iPRES 2023

Do Unacceptable Formats Exist?

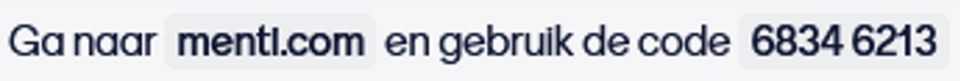
Policies, Risks and Strategies: A File Format Debate

Sam Alloing, Valentijn Gilissen, Leslie Johnston, Kate Murray, Micky Lindlar, Tyler Thorsted

Wednesday, September 2023 - 3:30 pm - 5:00 pm
SUS-3 - Heritage Hall 2

Agenda

- Viewpoints of panellist
  - Online participants add your viewpoint on unacceptable/preferred formats in the chat
- Your opinion on some questions
  - Online and in-person using Mentimeter

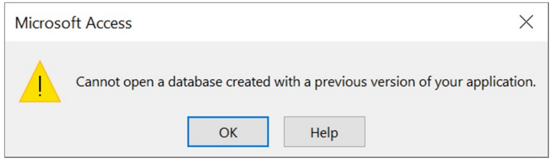


Ga naar menti.com en gebruik de code 6834 6213

- Discussion
  - In-person: use the microphones!
  - Online: use the chat or hand raise option
- Conference Bingo
  - Panellist try to use Bingo Words

https://openpreservation.org/blogs/monitoring-disappearing-file-formats-4-dans/

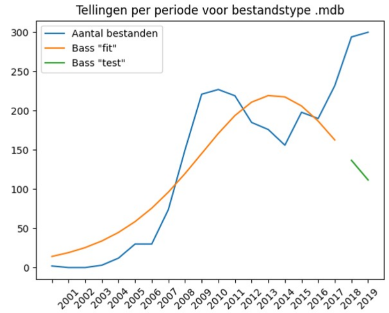
Valentijn’s vision

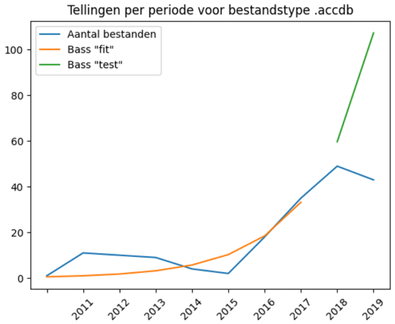

Microsoft Access

Cannot open a database created with a previous version of your application.

netwerk
digitaal
erfgoed

- Formats can still be in danger of falling into disuse and becoming unsupported. Long-term preservation should aim to make use of robust file formats
- Ideally, file formats most suitable for preservation have open specifications; are independent of specific software, developers or vendors; are frequently used
- If it is not possible to obtain preferred formats, non-preferred formats should still be accepted, although less long-term guarantees can be given
- When formats are converted/migrated, the original data should be preserved as well
- Certain non-preferred formats can be made available as current usage formats
- File format strategies need to be informed by expert input from data users and communities
- A policy with ‘non-acceptable formats’ risks receiving incomplete or lower-quality data, but a policy with ‘acceptable formats’ risks a lack of any effort made for sustainability
- Emulation may certainly offer helpful solutions, but relying on emulation may simply shift preservation concerns from formats to software





The 4CH project will design and prepare for a European Competence Centre (CC) on the Conservation of Cultural Heritage which will work proactively for the preservation and conservation of cultural heritage (CH).

https://dans.knaw.nl/en/file-formats/

4CH COMPETENCE CENTRE FOR THE CONSERVATION OF CULTURAL HERITAGE

Data Archiving and Networked Services
DANS

Sam’s vision

- We accept every file format from a publisher
- Preferred file formats policy created problems with bad file format migration
- The reason to implement preferred file format policy was ‘we can’t know all formats and support all formats with tools’
    - We have 1000+ different file extensions
- Switched to Knowledge Levels
    - How much file format information is known for our formats
    - Way to communicate the knowledge we have of a file format
    - Level 1: Stored/Level 2: Identified/Level 3: Know

Leslie’s vision

- After collecting born-digital records since 1971, at the US National Archives (NARA) we have seen that format obsolescence is real.
- NARA has a list of Preferred and Acceptable formats as guidance to agencies: https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html. It is regularly updated based on input from agencies about what format they are transferring.
- NARA has put considerable effort into developing extensive criteria for identifying preferred and acceptable formats and for Preservation Plans in its Digital Preservation Framework: https://github.com/usnationalarchives/digital-preservation. We also define essential characteristics for several categories of records. The plans are updated quarterly.
- NARA supplies file format guidance to agencies, but is not proscriptive and accepts variations of formats and formars not on its list. This may seem contradictory, but as a government archive accepting records that can be 5-15-25+ years old, we do our best to take what’s on offer.
- Transfer and Preservation and Access formats are not necessarily the same.
- NARA has not yet done any experimentation with emulation, given the several hundred variants of formats covering 40+ years of transfers.

Tyler’s Viewpoint

- BYU currently does not have an approved format policy. We have **preferred formats for our internal digitization and reformatting work**, but for born-digital we have no formal policy.
- I prefer action plans and strategies over general policies for born-digital collections.
- I think many of the preferred format policies out there are too restrictive and are not focused on the right things.
- Preservation actions should be based on **properties of a file format** not the format itself.
- The only way to know if a file is a risk is to understand more about the format **beyond its extension**.
- If a format is **well documented** and understood, the preservation risk is lower.
- I also lean towards levels of preservation strategy, some f**iles get more attention than others**.
- We are creating a new **file format strategy** to document the formats we have in our repository and assign action plans.

Kate’s Viewpoint

- **Context** is essential to format evaluation (which supports the idea that there are no inherently good or bad formats). “It depends” is a legit answer.
- In the **Sustainability of Digital Formats**, the Library of Congress uses seven sustainability factors (including disclosure and adoption) as well as more specific quality and functionality factors for content types, to assess and describe the ability of the Library to preserve content in a given format.
    - https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml
- We have take both **global and local factors** into account in our **Recommended Formats Statement** for identifying a format as preferred or acceptable (and acceptable doesn’t mean bad). RFS is updated regularly/yearly.
    - https://www.loc.gov/preservation/resources/rfs/index.html
- Evaluation matrix global factors (aka sustainability factors); local factors = Staff experience and expertise; Software/Hardware/OS available; Representation/extent in LC collections/storage; Established workflow/functionality; Access options
- One of the Digital Collections Strategy, FY 2022-2026 objectives is, “Transition to e-preferred collecting as appropriate.”

Micky’s viewpoint

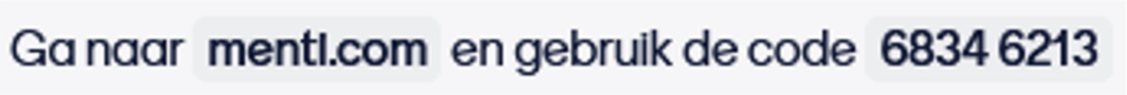
- TIB does not recommend or require specific file formats for born-digitals - currently 238 different file formats based on PUIDs

Unacceptable formats should not exist, because:

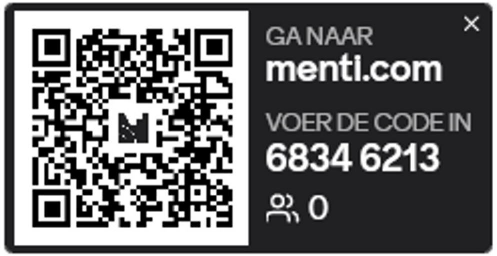
- They can be (unintentional) gatekeeping / not collecting vital information is much worse than collecting it in a „not ideal“ file format
- We makes assumptions about producers‘ technological knowledge & capability
- There are too many collection areas for which limitations are not feasible (e.g., data carrier images, software, web archives) – so why limit some workflows to recommended, but not others?
- Obsolescence is real, but always subjective & relative: it describes the effort that needs to go into making data interpretable via different channels and for different usage scenarios
- The risk of losing information during normalization is too big

Over to you!

https://www.menti.com/al5qdm23d3z2



Ga naar menti.com en gebruik de code 6834 6213



GA NAAR

menti.com

VOER DE CODE IN

6834 6213

0

Discussion

- What is the difference between Libraries and Archives when it comes to (preferred/unacceptable) file formats?

Discussion

- Is there such a thing as a “good, bad or unacceptable” file format?

Discussion

- What goes into risk assessment for file formats?

Discussion

- What preservation strategies can be used to manage ‘unacceptable’ formats?

Discussion

- When do you take a preservation action? At ingest or just in time or …?