Preserving Attachments from an Email Collection: The Good, the Bad, the Ugly & the Thought-Provoking

Cal Lee
University of North Carolina
School of Information & Library Science
http://www.ils.unc.edu/callee/

What to preserve? Significant Properties of Digital Objects
British Library
April 7, 2008

UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

Outline

- Significant Properties & Representation Information
- SIGPROPS Project Overview
- Hunt Email Collection
- Significant Properties of Microsoft Word Documents

Significant Properties

- “Properties of digital objects that affect their quality, usability, rendering, and behaviour”*
- Often a range of possible values, rather than discrete values, e.g. appearance of email text will vary, but limits to how different content can be & still be "same" message

*Hedstrom, Margaret, and Christopher A. Lee. "Significant Properties of Digital Objects: Definitions, Applications, Implications." In *Proceedings of the DLM-Forum 2002, Barcelona, 6-8 May 2002: @ccess and Preservation of Electronic Information: Best Practices and Solutions, 218-27. Luxembourg: Office for Official Publications of the European Communities, 2002.*

Types of Properties

Type	Description	Sources of Evidence
Supported	By the file format in general	Help files, online documentation, specifications, options in application interface, source code
Observed	Properties of files in given collection - generally addressing higher-level considerations such as types of components, layout & formatting	Inspection of files directly
Measured	Properties of files in given collection that computer can identify directly, without need for human observation	Programs that detect & identify existence of properties
Intended	What creating organization or individual intended to convey	Statements from creators, legal/organizational conventions, patterns of behavior

Identifying Significant Property Implications of a Digital Preservation Strategy

- Supported properties – what’s possible in principle, e.g.
  - PDF/A doesn’t support audio, video, embedded scripts
  - ODF 1.1 doesn’t support tables in presentation slides
  - CSV doesn’t support formatting of text
- Observed/measured properties in originals – what matters for this collection/class of materials
- Observed/measured properties in digital objects resulting from given preservation strategy – what available software actually supports (e.g. footnotes supported in both formats, but “lost in translation”)

Representation Information

- “Information that maps a Data Object into more meaningful concepts” (OAIS)[1]
- Allows significant properties to be enacted (reproduced) in given technical environment or set of environments
- “No computation without representation”[2]

1. *Reference Model for an Open Archival Information System (OAIS)*. Washington, DC: Consultative Committee for Space Data Systems. 2002. [ISO 14721:2003]
2. Smith, Brian Cantwell. "Limits of Correctness in Computers." In *Computerization and Controversy: Value Conflicts and Social Choices*, edited by Rob Kling, 810-25. San Diego, CA: Academic Press, 1996. 815.

Strategic Investigation of Government Preservation of Records & Online Publications in States (SIGPROPS)

- Partnering with State Archives in North Carolina & Kansas
- Focusing on significant properties of office formats widely used in state government

The Good

Hunt Email Collection

- Electronic Correspondence File (E-Mail), 1997-2001, James B. Hunt, Jr. – Governor’s Papers
- Contents of e-mail accounts of employees of Governor’s Office during Hunt’s 4th term
- Transferred to State Archives in 2002 as 41 Outlook personal folder files (.pst)

Attachment Extraction & Conversion

- Used Outlook Attachment Sniffer (plug-in to Outlook) to extract attachments from PST files
- Generated md5 hashes for extracted files
- Created subsets with only Microsoft Office files with extensions: .xls, .xlt, .xlw, .doc, .dot, .ppt, .pps
- Ran each set through OpenOffice Document Converter

Overview of Collection

- 61973 messages, with large variance across .pst files (min=0, max=13607)
- 13746 attachments (min=0, max=1935)
- 7665 unique files (56% of file count)

15 Most Common File Formats

Extension	Count	Percentage
.doc	**9534**	**71%**
.jpg/.jpeg	545	4%
.xls	**516**	**4%**
.vcf	463	3%
.gif	446	3%
.tif	261	2%
.htm/.html	242	2%
.bmp	195	1%
.pdf	160	1%
.txt	153	1%
.dot	**131**	**1%**
No file extension	124	1%
.ppt	**117**	**1%**
.url	112	1%
.att	94	1%

The Bad (Or, at Least, The Challenging)

- 5 of 41 accounts include viruses (legacy of Anna Kournikova)
- Several files names too long to be handled by the software – my favorite:

North Carolina Promise is supporting the goals of the Unified State Plan through its expansion into new communities across the state to become Communities of Promise and through its commitment to increase the focus on the five goals of America.doc

Supported Properties – Based on File Format Specifications

- Daunting number of potential properties to consider
- Very detailed ODF 1.1. spec (738 pages)
- Microsoft release of binary Office format specs is important development, but...

The Ugly – Microsoft Office Binary Specifications

Long Trip from Word Representation Information to Significant Properties

- Fields or field ranges that are:
  - Defined as “like” some application (e.g. Word 5.x for Macintosh)
  - “Internal Information”, “Private”, “Reserved”, “Undocumented”, “Used internally by Word” or “Internal Information Used by Word”
- For fields associated with properties, rarely sufficient to reproduce an exact rendering, e.g.:
  - Long list of “built-in style names” but no indication of specifically what styles imply for properties of text
  - Dependencies on other software to deal with elements, e.g. Windows Metafile Format, OLE Compound Storage

“Character and paragraph properties in Word documents are stored in a compressed format. The information stored on disk is not the properties of a particular sequence of text but the difference of the properties from a specific reference property.”

Much of specification is about where to find representation information in a file rather than how software should handle, interpret or render it

For example:

Algorithm to determine PARAGRAPH PROPERTIES for a paragraph in a complex file

Having found the index `i` of the `FC` in an `FKP` that marks the character stored in the file immediately after the paragraph‘s paragraph mark, use the word offset stored in the first byte of the `fkp.rgbx[i-1]` to find the `PAPX` for the paragraph. Using `papx.istd` to index into the properties stored for the style sheet, the paragraph properties of the style are copied to a local `PAP`. Then the `grpprl` stored in the `PAPX` is applied to the local `PAP`, and `papx.istd` along with `fkp.rgbx.phe` are moved into the local `PAP`. The process thus far has created a `PAP` that describes what the paragraph properties of the paragraph were at the last full save. Now apply any paragraph `sprms` that were linked to the piece that contains the paragraph‘s paragraph mark. If `pcd.prm.fComplex==0`, `pcd.prm` contains 1 `sprm` which should only be applied to the local `PAP` if it is a paragraph sprm. If `pcd.prm.fComplex==1`, `pcd.prm.igrpprl` is the index of a `grpprl` in the `CLX`. If that `grpprl` contains any paragraph `sprms`, they should be applied to the local PAP. After applying all of the `sprms` for the piece, the local `PAP` contains the correct paragraph property values.

“The bottom line is that there are thousands of developer years of work that went into the current versions of Word and Excel, and if you really want to clone those applications completely, you’re going to have to do thousands of years of work. A file format is just a concise summary of all the features an application supports.”

“All of these subtle bits of behavior cannot be fully documented without writing a document that has the same amount of information as the Excel source code.”

-Joel Spolsky*

*”Why are the Microsoft Office file formats so complicated? (And some workarounds).” Joel on Software. February 19, 2008. http://www.joelonsoftware.com/items/2008/02/19.html

The Thought-Provoking

Major Issues for Microsoft Word Documents

Which of these should be preserved?

Conditional Values

- Many Word properties designed for future editing of document, e.g.
  - Line wrapping, breaking & hyphenation rules
  - Spacing of text before/after/near given elements
  - Page breaking rules
  - Page numbers in table of contents

Application State at Time of Save

- Examples:
  - Document view
  - Revision marking
  - Window state
  - Last selection

Properties for Printing

- Printer driver information
- Paper size, margins & printable area
- Portrait vs. landscape
- Headers, footers, page numbers, footnotes/endnotes
- Revision marks & comments
- Rules for odd or even numbered pages
- “Print colors as black on non-color printers”
- “Only print data inside of form fields”
- “Use printer metrics to lay out the document”

Hidden Data*

- Authors, user names & author history
- **Comments**
- Custom properties
- Database queries
- Embedded objects – extra data contained in them
- **Fast save** – change history appended to end of file, rather than being applied to body of document
- GUID – globally unique identifier for computer
- Hidden cells, slides, text – purposely hidden but then possibly forgotten
- Outlook properties & routing slips
- Path information – audio & video paths, author history, linked objects, printers, hyperlinks, include fields, template
- Presentation notes
- RSID – Revision save ID (differentiates changes from different editing sessions)
- **Tracked changes**
- **Versions**
- **Visual Basic code** – including macros & viruses

*For further elaboration, see: "The Risks of Metadata and Hidden Information: Analysis of Microsoft Office Files from the Web Sites of the Fortune 100." Oracle, 2007.

Industry Trends Suggest (At Least Some) Hidden Text Best Not Preserved

- Increasing realization of hidden data has spawned massive industries for e-discovery & tools to remove data
- Several changes by Microsoft to Office:
  - Embedded PID GUID abandoned in Office 2000
  - Fast save turned off by default in Word 2000 & disabled in Word 2003
  - Document Investigator (Office 2007)
  - Appearance of comments & tracked changes by default when opening document (Office 2007)

Tackling Observed & Measured Properties

In General:

- “Acid test” documents with range of possible properties, then apply preservation strategy & see which properties retained/lost
- Examples:
  - Test file that “that exercised all the field tags and field values” in Lotus 1-2-3[1]
  - Acid1, Acid2, Acid3 for Web[2]
  - OpenXML Reference Document[3]

For Given Collection:

- For given set of files, need tools/techniques to detect which of known set of possible properties are actually present, e.g.
  - Do any of these Word documents contain macros?
  - Do any of these PowerPoint presentations have tables?
  - Do any of these Excel spreadsheets contain hidden cells?
- Determine how much they matter

[1] Lawrence, Gregory W., William R. Kehoe, Oya Y. Rieger, William H. Walters, and Anne R. Kenney. "Risk Management of Digital Information: A File Format Investigation.“ Washington, DC: Coalition on Library and Information Resources, 2000.

[2] http://www.w3.org/Style/CSS/Test/CSS1/current/test5526c.htm; http://www.webstandards.org/files/acid2/test.html; http://acid3.acidtests.org/

[3] Andrew Ziem, http://katana.oooninja.com/w/reference_sample_documents

Measured Properties – Challenges & Considerations

- Not use cases normally assumed by software developers
- Some detectable properties as proxies for others, e.g.
  - Page count & bitmap rendering of pages as likely indicators of differences in line & paragraph breaks, which likely to be result of font rendering, hyphenation & spacing*
  - Word count likely indicates "differences in hyphenation or different layout of the titles, table of contents or index"*
- Difficult to determine if differences "are caused by the transformation, or are merely an artifact of the tools used to examine the objects"*

*Clausen, Lars R. "Opening Schrödingers Library: Semi-Automatic Qa Reduces Uncertainty in Object Transformation." In *Research and Advanced Technology for Digital Libraries: 11th European Conference, ECDL 2007, Budapest, Hungary, September 16-21, 2007; Proceedings*, edited by László Kovács, Norbert Fuhr and Carlo Meghini, 186-97. Berlin: Springer, 2007.

There are two kinds of people in the world, my friend: Those with a rope around the neck, and the people who have the job of doing the cutting.

- Tuco, *The Good, the Bad and The Ugly*, 1966