InterPARES 3 Project

International Research on Permanent Authentic Records in Electronic Systems

TEAM Korea

XML: Examining the Criteria to be an Open Standard File Format

InterPARES 3 Project
4th International Symposium

17 September 2010
Oslo, Norway

Eun Park

IP3 TEAM Korea

McGill University

Table of Contents

- Why file format is important?
- Research questions
- Basics of XML
- Open standard format
- Criteria of reviewing file formats
    1) Autonomy family;
    2) Interoperability family;
    3) Authenticity family; and
    4) Functionality family.
- Issues

- “*The ever-growing complexity and heterogeneity of digital file formats together with rapid changes in underlying technologies have posed extreme challenges to the longevity of information*” (Becker, et al. 2008).

Why Is File Format Important?

- Most file formats are **proprietary and dependent** on various operating systems, hardware and software combinations.
- Preservation file format vs. Access file format.
- Three main file formats: TIFF (GIF, JPEG); PDF (and PDF/A); ODF; various XML subsets.
- In order to select the best file format, various criteria have been proposed.

Research Questions

1) What are the characteristics of XML file formats, according to the defined requirements?
2) What are the basic requirements of open standard file formats for long-term preservation?
3) What are the recommendations for open file formats that meet the criteria for long-term preservation?

XML

- XML (eXtensible Markup Language) developed under the direction of W3C.
- XML as an open specification.
- XML is compatible with SGML, human-legible, easy to create and clear to understand.
- The W3C officially recommended XML Version 1.0 in 2008.
- Numerous subsets of XML exist.
- The Office Open XML specification has been an open standard file format by ISO and IEC as an International Standard (ISO/IEC 29500).

Open Standard Format

- Defined as “formats for which the technical specifications has been made available in the public domain” (The National Archives, 2003).
- Refers to independence from outside proprietary or commercial control (Stanescu, 2005).

→ We need to review the characteristics that appear to be at the core of the open standard movement.

Criteria for Examining File Formats

- Grouping various criteria into four families:

1) Autonomy family;
2) Interoperability family;
3) Authenticity family; and
4) Functionality family.

1) Autonomy Family

- The document should...
  - be self-contained;
  - contain all information needed to access and process the content, structure, formatting and necessary metadata;
  - be independent of proprietary or commercial hardware and software configurations; and
  - be capable of preventing problems with software versions, outdated material or patent/copyright issues.
- Examples of criteria:
  - Metadata support, self-documentation
  - Openness, open availability
  - Dependencies, device independencies, external-dependency, etc.

2) Interoperability Family

- The ability of a file format to be compatible with other formats and exchange documents without loss of information (the National Archives, 2003; ECMA, 2006).
- Specifically, the ability of a given software application to open a document without requiring any special application, plug-in, codec, or proprietary add-on.
- All XML-derived specifications are compatible.
- Examples: Robustness, data interchange, etc.

3) Authenticity Family

- The ability to guarantee that a file is what it originally was without any corruption or alteration and that it faithfully represents the original content (Becker, et al., 2008; the National Archives, 2003)
- Assessing the integrity of the file through:
  - Validating the traceability of a file; or
  - Reviewing external log files.
- Examples: Integrity of layout, integrity of structure, etc.

4) Functionality Family

- The ability of a format to do exactly what it is supposed to be doing.
- This is why it is important to distinguish between two broad uses: preservation of the document structure and formatting, and preservation of useable content.
- Examples: Technical protection mechanism, adoption, component reuse, etc.

Criteria Table

Criteria	Definition/Notes	Referred by	XML Yes/No
Disclosure	Authoritative specification publicly available.	Abrams et al. (2005)	N/A
Disclosure	Existence of complete documentation.	CENDI (2007) Hodge & Anderson (2007)	Yes
Open Availability	No proprietary formats.	Barnes (2006)	N/A
Open Availability	Any manufacturer or researcher should have the ability to use the standard, rather than having it under the control of only one company.	Lesk (1995)	N/A
Openness	Standardization, Restrictions on the interpretation of the file format, Reader with freely available source.	Rog & Wijk (2007) Wijk & Rog (2007)	N/A
Open Standard	Formats for which the technical specification has been made available in the public domain.	Brown (2003)	N/A

Issues

- Although XML is not proprietary, it has many subsets with different technical specifications, which are dependent on a specific file provider.
- Which file format is most appropriate to us?
- We will look at the basic characteristics of open standard file formats rather than specific subsets of XML.