The Development of Digital Preservation Best Practices in EPrints

Slavko Manojlovich
Associate University Librarian (QEII Library and IT)
Memorial University of Newfoundland
with contributions by Casey Hilliard

OR2012 : The 7th International Conference on Open Repositories

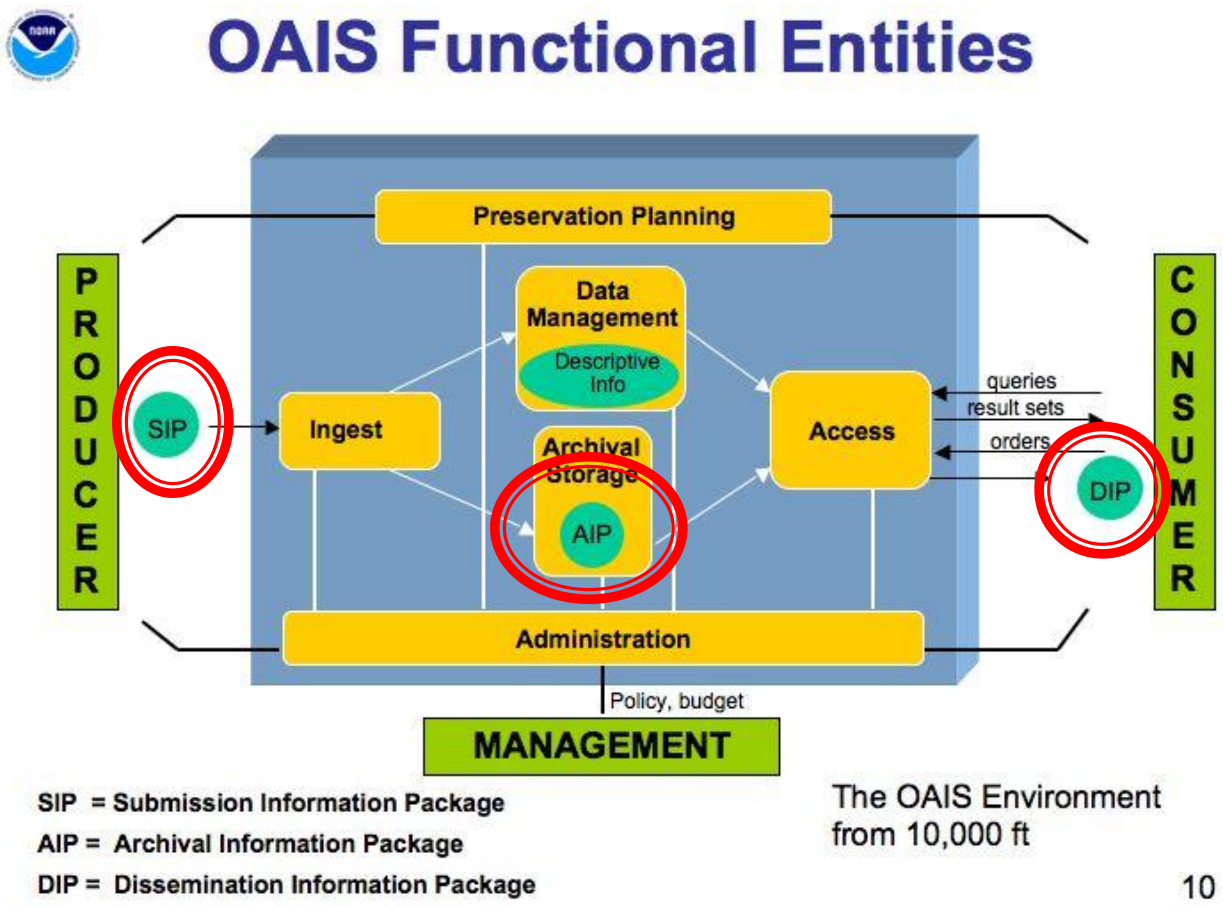
Where is Newfoundland?

Birthplace of the Internet?

Marconi and his receiving apparatus at Signal Hill, St. John's, December 1901.

Open Archival Information System

OAIS Reference Model - Objects

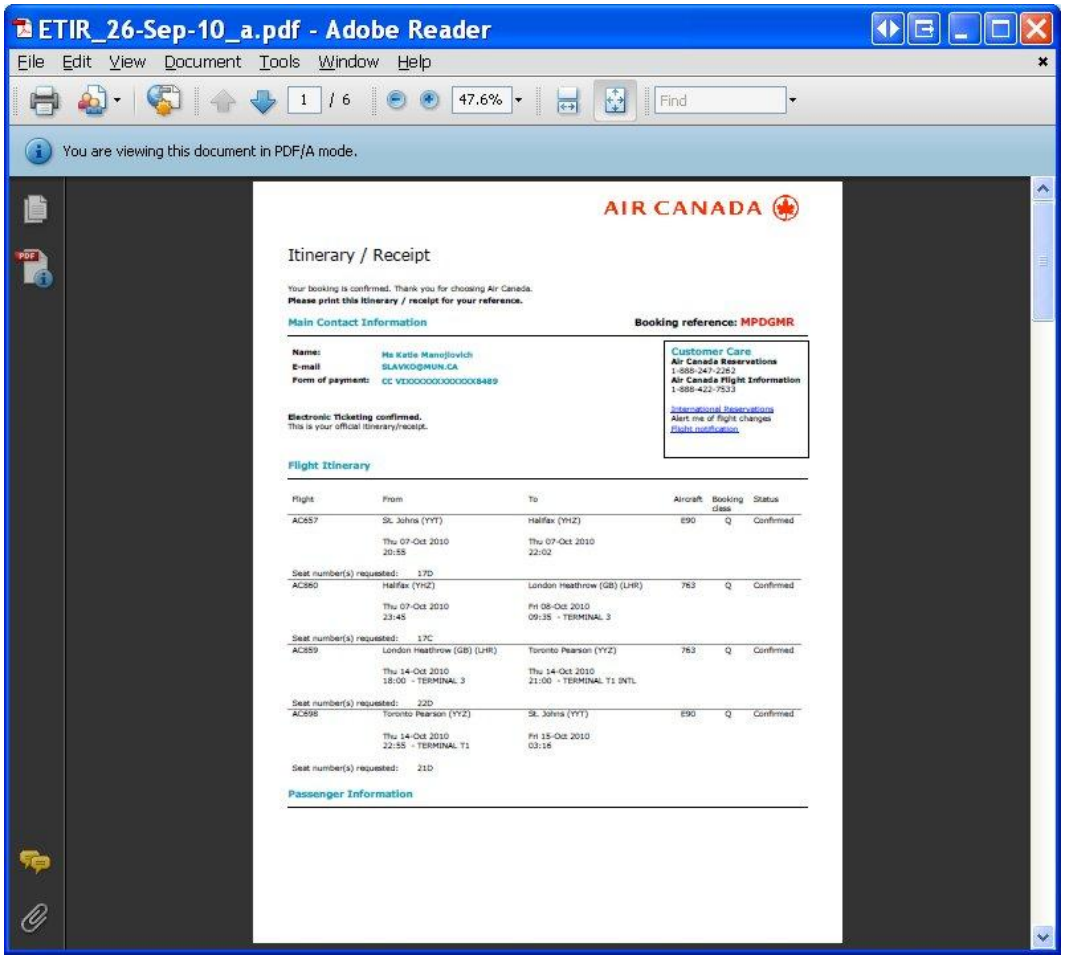


Digital Preservation Goals

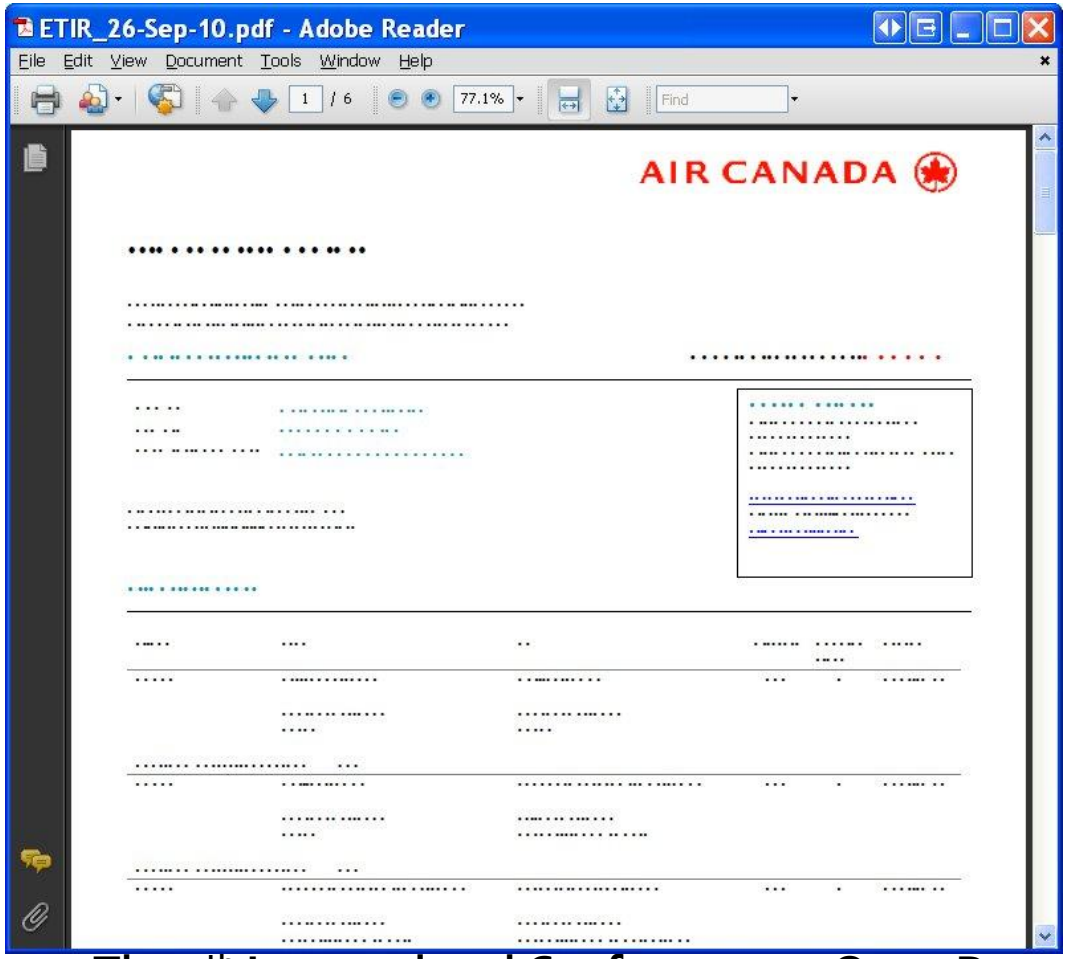
- Long-term meaningful access to file formats across all browsers, operating systems and devices.
- Display/play original look and feel (colour, layout, etc.)
- Reusability of born digital objects

Caveat: to the extent possible

Preserve Look and Feel



Preserve Look and Feel (missing fonts)



Preserve Look and Feel

You are viewing this document in PDF/A mode.

CLOTHES MOTH FACTS:

The BioCare Clothes Moth Trap attracts and captures adult male webbing clothes moths. Clothes Moths occasionally occur in homes, where the larval stage damages clothing and other personal objects that are made of wool, hair, fur, feathers or other natural fibers.

THIS TRAP SIGNALS WHEN A DAMAGING INFESTATION IS PRESENT: IT DOES NOT CONTROL THE LARVAE WHICH FEED ON CLOTHING. IF MOTHS ARE FOUND IN THE TRAP, TAKE IMMEDIATE MEASURES TO CONTROL THE LARVAE CAUSING THE DAMAGE.

Preserve Look and Feel (font substitution for missing fonts)

CLOTHES MOTH FACTS:

The BioCare Clothes Moth Trap attracts and captures adult male webbing clothes moths. Clothes Moths occasionally occur in homes, where the larval stage damages clothing and other personal objects that are made of wool, hair, fur, feathers or other natural fibers.

THIS TRAP SIGNALS WHEN A DAMAGING INFESTATION IS PRESENT: IT DOES NOT CONTROL THE LARVAE WHICH FEED ON CLOTHING. IF MOTHS ARE FOUND IN THE TRAP, TAKE IMMEDIATE MEASURES TO CONTROL THE LARVAE CAUSING THE DAMAGE.

Female moths are especially attracted to oily or soiled fabrics, where they

Preservation Planning

- Monitor designated community (consumer needs and expectations).
- Monitor technology.
- Develop preservation strategies and standards.
- Develop packaging designs and migration plans.

Monitor Designated Community

- Google Analytics (Browser)

	Browser	Visits	% Visits
1.	Internet Explorer	7,031	49.83%
2.	Firefox	2,832	20.07%
3.	Chrome	2,200	15.59%
4.	Safari	1,621	11.49%
5.	Mozilla Compatible Agent	123	0.87%
6.	Android Browser	109	0.77%
7.	IE with Chrome Frame	80	0.57%
8.	Opera	56	0.40%
9.	Opera Mini	29	0.21%
10.	Mozilla	16	0.11%

Monitor Designated Community

- Google Analytics (Operating System)

	Operating System	Visits	% Visits
1.	Windows	11,558	81.92%
2.	Macintosh	1,538	10.90%
3.	iPad	409	2.90%
4.	iPhone	238	1.69%
5.	Android	111	0.79%
6.	Linux	92	0.65%
7.	(not set)	76	0.54%
8.	iPod	51	0.36%
9.	BlackBerry	21	0.15%
10.	AIX	4	0.03%

Monitor Designated Community

- Google Analytics (Mobile Operating System)

	Operating System	Visits	% Visits
1.	iPad	409	48.75%
2.	iPhone	238	28.37%
3.	Android	111	13.23%
4.	iPod	51	6.08%
5.	BlackBerry	21	2.50%
6.	SymbianOS	3	0.36%
7.	Windows Phone	3	0.36%
8.	Samsung	2	0.24%
9.	Nokia	1	0.12%

Monitor Designated Community

- Google Analytics (Mobile Screen Resolution)

	Screen Resolution	Visits	% Visits
1.	768x1024	409	48.75%
2.	320x480	288	34.33%
3.	1280x800	13	1.55%
4.	480x800	10	1.19%
5.	320x240	9	1.07%
6.	480x360	7	0.83%
7.	0x0	5	0.60%
8.	720x1280	4	0.48%
9.	360x640	3	0.36%
10.	480x640	3	0.36%

Monitor Technology

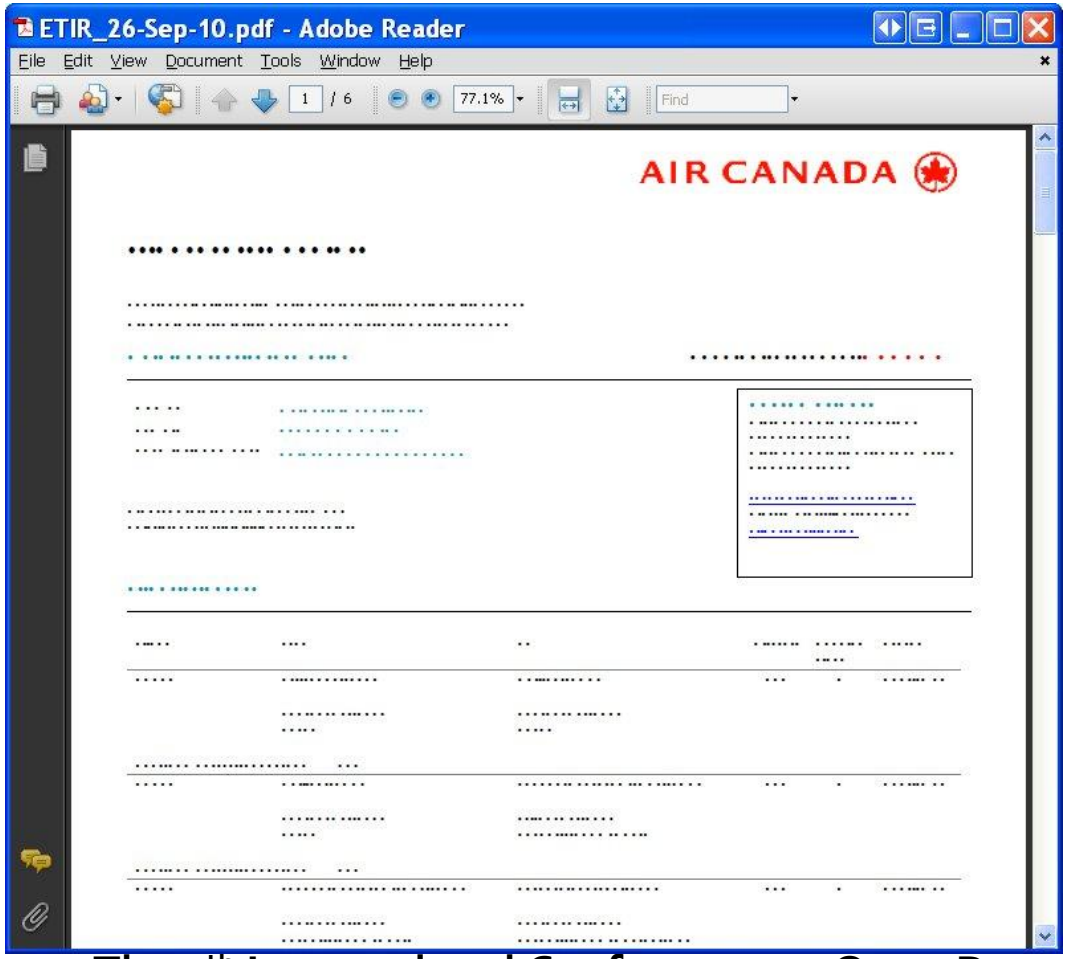
- Adobe Multimedia Flash Format

``At the town hall meeting where he [Steve Jobs] attacked Google, he also assailed Adobe`s multimedia platform for websites, Flash, as a ``buggy`` battery hog made by ``lazy people``. The iPod and iPhone, he said, would never run Flash.`` Steve Jobs by Walter Isaacson, 2011, p. 514.

News: Adobe drops Flash development for mobile devices. H.264 (MPEG-4) becomes de facto standard.

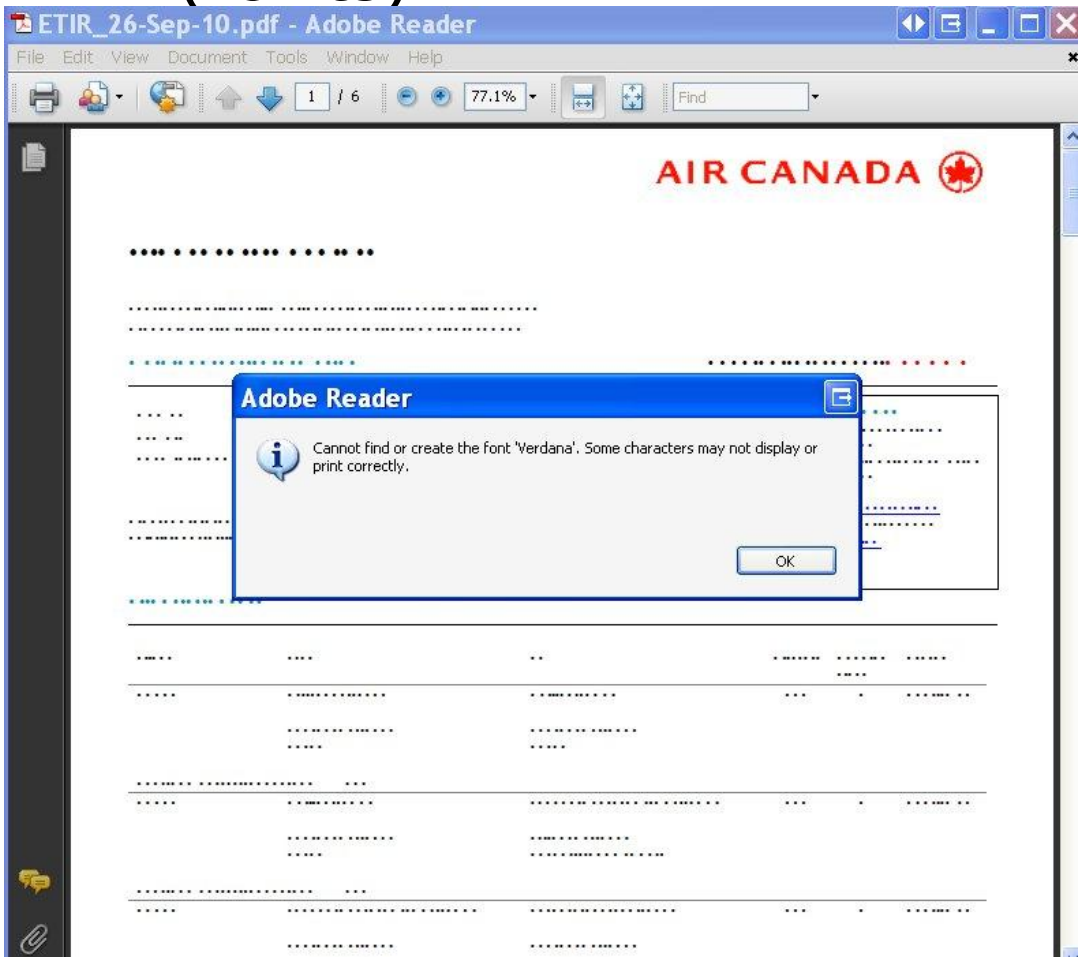
Monitor Technology

- Adobe PDF (fonts)



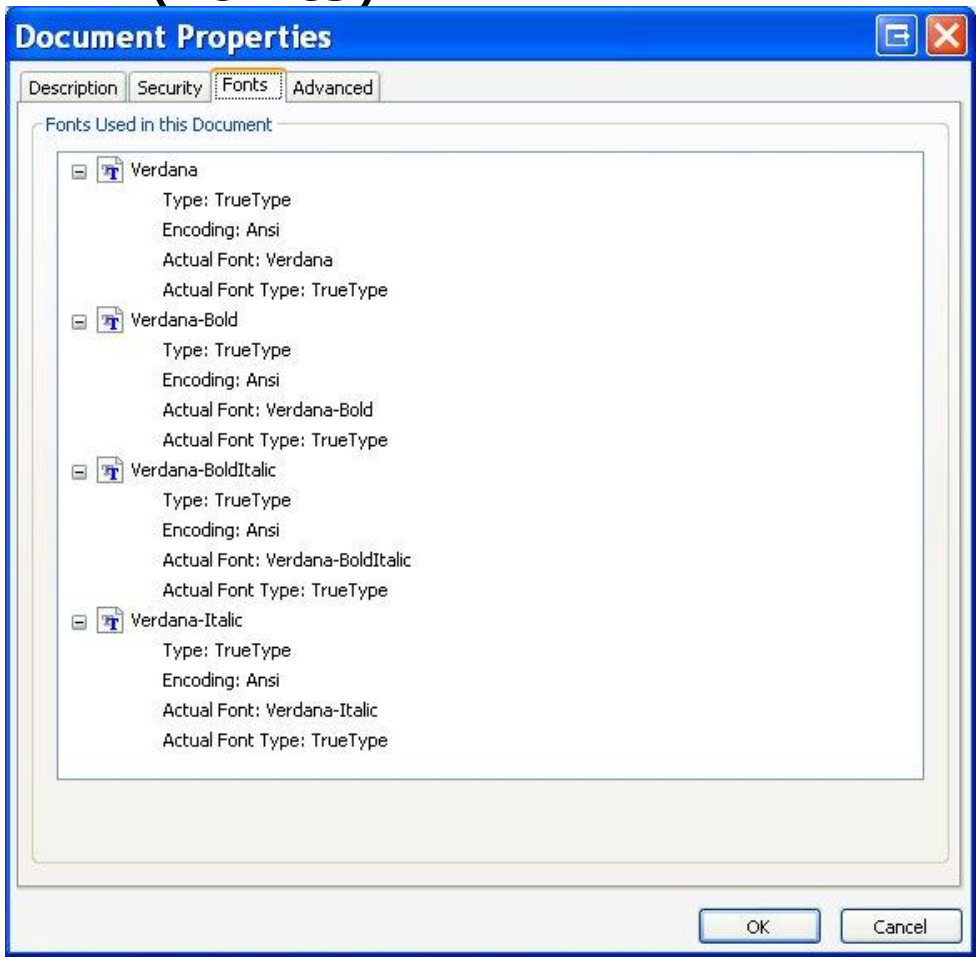
Monitor Technology

- Adobe PDF (fonts)



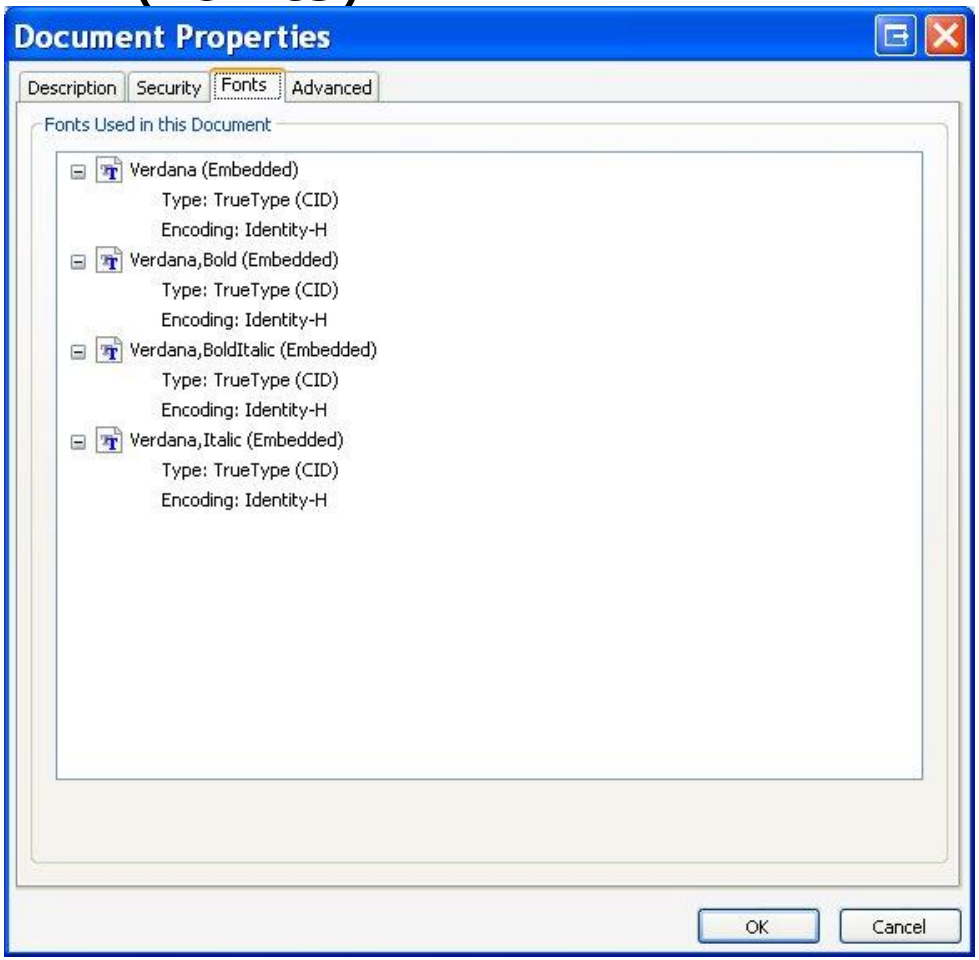
Monitor Technology

- Adobe PDF (fonts)



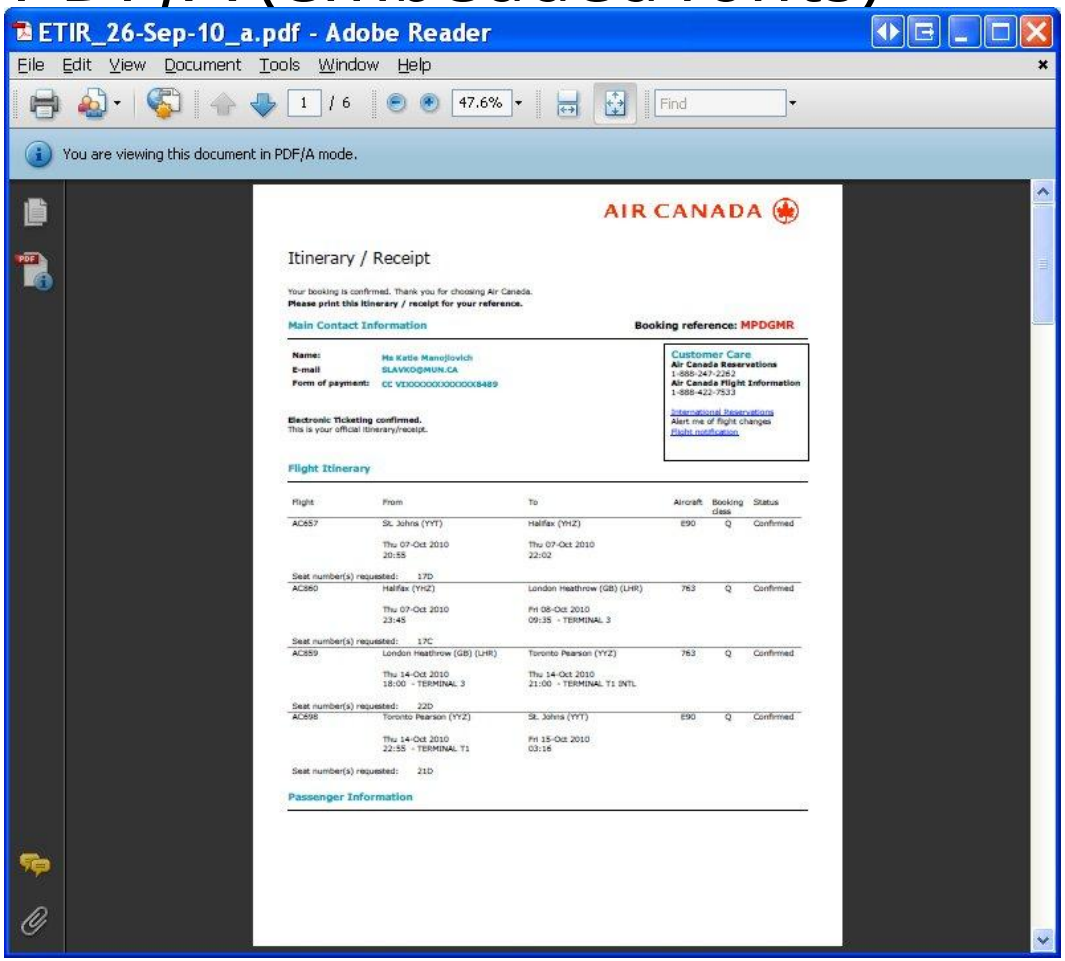
Monitor Technology

- Adobe PDF (fonts)



Monitor Technology

- Adobe PDF/A (embedded fonts)

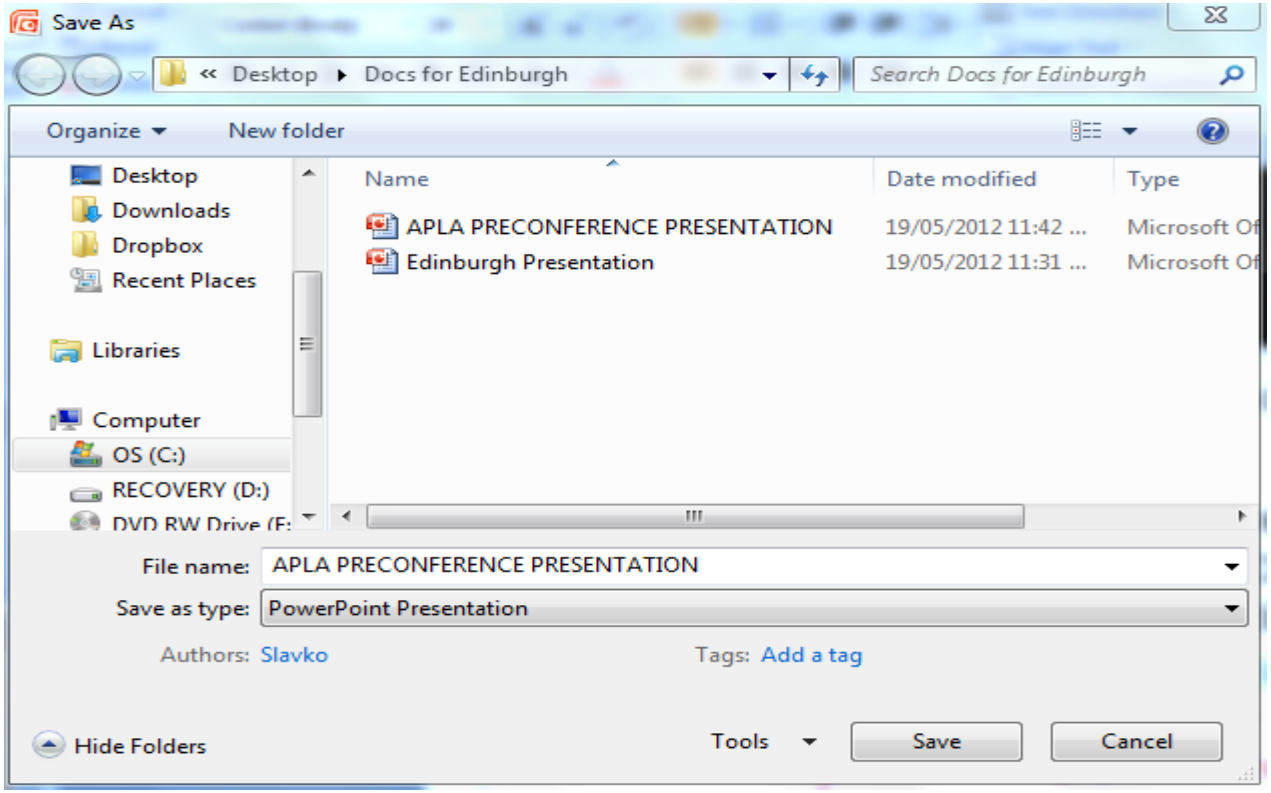


Monitor Technology

- Adobe PDF File Format Characteristics
  - Scanned image or born digital PDF
  - Searchable (OCR required for scanned image PDF)
  - PDF/A long-term preservation standard for page-oriented documents.
    Library of Congress
  - PDF/A constraints (no embedded audio, video or javascript).
  - PDF/A embeds the fonts and color spaces.

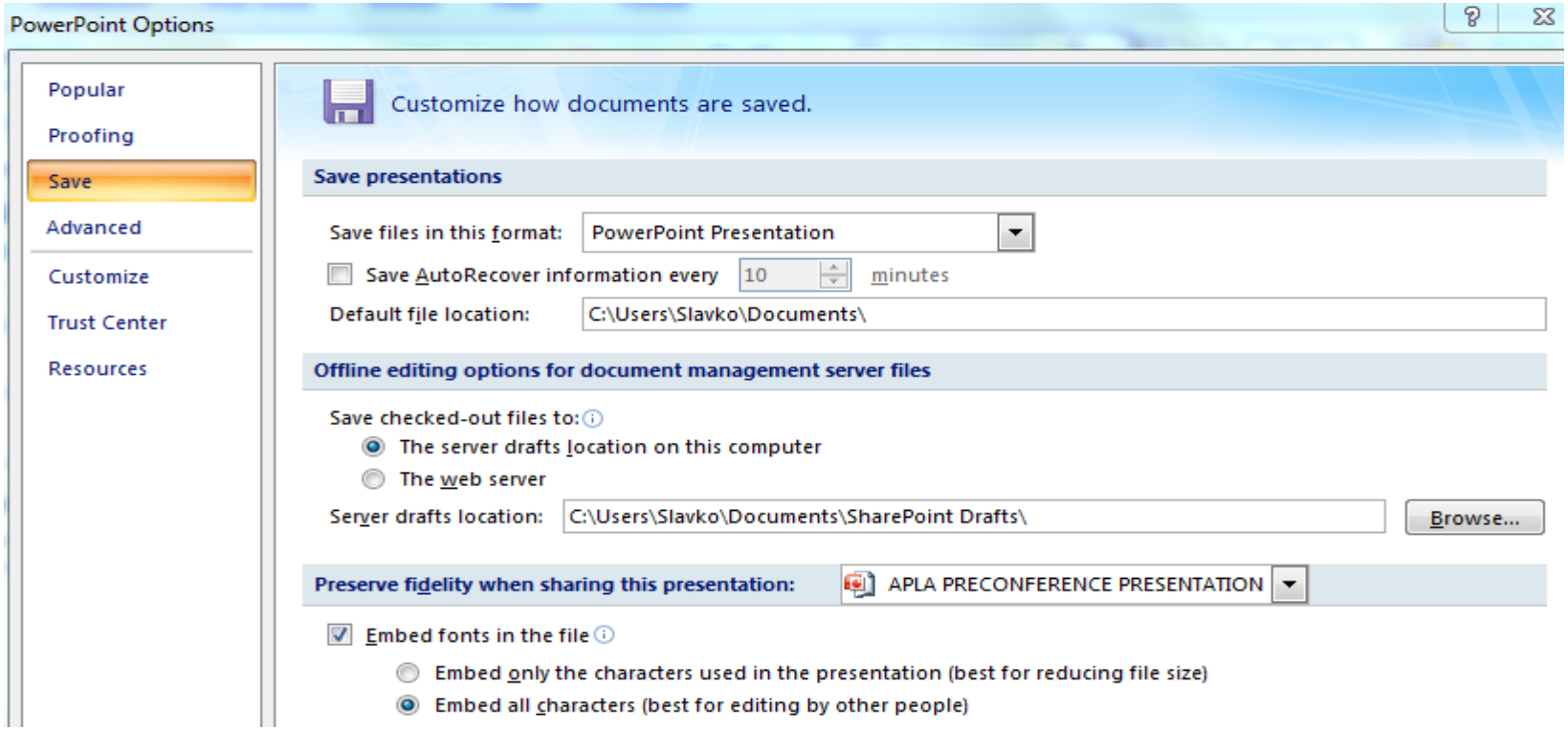
Monitor Technology

- PowerPoint File Format Characteristics
  - Microsoft Office > SAVE AS > Tools



Monitor Technology

- PowerPoint File Format Characteristics
  - Microsoft Office can embed all characters in a font



Monitor Technology

- Adobe PDF File Format Characteristics
  - Make sure all fonts are embedded

    "By default, when creating a Screen Optimized PDF with Distiller or any PDF from PDF Writer, the Base 14 Fonts are not embedded in the document. Since these fonts are available in Acrobat Reader it is assumed that they will be available to any viewer and embedding would simply add unnecessarily to the file size. However, Distiller (4.x) settings and PDF Writer for Windows settings can be changed in order to embed the Base 14 fonts."

Monitor Technology

- TIFF Versus JPEG 2000 Image Format
  "Certainly, one of JPEG2000's major attractions is its reduced storage requirement for losslessly compressed images. Tests have consistently shown that storage savings of around 50% (2:1) as compared to uncompressed TIFF files can be anticipated without consequent loss of image quality."

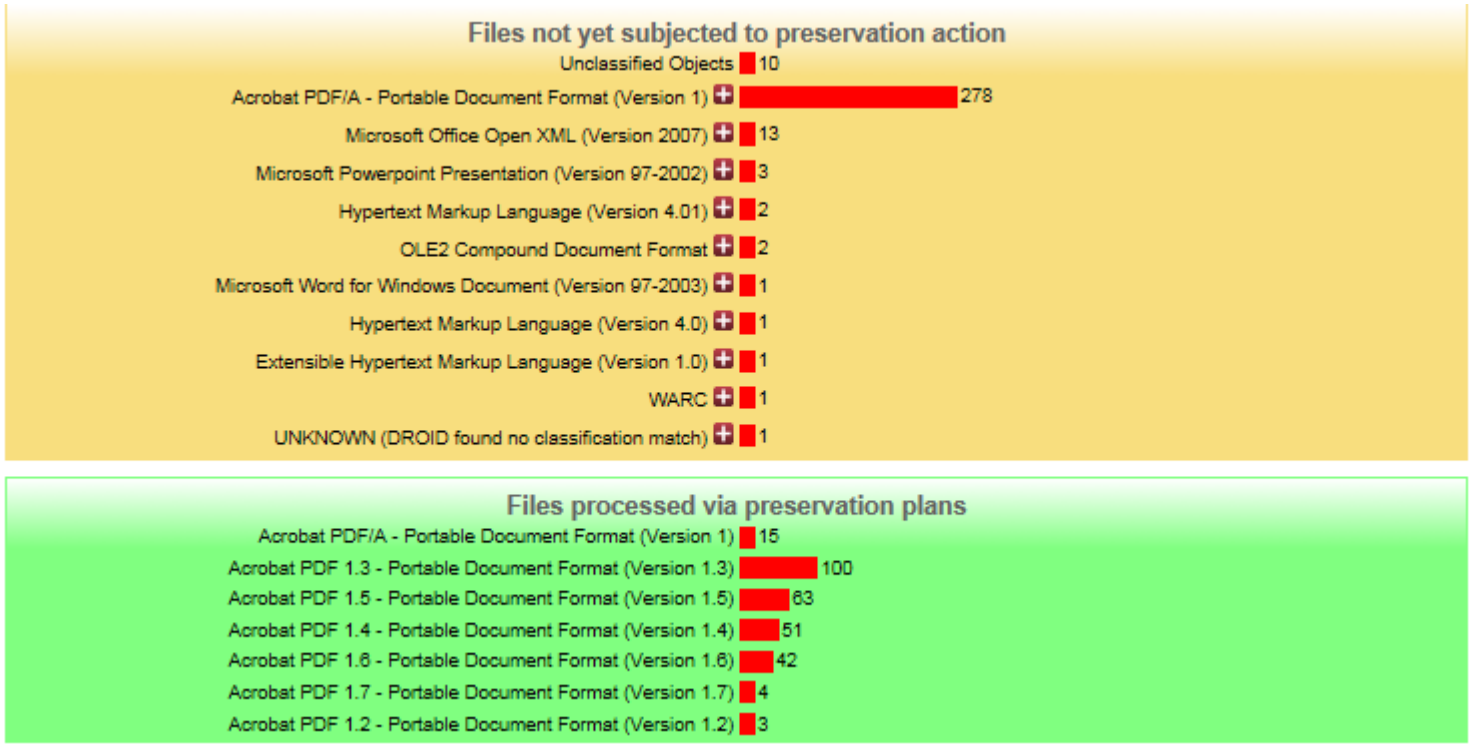
  JPEG2000 Preservation File Format Working Group
  Library and Archives Canada
  http://www.collectionscanada.gc.ca/digital-initiatives/012018-2100.01-e.html#anc5

Develop Preservation Strategies

- Create an inventory of files and file formats.
- File extension is not a reliable indicator of the file format:
  - .pdf = pdf or pdf/a?
  - Malicious files can masquerade behind a file extension.
  - Use a file format identification tool like DROID, JHOVE, JHOVE2 or FITS to identify files. All tools reference the PRONOM or UDFR File Format registries of file formats.

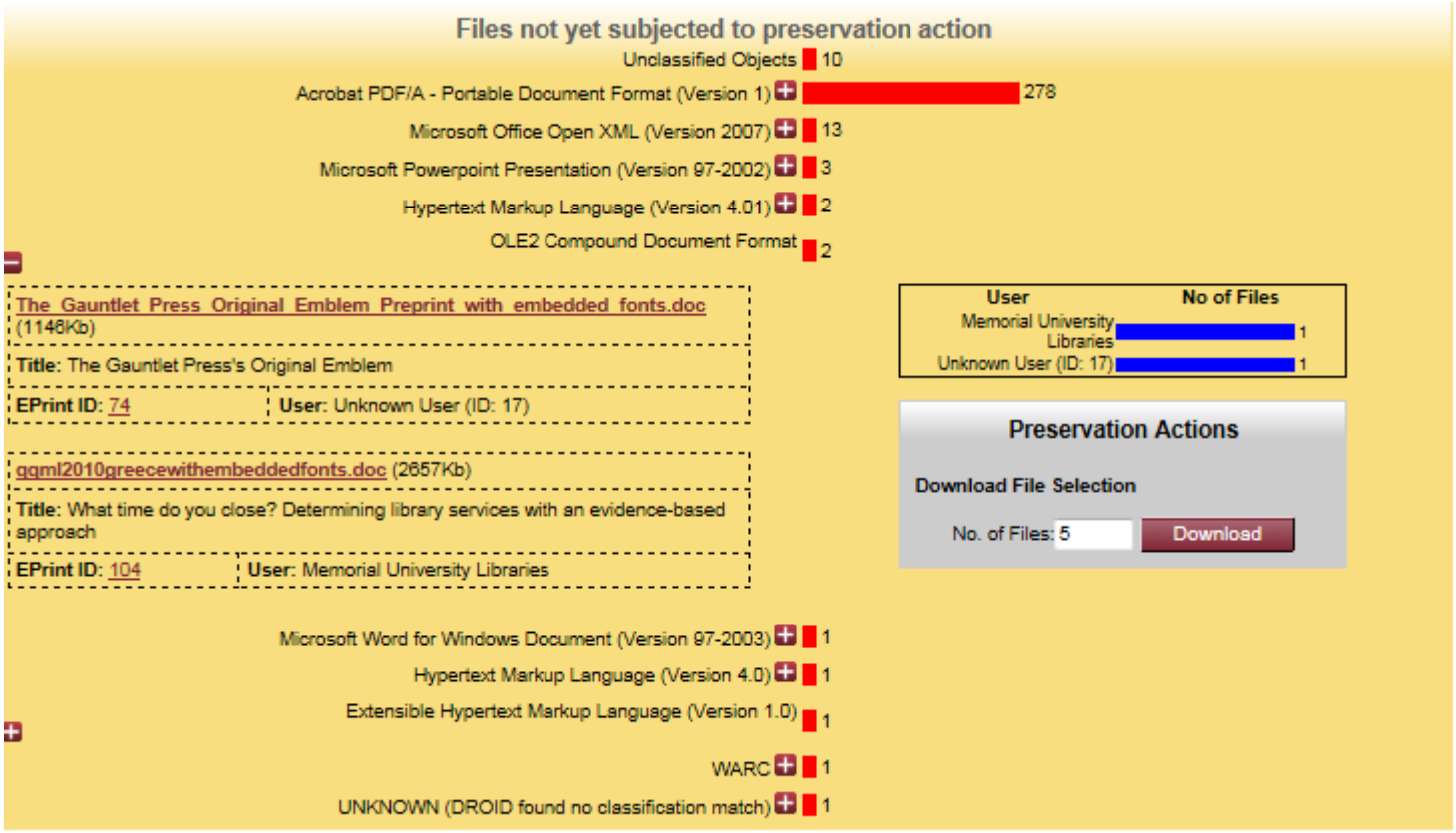
Develop Preservation Strategies

- DROID integration in an EPrints repository.



Develop Preservation Strategies

- DROID integration in an EPrints repository.



Develop Preservation Strategies

- Create a media type preservation plan.

Media type	Supported Ingest File Format Extensions	Long-Term Preservation Format(s)	Access Format(s)	Normalization tool
Audio	MP3, WMA, WAV	WAV (LPCM)	MP3, WMA	Switch Sound File Converter version 4
HTML documents	HTML	MHT, WEBARCHIVE	MHT, WEBARCHIVE	Browser SAVE AS
Portable Document Format without embedded video, audio or javascript	PDF	PDF/A searchable	PDF/A searchable	PDF/A Manager (PDFTRON), Solid PDF Tools, Adobe Pro

Develop Preservation Strategies

- Create a media type preservation plan.

Media type	Supported Ingest File Format Extensions	Long-Term Preservation Format(s)	Access Format(s)	Normalization tool
Audio	MP3, WMA, WAV	WAV (LPCM)	MP3, WMA	Switch Sound File Converter version 4
HTML documents	HTML	MHT, WEBARCHIVE	MHT, WEBARCHIVE	Browser SAVE AS
Portable Document Format without embedded video, audio or javascript	PDF	PDF/A searchable	PDF/A searchable	PDF/A Manager (PDFTRON), Solid PDF Tools, Adobe Pro

Develop Preservation Strategies

- Create a media type preservation plan.

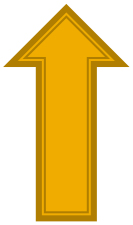
Media type	Supported Ingest File Format Extensions	Long-Term Preservation Format(s)	Access Format(s)	Normalization tool
Audio	MP3, WMA, WAV	WAV (LPCM)	MP3, WMA	Switch Sound File Converter version 4
HTML documents	HTML	MHT, WEBARCHIVE	MHT, WEBARCHIVE	Browser SAVE AS
Portable Document Format without embedded video, audio or javascript	PDF	PDF/A searchable	PDF/A searchable	PDF/A Manager (PDFTRON), Solid PDF Tools, Adobe Pro

Develop Preservation Strategies

- Create a media type preservation plan.

Media type	Supported Ingest File Format Extensions	Long-Term Preservation Format(s)	Access Format(s)	Normalization tool
Audio	MP3, WMA, WAV	WAV (LPCM)	MP3, WMA	Switch Sound File Converter version 4
HTML documents	HTML	MHT, WEBARCHIVE	MHT, WEBARCHIVE	Browser SAVE AS
Portable Document Format without embedded video, audio or javascript	PDF	PDF/A searchable	PDF/A searchable	PDF/A Manager (PDFTRON), Solid PDF Tools, Adobe Pro

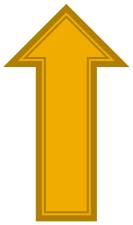
Includes original files



Develop Preservation Strategies

- Create a media type preservation plan.

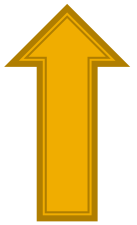
Media type	Supported Ingest File Format Extensions	Long-Term Preservation Format(s)	Access Format(s)	Normalization tool
Audio	MP3, WMA, WAV	WAV (LPCM)	MP3, WMA	Switch Sound File Converter version 4
HTML documents	HTML	MHT, WEBARCHIVE	MHT, WEBARCHIVE	Browser SAVE AS
Portable Document Format without embedded video, audio or javascript	PDF	PDF/A searchable	PDF/A searchable	PDF/A Manager (PDFTRON), Solid PDF Tools, Adobe Pro



Develop Preservation Strategies

- Create a media type preservation plan.

Media type	Supported Ingest File Format Extensions	Long-Term Preservation Format(s)	Access Format(s)	Normalization tool
Audio	MP3, WMA, WAV	WAV (LPCM)	MP3, WMA	Switch Sound File Converter version 4
HTML documents	HTML	MHT, WEBARCHIVE	MHT, WEBARCHIVE	Browser SAVE AS
Portable Document Format without embedded video, audio or javascript	PDF	PDF/A searchable	PDF/A searchable	PDF/A Manager (PDFTRON), Solid PDF Tools, Adobe Pro



Develop Preservation Strategies

1. Create a media type preservation plan.

Media type	Supported Ingest File Format Extensions	Long-Term Preservation Format(s)	Access Format(s)	Normalization tool
Audio	MP3, WMA, WAV	WAV (LPCM)	MP3, WMA	Switch Sound File Converter version 4
HTML documents	HTML	MHT, WEBARCHIVE	MHT, WEBARCHIVE	Browser SAVE AS
Portable Document Format without embedded video, audio or javascript	PDF	PDF/A searchable	PDF/A searchable	PDF/A Manager (PDFTRON), Solid PDF Tools, Adobe Pro

Memorial University Draft Media Type Preservation Plan

Develop Preservation Strategies

- Media Type: html document
  - Supported ingest file format extension: html
  - Long-term preservation format(s): mht, webarchive
  - Access formats:
    - mht: web archive format supported by most Windows browsers -- Firefox requires UnMHT plugin
    - webarchive: Safari browsers (Windows and Mac) and iPAD Safari with the Downloads Lite for iPAD app
  - Normalization tool: browser SAVE AS
  - Link to EPrints record

Develop Preservation Strategies

- Media Type: web site
  - Supported ingest file format extension: html link to web site home page
  - Long-term preservation format(s): warc
  - Access formats:
    - html link to archived web site
    - warc
  - Normalization tool: heretrix, wayback
  - Link to EPrints record

Develop Preservation Strategies

- Media Type: Portable Document Format (PDF) without embedded audio, video or javascript
  - Supported ingest file format extension: pdf
  - Long-term preservation format(s): PDF/A + original file
  - Access formats:
    - PDF/A
  - Normalization tool: PDF/A Manager (PDFTRON), Solid PDF Tools, Adobe Pro
  - Link to EPrints record

Develop Migration Plans

- EPrints: enhanced digital preservation module batch processes self/mediated deposit of born digital content. PDF module migrates ingested PDFs to searchable PDF/A with embedded fonts and records the transformations in the preservation metadata.

EPrints Preservation Planning Services at Memorial University

- Clean Filename Service
  - Part of: upload
  - Makes use of: Perl regular expressions
  - Logic: Replace offending characters with "-" and "—" with "-"
  - History log: No, work in progress

EPrints Preservation Planning Services at Memorial University

- Virus Check Service
  - Part of: upload
  - Makes use of: ClamAV virus checker
  - Logic: terminate upload if virus found
  - History log: No (work in progress)

EPrints Preservation Planning Services at Memorial University

- Searchable PDF Service
  - Part of: independent batch job
  - Makes use of: pdffonts from xpdf
  - Logic: look for fonts in PDF and output a list of files which require OCR
  - History log: No (work in progress)

EPrints Preservation Planning Services at Memorial University

- File Format Identification Service
  - Part of: EPrints preservation module
  - Makes use of: Formats/Risks module and DROID repository
  - Logic: identify file formats using file signatures to facilitate further preservation actions
  - History log: Yes

EPrints Preservation Planning Services at Memorial University

- PDF to PDF/A Migration Service
  - Part of: independent cron job
  - Makes use of: modified Formats/Risks main processor and PDFTron software
  - Logic: find all PDFs which have not been migrated; check PDF/A compliance; migrate PDFs to PDF/As; add migrated file to EPrint record inheriting metadata of original file
  - History log: Yes

EPrints Preservation Planning Services at Memorial University

- Digital Integrity Service
  - Part of: independent cron job
  - Makes use of: Audit Control Environment (ACE) under construction
  - Logic: periodically calculates and compares checksums for all files in addition to checking the integrity of the checksum database
  - History log: No

EPrints Preservation Planning Services at Memorial University

- Next Steps
  - Continue developing migration plans and automated processes for additional file types
  - Work toward the development of an integrated preservation planning solution for the 2013 Open Repositories Conference to be held in Charlottetown, Prince Edward Island, Canada

Contact: slavko@mun.ca

Memorial University Research Repository