Digital file formats for long-term preservation

Tracey Krause
Yvonne Loiselle
Nov. 16, 2005

Introduction

- Our position within InterPARES
- Definitions
- Organization of the research
    - Part One
        - Research Questions
            - File format criteria
            - Recommendations
    - Part Two
        - Email Survey
- Current Status

Definitions: “File format”

- The term has been accepted by the Terminology Cross-Domain for definition; definitions awaiting acceptance:
  - The organization of data within files, usually designed to facilitate the storage, retrieval, processing, presentation, and/or transmission of the data by software. - [Archives]

Definitions: “File format” cont.

- Waiting acceptance cont.
  - The way in which the information in a file is encoded. There are many proprietary formats – nearly every application has its own, often changing with new versions – as well as standard file formats such as RTF, TIFF, and EPS. In some systems, such as Apple Macintosh, the information about file format and originating application is part of the file, but in other systems it is up to the user to know what the format is, although there are more-or-less strict file-naming conventions. The multiplicity of file formats is a continuing problem for both software developers and users. - [Computer and Information Sciences]

Definitions: “File format” cont.

- The institutions studied did not define “file format” in their policies and recommendations.

Are HTML and XML file formats?

- HTML is not a file format, but rather a set of tags and attributes used to structure data in a file in order for it to be read by a web browser.
- XML is not a file format, but rather a metalanguage that can be used to structure data – i.e. to create file formats.

Definitions: “Open” file format

- The terminology varies:
  - Open format
  - Open standard format
  - Non-proprietary format
  - Standard format
- What are the criteria?

Definitions: “Standard file format”

- A format for which there is a published specification, and/or
- A format that is widely used, and/or
- A format for which there has been an initiative to adopt it as an industry standard, and/or
- A format that has been standardized by a recognized organization, such as ISO or ANSI.

The research

- The recommendations of approximately 40 institutions were surveyed, with a view to discerning the existence of any consensus on desirable format attributes. 21 institutions were examined in detail.
- Survey conducted via email of 15 artistic and 10 scientific institutions comprised of 5 questions.

Part One: The research questions

- Do existing guidelines on recommended file formats for electronic records preservation constitute a coherent body of knowledge on which to base model policies?
- Should existing guidelines on recommended file formats for electronic records preservation be translated into model policies on file formats for electronic records creation?

Criteria for selection of file formats for long-term preservation

- Widespread use (15 institutions)
- Non-proprietary or “open” (14 institutions)
  - Proprietary OK (7 institutions)
- Published file specification (5 institutions)
  - “Well documented” (6 institutions)
- Platform independence (8 institutions)

Criteria for selection con’t

- No compression (5 institutions)
  - Lossless compression only (7 institutions)
  - Lossy compression OK (1 institution)
  - Non-proprietary compression algorithm only (2 institutions)

Criteria for selection con’t

- Supported by standards (4 institutions)
- Backward compatibility (3 institutions)
- No encryption (3 institutions)
- Accommodates internal metadata (3 institutions)
- Ability to map embedded metadata to other formats (2 institutions)

Criteria for selection con’t

- Freely available viewers (2 institutions)
- Low proliferation of versions (2 institutions)
- “Stable” (2 institutions)
- “Well-supported” (2 institutions)
- Highly tested (2 institutions)

Criteria for selection con’t

- Small file size (1 institution)
- Rendering accuracy (1 institution)
- Ability to create deliverable images from digital masters (1 institution)
- Robust (1 institution)
- “Good software support” (1 institution)

Suggested answers to the research questions

- Do existing guidelines on recommended file formats for electronic records preservation constitute a coherent body of knowledge on which to base model policies?
  - Yes, there is consensus on a number of points. However, the language needs to be clarified so every institution is talking about the same thing.

Suggested answers to the research questions

- Should existing guidelines on recommended file formats for electronic records preservation be translated into model policies on file formats for electronic records creation?
  - Yes, if archives are going to limit the number of file formats they will accept for long-term preservation.

Part Two: Survey; question one

1. Does your organization preserve digital records over the long term?
   - What do you consider to be the long term?
   - Why (for whom and for what purpose) do you preserve the records?

Part Two: Survey; question two

2. Does your organization limit the number of file formats that you select for long term preservation?

Part Two: Survey; question three

3. If no, how are you planning to preserve all the different file formats?

Part Two: Survey; question four

4. If yes, what are the formats, and what are the criteria you use for selecting these formats?

Part Two: Survey; question five

5. Do you have any written policies or procedures on this subject that you can send us? May we refer to these written materials in our research?

Institutions/research groups studied

- Digital Preservation Coalition, United Kingdom
- Arts and Humanities Data Service, United Kingdom
- Research Libraries Group and Digital Library Federation, United Kingdom
- California Digital Library, United States
- EU-US Working Group on Spoken-Word Audio Collections, International
- On-line Computer Library Center, United States
- Bible for the Future, United Kingdom

Institutions/research groups studied

- Cornell University Library, United States
- DAVID (Digital Archiving: Guideline and Advice), Belgium
- Digital Image Archive of Medieval Music (DIAMM), Universities of Oxford and London, United Kingdom
- Swiss Federal Archives
- Netherlands Institute for Scientific Information Services
- Technical Advisory Service for Images, United Kingdom
- National Archives of Australia

Institutions/research groups studied

- U.S. National Archives and Records Administration
- Ohio Electronic Records Committee
- The State and University Library, Arhus; The Royal Library, Copenhagen, Denmark
- VERS - Victorian Electronic Records Strategy, Australia
- Library and Archives Canada
- MIT, Hewlett-Packard (DSpace), United States
- Florida Center for Library Automation, United States

Survey Institutions - Respondents

- Artistic
  - Beethoven-Haus Bonn
  - MediaRights
  - Digital Media (UC Berkeley – Pacific Film Archive)
- Scientific
  - USGS – Gigapolis Project
  - Cornell University Geospatial Information Repository (CUGIR)

Current Status

- Re-visiting survey respondents
- Target
  - Ten institutions
  - In-depth responses