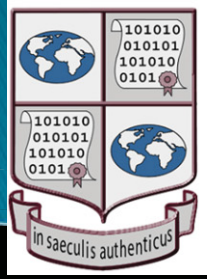

InterPARES 3 Project

International Research on Permanent Authentic Records in Electronic Systems

TEAM Korea

XML

Eun Park

McGill University

InterPARES 3 Project, 3rd International Symposium

Vancouver, BC, Canada

29 May 2010

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

Why File Format Is Important?

- "*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, Rauber, Heydegger, Schnasse and Thaller, 2008).

- Most of file formats are **proprietary and dependent** on various operating systems, hardware and software combination.
- Three main file formats: TIFF (GIF, JPEG); PDF (and PDF/A); ODF; various XML subsets.
- Many researchers have proposed their own criteria. What are the common characteristics among them?

Research Questions

- What are the characteristics of XML implementations for long term preservation?
- What is the best strategy to preserve XML documents?

Procedures

1. Identifying the characteristics of XML;
2. Identifying the criteria of XML documents to be an open standard file format;
3. Compiling the criteria to compare XML file specifications;
4. Comparing XML to other file formats;
5. Identifying the characteristics of being an open standard format for long term preservation; and
6. Developing a long term strategy for preserving XML documents for a permanent period of time.

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).
- Considering open standards from a point of the view of institutional support, relying on the user community for those standards that are widely available and used (Folk and Barkstrom, 2003).
- Refers to independence from outside proprietary or commercial control (Stanescu, 2005).

→ We need to review the characteristic that appears to be at the core of the open standard movement

Criteria of Reviewing File Formats

- To better define open standard formats, summarizing the various criteria into four major families for now:
    1) Autonomy family;
    2) Interoperability family;
    3) Authenticity family; and
    4) Functionality family.

1) Autonomy Family

- The document should be self-contained.
- Documents with all the information to access and process the content, the structure, the formatting, and necessary metadata.
- The independence of this document from proprietary or commercial hardware and software configurations
- Enable to prevent any issues with software versions, outdated material or patent and copyright issues.

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 to open a document without requiring any special application, plug-in, codec, or proprietary add-ons.
- All these XML-derived specifications are compatible.
- Practical applications of XML standards are in exchange information protocols.

3) Authenticity family

- The ability to guarantee that a file is what it originally was without any corruption or alteration and represents the content (Becker, Rauber, et al., 2008; the National Archives, 2003).
- And to uniquely identify each file.
- Specifically, data integrity which assesses the integrity of the file through:
  - An internal method to validate the authenticity of a document is to look at its traceability (i.e. tracing the original author, those who modified a file, those who opened a file, etc.); and
  - External log files.

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.
- The decision to preserve one over the other will rest with the author, the records manager or the archivist.
- And a file format will need to be chosen to better suit that need.

Issues

- XML is a proprietary specification, being dependent on a specific file provider.
- XML has many subsets with different technical specifications.

- We will look at the basic characteristics of open standard file formats rather than specific subsets of XML.
  - Four families are examined in depth;
  - Will be extended.