving the grey literature explosion: PDF/A and the digital archive. *Information Standards Quarterly* 25(3), 20-27. https://doi.org/10.3789/isqv25no3.2013.04
- Moore, R., & Evans, T. (2014). The use of PDF/A in digital archives: A case study from archaeology. *International Journal of Digital Curation* 9(2), 123-138. https://doi.org/10.2218/ijdc.v9i2.267
- Oates, A. (2018). Navigating the PDF/A standard: a case study of theses in the University of Oxford's institutional repository, [Master's thesis, University of Illinois at Urbana-Champaign]. *IDEALS*. http://hdl.handle.net/2142/100913
- Oettler, A. (2013). PDF/A in a Nutshell. PDF Association. http://www.pdfa.org/resource/pdfa-in-a-nutshell-2.0

Resources: Standards

- International Organization for Standardization. (2005). Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF 1.4 (PDF/A-1) (ISO 19005-1:2005). https://www.iso.org/standard/38920.html
- International Organization for Standardization. (2008). Document management—Portable document format—Part 1: PDF 1.7 (ISO 32000-1:2008). https://www.iso.org/standard/51502.html
- International Organization for Standardization. (2011). Document management—Electronic document file format for long-term preservation—Part 2: Use of ISO 32000-1 (PDF/A-2) (ISO 19005-2:2011). https://www.iso.org/standard/50655.html
- International Organization for Standardization. (2012). Document management—Electronic document file format for long-term preservation—Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3) (ISO 19005-3:2012). https://www.iso.org/standard/57229.html
- International Organization for Standardization. (2017). Document management—Portable document format—Part 1: PDF 2.0 (ISO 32000-2:2017). https://www.iso.org/standard/63534.html
- International Organization for Standardization. (n.d.). Document management—Electronic document file format for long-term preservation—Part 4: Use of ISO 32000-2 (PDF/A-4) (ISO/PRF 19005-4). https://www.iso.org/standard/71832.html