Policies, Risks and Strategies:

A File Format Debate

Sam Alloing
National Library of the Netherlands (KBNL)
sam.alloing@kb.nl
0000-0002-1254-1483

Valentijn Gilissen
Data Archiving and Networked Services (DANS)
the Netherlands
valentijn.gilissen@dans.knaw.nl
0000-0003-2399-7598

Leslie Johnston
National Archives and Records Administration (NARA)
United States
leslie.johnston@nara.gov
0000-0001-9908-0183

Kate Murray
Library of Congress (LoC)
United States
kmur@loc.gov
0000-0003-1325-0829

Tyler Thorsted
Brigham Young University (BYU)
United States
thorsted@byu.edu
0000-0003-0292-0962

Paul Wheatley
Digital Preservation Coalition (DPC)
United Kingdom
paul@dpconline.org
0000-0002-3839-3298

The digital preservation community has been developing approaches to preserving the meaning of digital content for a number of decades. But questions still remain as to the most accurate, practical, timely and cost effective way of keeping our data usable. Collating and presenting file format policies from several organizations triggered a lively panel discussion in early 2023. This panel session will build on the success and popularity of that debate by bringing in new voices and topics raised by the audience. This subject is a critical one to understand if we are to be successful in preserving our data for future generations.

Keywords – File formats, file format policy, file format assessment, preservation planning, preservation strategy

Conference Topics – Sustainability: Real and Imagined, From Theory to Practice

1. Background

The International Comparison of Recommended File Formats [1] collates file format policies from 28 organizations from around the world. Paul Wheatley published a blog post which questioned a number of aspects of this work [2]. On February 9th 2023 Sam Alloing (KBNL) moderated a panel debate between Valentijn Gilissen (DANS) and Paul Wheatley (DPC), entitled "Do unacceptable file formats exist?". The event drew a crowd of 200 people and provoked an almost overwhelming degree of comment and engagement from the audience. This panel session aims to build on the success of that debate by bringing an extended panel of diverse opinions to the iPres Conference.

2. Contrasting Approaches To Preservation Strategy

A strategy used for file format preservation is the use of preferred file formats. In this strategy a file format is identified as preferred if it complies with some defined criteria. DANS has such a Preferred Formats policy [3], as does NARA [4], the LoC [5], KBNL [6] and others. For both DANS and NARA, a file acquired in a non-preferred file format is migrated to a preferred file format if possible and the original retained. At the KBNL, all file formats of a publisher are allowed and preserved. The KBNL's policy is to assign file formats a 'knowledge level'. This is the status of a file format in the repository and indicates what preservation operations are possible. For

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121082

example the first level is 'stored file', this means that the file is only bit-preserved. The third and last level is known 'file format', where the results of identification, validation and technical metadata extraction can be interpreted and guidelines for each format have been formulated. Preferred formats are typically identified through generic criteria such as age, tool support, complexity, documentation, risk and context. Examples include the LoC Recommended Formats Statement evaluation matrix and the NARA Risk Assessment Matrix [5, 7]. The resulting data would then be used to determine file format policy and ultimately which formats to migrate. An opposing view was offered by van der Knijff who argued that such risk factors were largely theoretical [8]. Rosenthal argued that "format obsolescence is a rare problem" due in large part to the availability of open source rendering tools [9]. Just Solve has focused on documenting and web archiving sources of information on file formats [10].

3. Broadening The Debate

The variety of perceptions in the world of digital preservation may seem to conflict with each other. Having an open debate about these subjects provides a fruitful basis for sharing knowledge and gaining consensus. This panel session will continue, broaden and extend the debate held in February 2023. It will incorporate the diverse viewpoints of several members of the audience of that original debate. Leslie Johnston (NARA) brings the stark challenges of the long-term preservation of an ominously large range of different file formats. Kate Murray (LoC) brings experience of researching and accessing file formats through her leadership of the Sustainability of Digital Formats and Recommended Formats Statement. Tyler Thorsted (BYU) brings a track record of contributions to the PRONOM and Just Solve registries. Leslie, Kate and Tyler will join the panelists of the original debate: Valentijn Gilissen oversees the file format guidelines of the Dutch national centre of expertise and repository for research data (DANS) in his role as preservation officer. Paul Wheatley is Head of Research and Practice at the Digital Preservation Coalition. Sam Alloing (KBNL) actively contributed to the Guide to Preferred File Formats of the Dutch Digital Heritage Network (DDHN) and the analysis of the File Format Lifecycle also from the DDHN.

4. Format Of The Panel

Following short introductions from each of the panelists the session will move to a question and discussion format, moderated by Sam Alloing. It will focus primarily on questions of file format policy and digital preservation strategy. The considerable text chat from the February panel discussion will be used as a source of topics for discussion. The panel will ensure strong audience participation by both accepting questions from them and posing live poll questions to them. This will provide an impression of the state of play for preservationists represented at iPres alongside the viewpoints of the panel members. Activating the audience with poll questions demonstrated a meaningful and active discussion in the February debate, so we would like to replicate that approach here. The format allows remote and in-person participation.

Key questions for discussion by the panel members include: 1) What are the criteria for file format assessment in the global and institutional contexts? 2) Is there such a thing as a "good, bad or unacceptable" format? 3) What goes into risk assessment for file formats? 4) How do file format risks compare to other risks in the field of digital preservation? 5) What strategies are used for assessing file format risks?

5. REFERENCES

[1] International Comparison of Recommended File Formats [Online] https://openpreservation.org/resources/member-groups/international-comparison-of-recommended-file-formats/

[2] Wheatley, P. "File format recommendations…" Blog post, DPC, [Online] https://www.dpconline.org/blog/file-format-recommendations

[3] Laagland et al, "White paper on preferred formats" http://doi.org/10.5281/zenodo.4518486, p 19

[4] DANS Preferred Formats guidelines [Online], https://dans.knaw.nl/en/file-formats/

[5] NARA Format Guidance for the Transfer of Permanent Electronic Records [Online], https://www.archives.gov/records-mgmt/bulletins/2014/2014-04.html

[6] Library of Congress Recommended Formats Statement [Online], https://www.loc.gov/preservation/resources/rfs/

[7] NARA Digital Preservation Framework [Online], https://github.com/usnationalarchives/digital-preservation

[8] Van der Knijff, J. "Assessing file format risks: searching for Bigfoot?", Blog post, Bitsgalore [Online] https://www.bitsgalore.org/2013/09/30/assessing-file-format-risks-searching-bigfoot

[9] Rosenthal, D. "Format Obsolescence: Assessing the…" https://web.stanford.edu/group/lockss/resources/2010-06_Format_Obsolescence.pdf

[10] ArchiveTeam File Format Wiki [Online] http://fileformats.archiveteam.org/wiki/Main_Page