Digital Preservation Best Practices: Lessons Learned From The Experts

Preconference Workshop (Monday July 8, 1:00 – 5:00 pm)

Casey Hilliard (Memorial University of Newfoundland)
Anthony Leroy (Université Libre de Bruxelles)
Slavko Manojlovich (Memorial University of Newfoundland)
Courtney Mumma (Artefactual Systems, Inc.)
Benoit Pauwels (Université Libre de Bruxelles)
David Tarrant (Southampton University)

Open Repositories 2013

Repository Island, Charlottetown, PEI, Canada

Schedule

- Introduction to Digital Preservation
- Media Type Preservation Planning
- Current State of Preservation Tools
- E-Theses Preservation
- PDF to PDF/A Migration Workflow

  <BREAK>
- Archivematica
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- Q and A

The Speakers

Speakers' Repository Platforms

- Archivematica
- CONTENTdm
- DSpace
- EPrints

Schedule

- **Introduction to Digital Preservation**
- Media Type Preservation Planning
- Current State of Preservation Tools
- E-Theses Preservation
- PDF to PDF/A Migration Workflow

  <BREAK>
- Archivematica
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- Q and A

Digital Preservation Best Practices

Courtney C. Mumma, MAS/MLIS, Systems Analyst and Archivematica Product Manager

Who am I?

Home Services Team Clients Contact

artefactual systems inc.

Courtney C. Mumma, MAS/MLIS

Systems Analyst & Archivematica Product Manager

Courtney is responsible for managing Archivematica system requirements, product design, technical support, training, and community relations.

Courtney is a graduate of the University of British Columbia's Master of Archival Studies and Master of Library and Information Studies programs (2009). Prior to joining Artefactual, Courtney worked at the City of Vancouver Archives implementing their digital archives system while managing the acquisition of the hybrid digital-analog 2010 Winter Olympic Games archives.

She has been a researcher and co-investigator on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES 3 Project), researcher on the UBC-SLAIS Digital Records Forensics Project, and a member of the Professional Experts Panel on the BitCurator Project. Courtney has been published in Archivaria and has delivered many presentations on the practical application of digital preservation strategies.

artefactual systems inc.

digital preservation consulting
open-source sofware for archives and libraries

archivematica®

ICAAtoM

Peter Van Garderen
President

Evelyn McLellan
Director, Consulting Services

David Juhasz
Director, Technical Services

Courtney C. Mumma
Archivematica Product Manager

Jessica Bushey
AtoM Product Manager

Jesús García Crespo
Software Developer

Joseph Perry
Software Developer

Austin Trask
Systems Technician

Mike Cantelon
Software Developer

Dan Gillean
Systems Analyst

Justin Simpson
Software Developer

Mike Gale
Software Developer

The Digital Preservation Problem:

1. Rapid technological change drives constant system upgrades, migrations and retirement of legacy technologies.

2. Incompatible, obsolete, obscure or proprietary systems and file formats.

3. Loss or damage to bitstreams due the fragility of digital storage media, system error, or human error.

The Digital Preservation Problem:

4. The overwhelming volume of digital information objects created daily, each with many possible copies and versions.

5. The lack or loss of adequate metadata describing digital information objects.

6. Accidental or malicious content alteration.

The Digital Preservation Problem:

7. Doubts about the reliability and integrity of electronic records and the inability to vouch for their authenticity.

8. The complexity of digital information objects which requires preservation of their content, structure, context, presentation, behaviour as intellectual entities as well as bitstreams.

9. The lack of formally recognized organizational responsibility, resources and enterprise architecture components that facilitate digital curation, preservation and long-term access.

content
context
structure
presentation
behaviour

digital information
1010 1000 1011 1101
1010 1000 1011 1101
1010 1000 1011 1101
1010 1000 1011 1101
1010 1000 1011
1010 1000 1011

now

contextualize
authenticate
relate / bind
find

file system
file format
codec
character encoding
fonts
packaging
decryption
error correction
operating system
compression
metadata

storage media
storage driver
input / output devices
bitstream
storage device
application software
user interface

stored
copied
protected

digital information
1010 1000 1011 1101
1010 1000 1011 1101
1010 1000 1011 1101
1010 1000 1011 1101
1010 1000 1011
1010 1000 1011

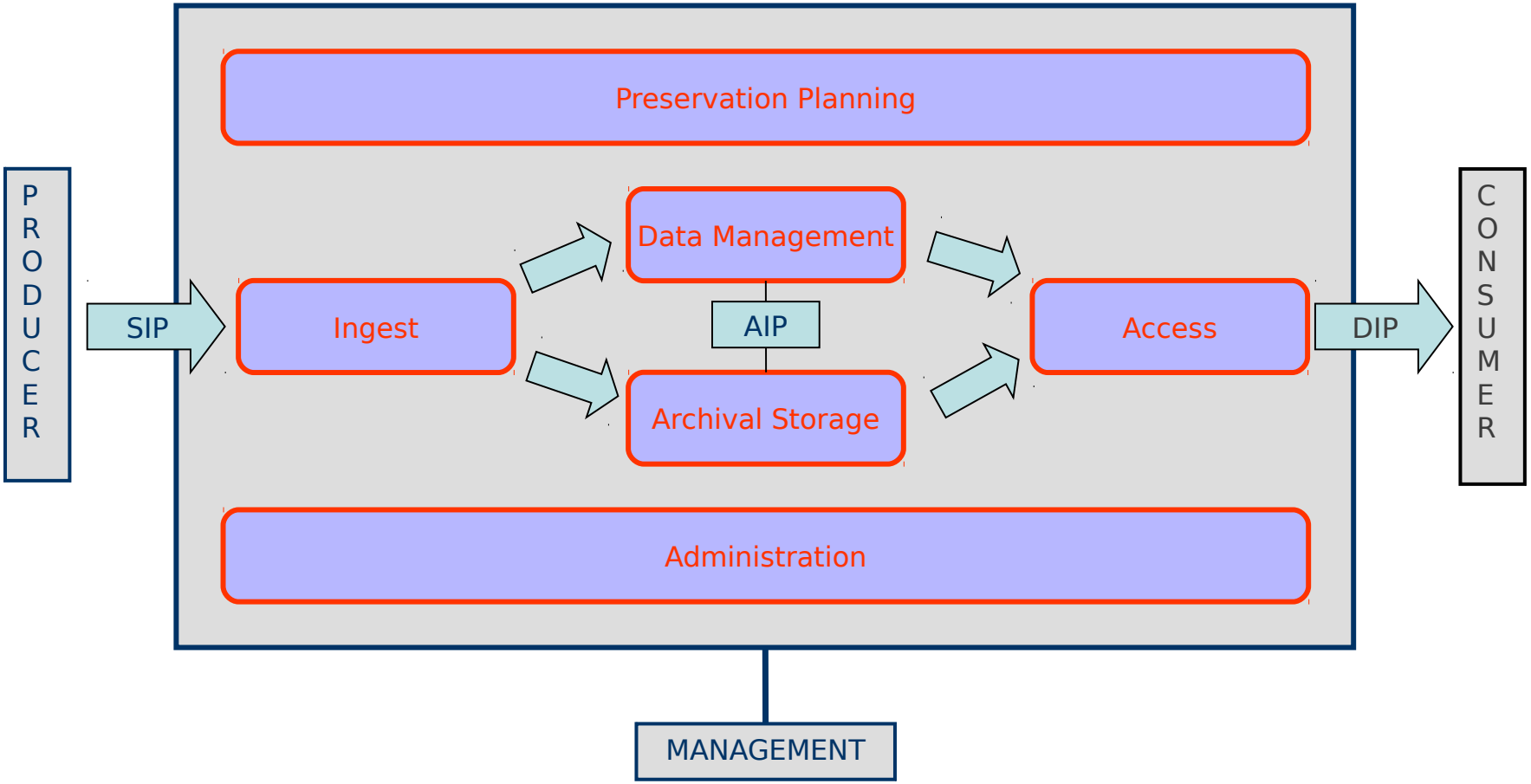
future

Accessible?
Usable?
Authentic?

Responsible?
Architecture?
Resources?


Bill Hanna
Joe Barbera
228/300

Open Archival Information System (OAIS) reference model (ISO-STD 14721)

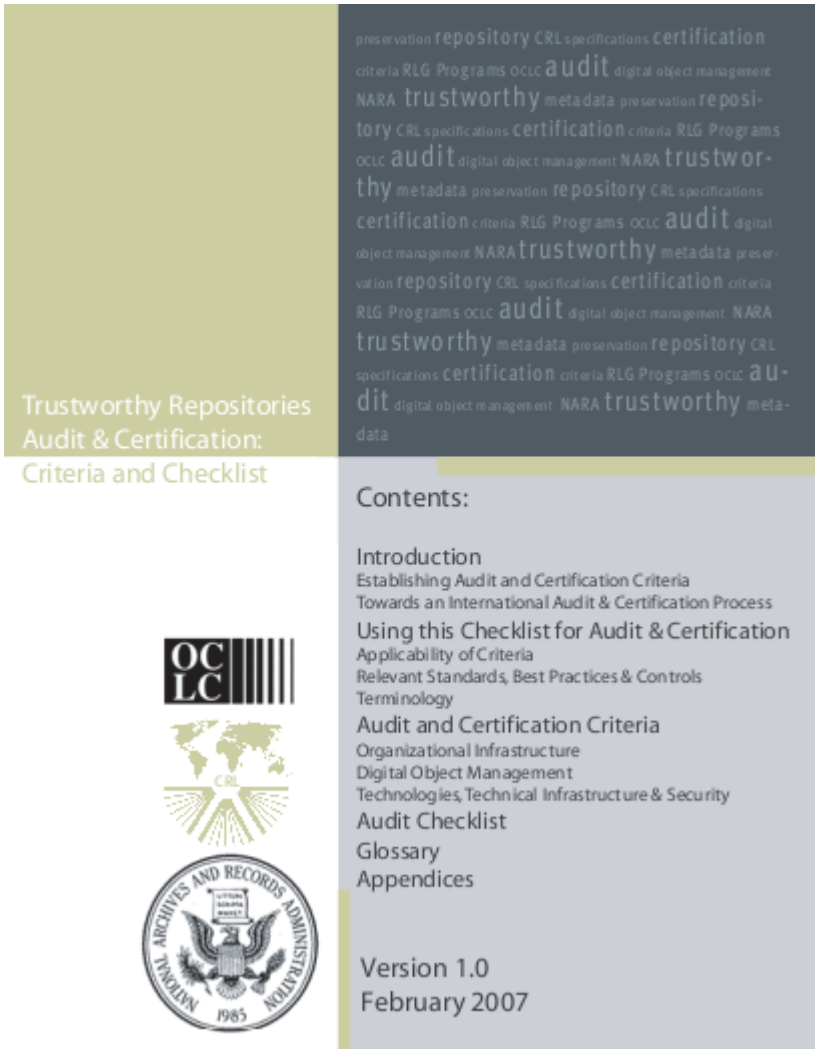


https://www.archivematica.org/wiki/OAIS_Use_Cases
https://www.archivematica.org/wiki/OAIS_Activity_Diagrams

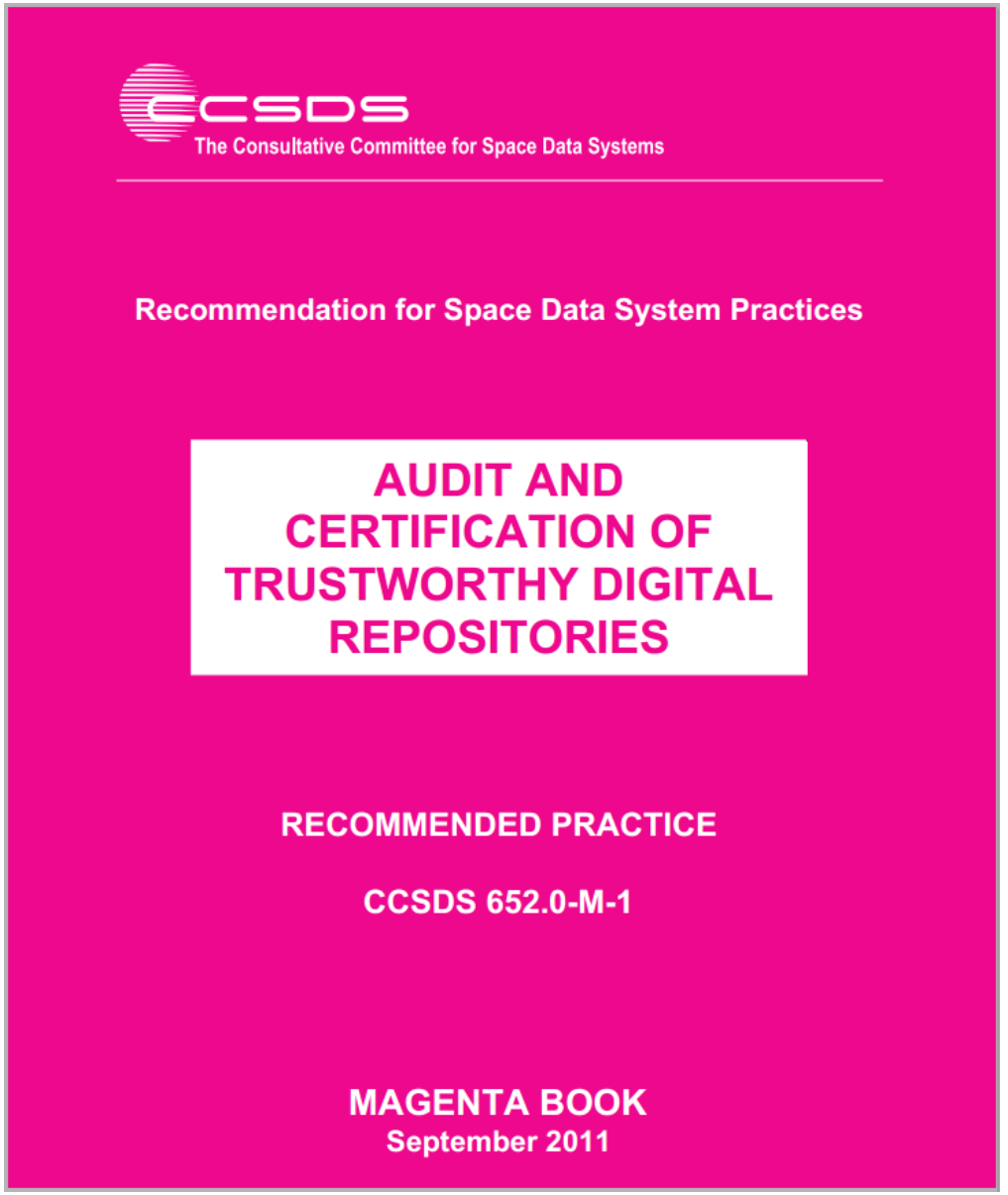
Capacity Gap

- No preservation features in key systems
- No digital preservation planning
  - Format obsolescence
  - System & platform incompatibility/obsolesence
  - Digital preservation metadata
  - External media processing
  - Dedicated storage and geo-remote backup
- No Trusted Digital Repository (TDR)
- No obvious next steps to improve capacity

TRAC: Trustworthy Digital Repositories



ISO 16363:2012



Schedule

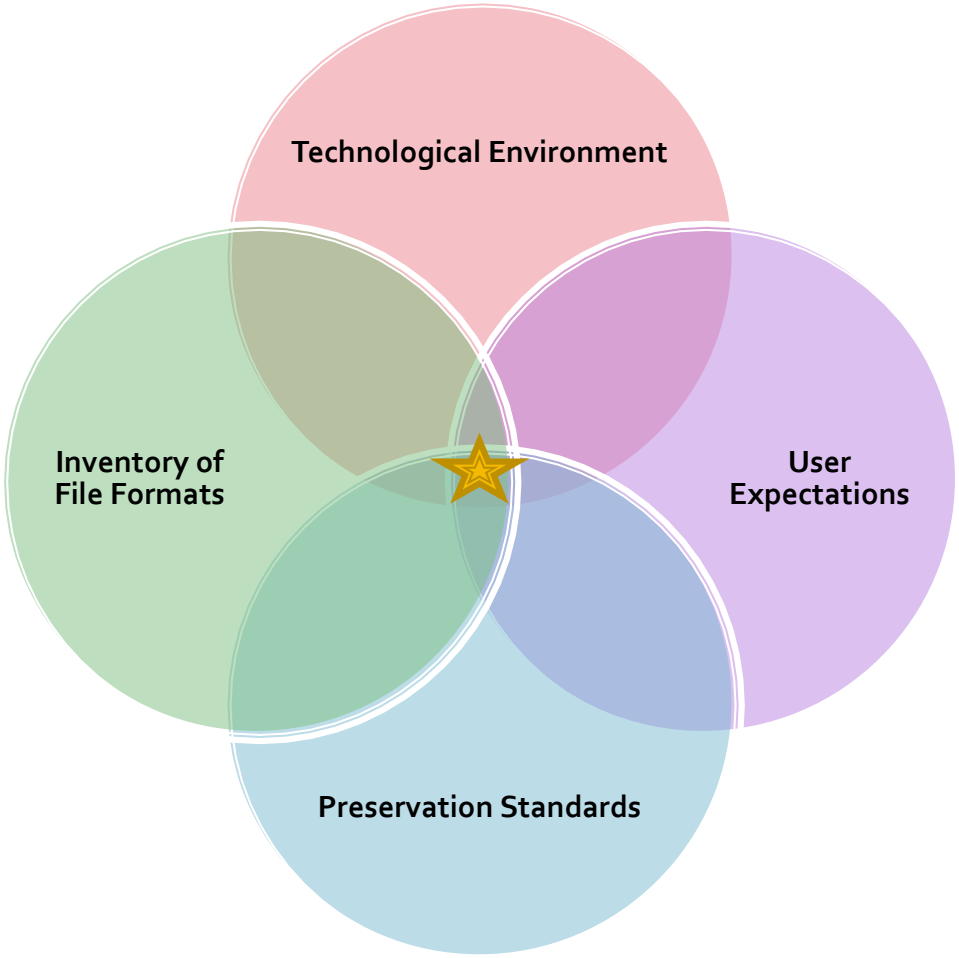
- Introduction to Digital Preservation
- **Media Type Preservation Planning**
- Current State of Preservation Tools
- E-Theses Preservation
- PDF to PDF/A Migration Workflow

  <BREAK>
- Archivematica
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- Q and A

Media Type Preservation Planning

- The Media Type Preservation Plan (MTPP), aka "Format Policies", determines what file formats to use for digitizing analog content and preserving born digital content. It is created from the intersection of supported media and file formats and their associated significant characteristics, the needs and expectations of our users, the underlying technological environment and digital preservation standards.

Media Type Preservation Plan (MTPP)



MTPP Digital Preservation Goals

- Provide long-term meaningful access to file formats across all browsers, operating systems and devices.
- Enable the display/play of the original look, feel, experience (colour, layout, etc.).
- Support reuse of born digital objects.

Caveat: to the extent possible

Media Type Preservation Plan (MTPP)

- Initial Steps
  1. Media Type: perform an inventory of media types and file formats in your physical collection/digital repository.
     a) File extensions don't provide enough information regarding a file (e.g. .pdf could be PDF or PDF/A)
     b) Use a file identification / validation tool like DRIOD or JHOVE2.

Media Type Preservation Plan (MTPP)

- Initial Steps
  1. Media Type: an inventory of file formats. digital

Media Type Preservation Plan (MTPP)

- Initial Steps
  2. Use Google Analytics or other web site tools to determine the predominant devices, operating systems and browsers used to access your services.

Media Type Preservation Plan (MTPP)

	Operating System	Visits	% Visits
1.	Windows	9,758	73.07%
2.	Macintosh	1,620	12.13%
3.	iOS	1,383	10.36%
4.	Android	304	2.28%
5.	Linux	89	0.67%
6.	BlackBerry	86	0.64%
7.	(not set)	71	0.53%
8.	Series40	12	0.09%
9.	Firefox OS	11	0.08%
10.	Chrome OS	6	0.04%

Media Type Preservation Plan (MTPP)

	Browser	Visits	% Visits
1.	Internet Explorer	4,799	35.93%
2.	Chrome	3,315	24.82%
3.	Firefox	2,292	17.16%
4.	Safari	2,171	16.26%
5.	Android Browser	236	1.77%
6.	Safari (in-app)	196	1.47%
7.	Opera Mini	86	0.64%
8.	BlackBerry	79	0.59%
9.	Opera	60	0.45%
10.	Mozilla Compatible Agent	50	0.37%

Media Type Preservation Plan (MTPP)

Mobile

	Operating System	Visits	% Visits
1.	iOS	1,383	73.96%
2.	Android	304	16.26%
3.	BlackBerry	86	4.60%
4.	(not set)	64	3.42%
5.	Series40	12	0.64%
6.	Firefox OS	11	0.59%
7.	Windows Phone	4	0.21%
8.	SymbianOS	3	0.16%
9.	Nokia	2	0.11%
10.	Samsung	1	0.05%

Media Type Preservation Plan (MTPP)

- Initial Steps
  3. Monitor the high-tech environment for changes which may impact your services.

Media Type Preservation Plan (MTPP)

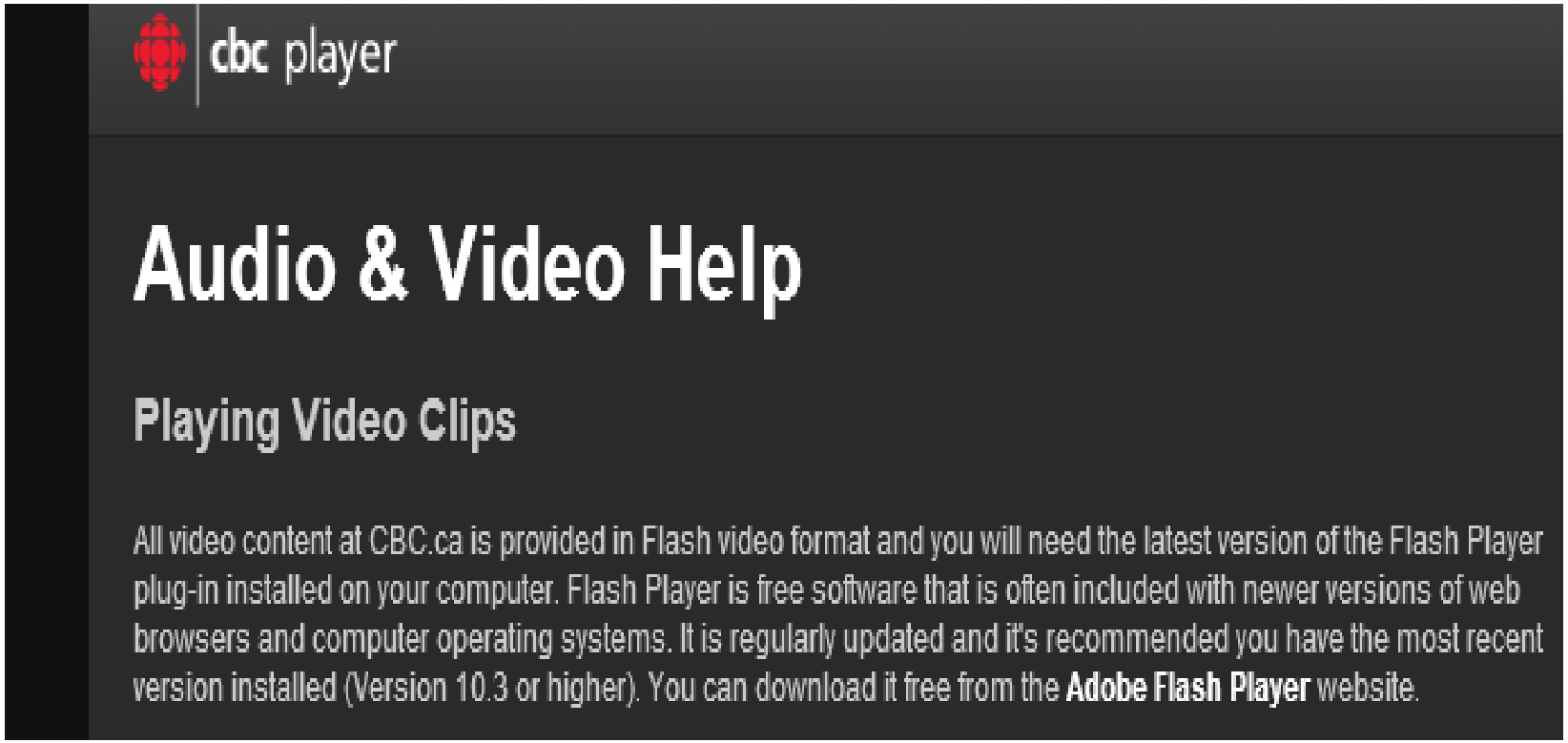
Steve Would Be Proud: How Apple Won The War Against Flash

RYAN LAWLER

Saturday, June 30th, 2012 **64 Comments**

Late Thursday, an extraordinary thing happened: Adobe announced in a blog post that it would **not provide Flash Player support for devices running Android 4.1**, and that it would pull the plugin from the Google Play store on August 15. The retreat comes five years after the introduction of the iPhone, the device which thwarted Flash's mobile ambitions, almost even before they began.

Media Type Preservation Plan (MTPP)



Media Type Preservation Plan (MTPP)

<Audio> element format support [edit]

This table documents the current support for audio codecs by the `<audio>` element.

Browser	Operating system	Formats supported by different web browsers					
		Ogg Vorbis	WAV PCM	MP3	AAC	WebM Vorbis	Ogg Opus
Google Chrome	All supported	9	Yes	Yes	Yes	Yes	25
Internet Explorer	Windows	No	No	9.0	9.0	No	No
Mozilla Firefox	All supported	3.5	3.5	21.0, Windows only	21.0, Windows only	4.0	15.0
Opera	All supported	10.50	11.00	No	No	10.60	No
Safari	OS X	No	3.1	3.1	3.1	No	No

Media Type Preservation Plan (MTPP)

Blackberry Z10 Supported File Formats

Image formats	BMP, WBMP, JPG, GIF, PNG, TIFF, SGI, TGA
Audio & video formats	3GP, 3GP2, M4A, M4V, MOV, MP4, MKV, MPEG-4, AVI, ASF, WMV, WMA, MP3, MKA, AAC, AMR, F4V, WAV, MP2PS, MP2TS, AWB, OGG, FLAC
Audio & video encoding/decoding	H.264, MPEG-4, H.263, AAC-LC, AAC+, eAAC, MP3, PCM, Xvid, AMR-NB, WMA 9/10, WMA10 professional, WMA-LL, VC-1, VP6, SPARK, PCM, MPEG-2, MJPEG (mov), AC-3, AMR-WB, QCELP, FLAC, VORBIS

Media Type Preservation Plan (MTPP)

- Initial Steps
  4. Monitor communications within the digital preservation community regarding recommended long-term preservation file formats.

Media Type Preservation Plan (MTPP)

Sustainability of Digital Formats
Planning for Library of Congress Collections

Introduction | Sustainability Factors | Content Categories | Format Descriptions | Contact

The Digital Formats Web site provides information about digital content formats. The analyses and resources presented here will increase and be updated over time. The compilers, Caroline R. Arms and Carl Fleischhauer, invite feedback on the content.

Introduction

Background information and overview: What is a format? How shall we evaluate formats? What projects in other organizations are addressing these questions? >>

Overview | Formats, Evaluation Factors, and Relationships | Papers and Presentations | Related Resources

Sustainability Factors

What affects the ability of the Library to preserve content in a given format? These sustainability factors apply to all formats. >>

Disclosure | Adoption | Transparency | Self-documentation | External Dependencies | Impact of Patents | Technical Protection Mechanisms

Content Categories

The evaluation of formats must take into account quality and functionality. These factors vary according to the type of content under consideration and the categories will be expanded as time passes. >>

Still Image | Sound | Textual | Moving Image | Web Archive | Datasets | Geospatial | Generic

Format Descriptions

Documents with more information about specific formats. >>

Browse categories | Browse alphabetical list

Media Type Preservation Plan (MTPP)

- Sustainability Factors for Long-Term Preservation Formats:
  - Disclosure: open specifications and validation tools.
  - Adoption: widely adopted, bundled with OS.
  - Transparency: open to direct analysis with basic tools.
  - Self-documentation: metadata embedded in file.
  - External dependencies: the fewer the better.
  - Impact of patents: licenses and royalty fees.
  - Technical protection mechanisms: the fewer the better.

Media Type Preservation Plan (MTPP)

- Sustainability Factors for MP3 Audio Format:

Sustainability factors

Disclosure	Open standard. Developed by the Motion Pictures Expert Group (MPEG), Coding of audio, picture, multimedia and hypermedia information.
Documentation	(1) MPEG-1: ISO/IEC 11172-3. Information technology -- Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s -- Part 3: Audio. (2) MPEG-2: ISO/IEC 13818-3. Information technology -- Generic coding of moving pictures and associated audio information -- Part 3: Audio. These specifications describe the syntax and semantics for three classes of compression methods known as Layers I, II, and III. MP3 is Layer III. See list of ISO documents in Format specifications below; see also MPEG-1 and MPEG-2.
Adoption	Widely adopted for World Wide Web dissemination and playback on specialized devices. Many software tools exist for encoding and decoding.
Licensing and patents	Various authorities cite a number of patent claims associated with MP3; see for example A Big List of MP3 Patents (and supposed expiration dates) (consulted in March 2008). The practical impact of these claims is not clear to the compiler of this document.
Transparency	Depends upon algorithms and tools to read; requires sophistication to build tools.
Self-documentation	Technical (coding) information is contained in the headers for the "frames" that make up the MP3 bitstream. The lack of *descriptive* metadata motivated the producer community to develop ID3, a separately specified structure for metadata to support discovery and other purposes.
External dependencies	None
Technical protection considerations	None

Media Type Preservation Plan (MTPP)

- Sustainability Factors for WAV Audio Format:

Sustainability factors

Disclosure	Fully documented. Proprietary format developed by Microsoft and IBM as part of the Resource Interchange File Format (RIFF) for Windows 3.1, with documentation freely available.
Documentation	Multimedia Programming Interface and Data Specifications 1.0. IBM Corporation and Microsoft Corporation, August 1991. Available online, e.g., at http://www.tactilemedia.com/info/MCI_Control_Info.html Multimedia Data Standards Update April 15, 1994 at http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/Docs/RIFFNEW.pdf
Adoption	Widely adopted. With LPCM encoding, a preferred or recommended format for sound in many long-term archives. Examples include Florida Digital Archive, DSpace at MIT, Libraries and Archives Canada.
Licensing and patents	No licensing required.
Transparency	Depends on audio codec employed for bitstream encoding (which may incorporate compression); see LPCM, μ-Law, A-Law, DPCM, and ADPCM.
Self-documentation	Metadata can be placed in the INFO chunk (aka "LIST" chunk with a list type of "INFO") associated with all RIFF files. Additional metadata is a feature of the *bext* (Broadcast Audio Extension) associated with WAVE_LPCM_BWF (Broadcast WAVE Audio File Format). Additional metadata chunks have been defined: aXML, iXML, and the CART/Audio Delivery Extension to BWF, from the Audio Engineering Society in AES46-2002.
External dependencies	None
Technical protection considerations	None

Media Type Preservation Plan (MTPP)

- MUN File Format Policy for Audio
  1. Media Type: Audio.
  2. Supported ingest format extensions: MP3, WMA, WAV
  3. Long-term preservation format(s): WAV (LPCM)
  4. Access format(s): MP3, WAV
  5. Normalization tool: Soundforge Audio Studio Version 10
  6. Analog digitization / born digital preservation audio standards

Media Type Preservation Plan (MTPP)

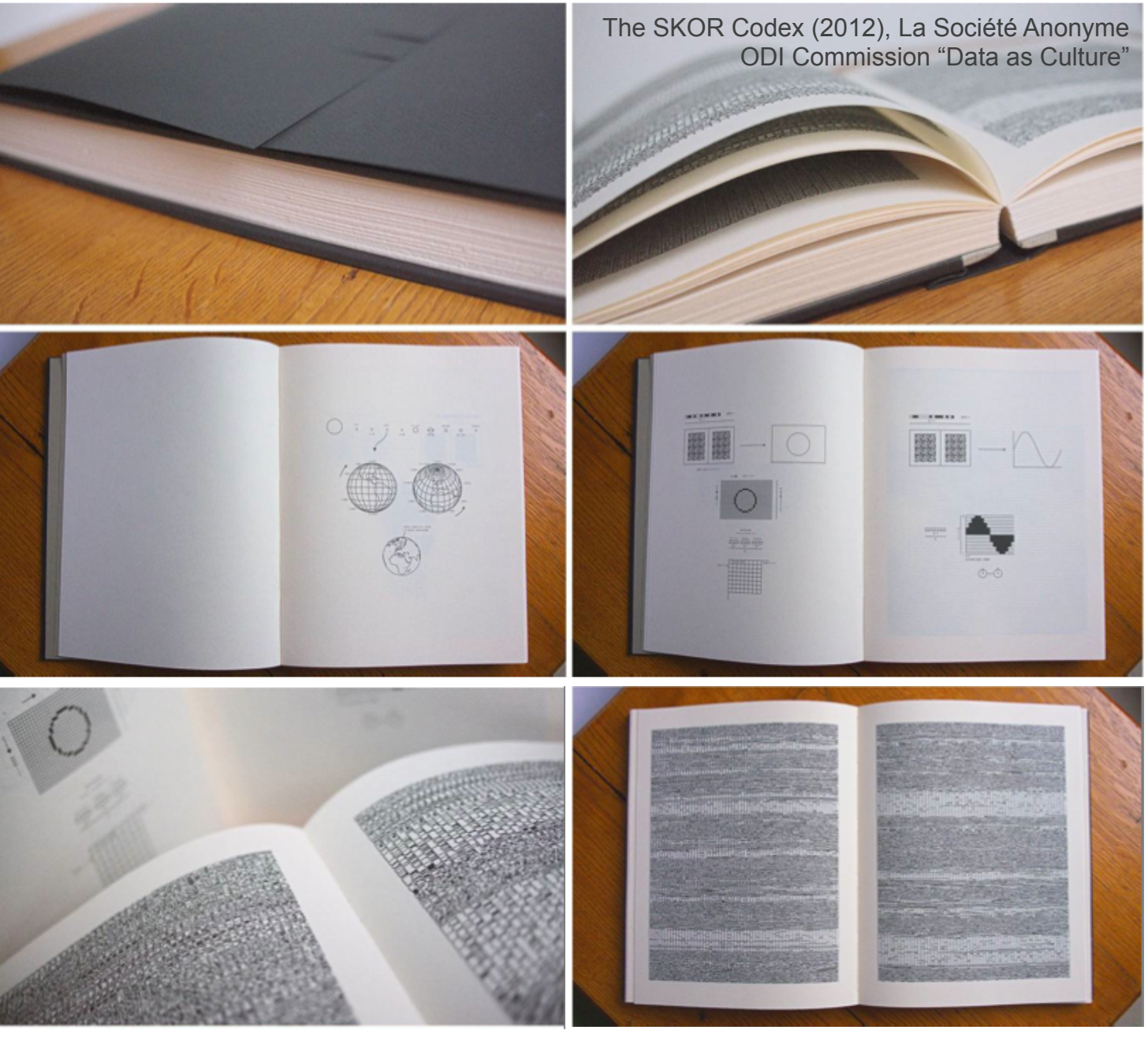
- Analog digitization / born digital preservation audio standards:

Category	File Format	Resolution	Notes
Long-term preservation high resolution standard	-Linear PCM bit stream -Uncompressed WAV	24 Bit @ 48 KHz (minimum), mono/ dual or mono/ stereo, interleaved	Higher resolutions encouraged if possible.
Deliverable "hard copy" standard (CD)	-Linear PCM bit stream - Uncompressed .WAV	16 Bit @ 44.1 KHz Stereo Interleaved only	Derive from original
Deliverable Web-based access/ download standard	MP3, WAV	Minimum of 128 Kbps @ 44.1 KHz mono or 256 Kbps @ 44.1 Kbps stereo	Derive from original

Schedule

- Introduction to Digital Preservation
- Media Type Preservation Planning
- **Current State of Preservation Tools**
- E-Theses Preservation
- PDF to PDF/A Migration Workflow

  <BREAK>
- Archivematica
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- Q and A



Digital Preservation
State of the Art

Slides by David Tarrant

UNIVERSITY OF Southampton

ROT

flickr stuckincustoms/7981697149

WHAT


USE
lacatholique/8407322574
flickr

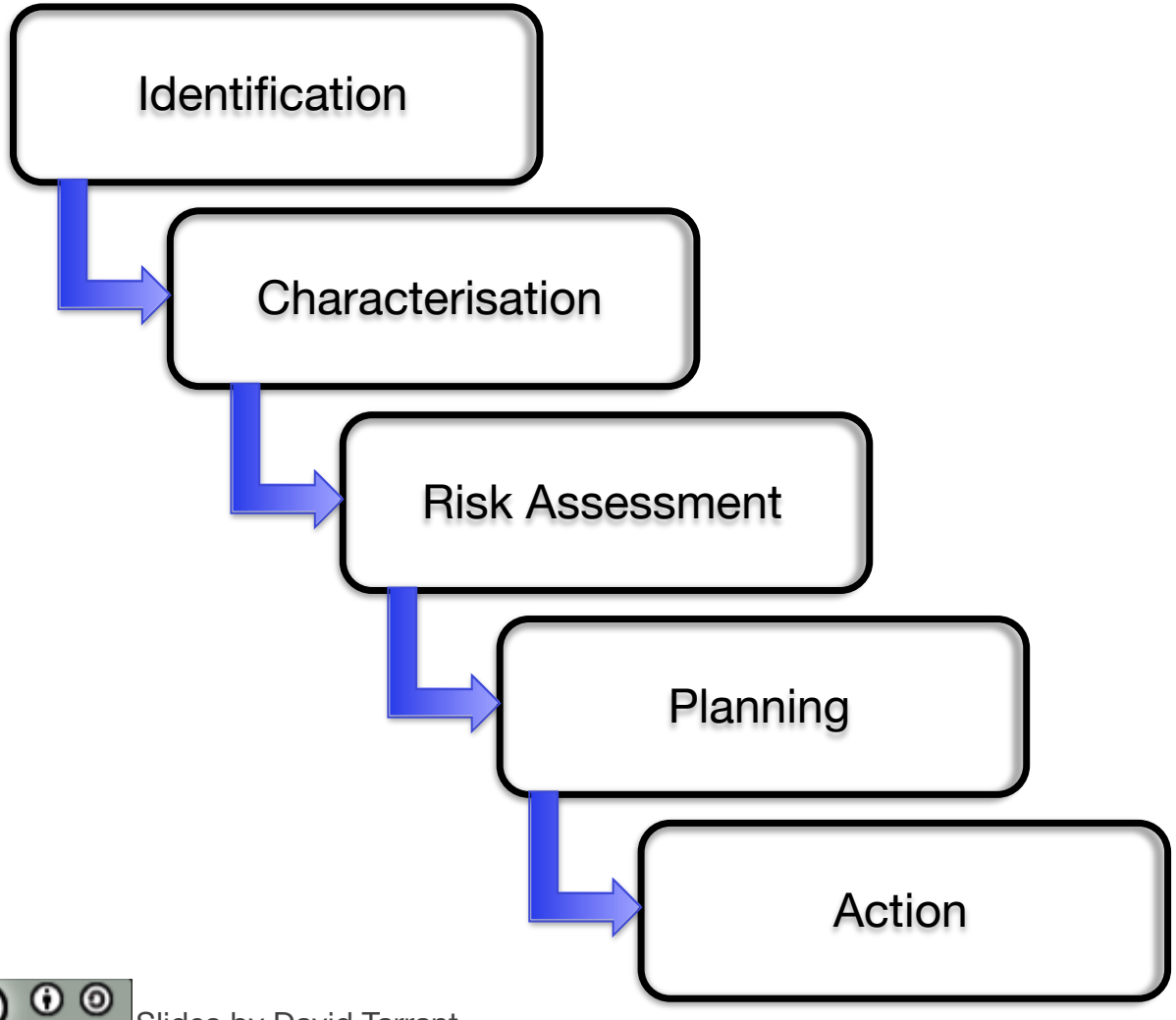

RISK
flickr fayjo/333325967

PLAN

ACTION

adilson_aracaju/5325377654 flickr

Flow...



DROID

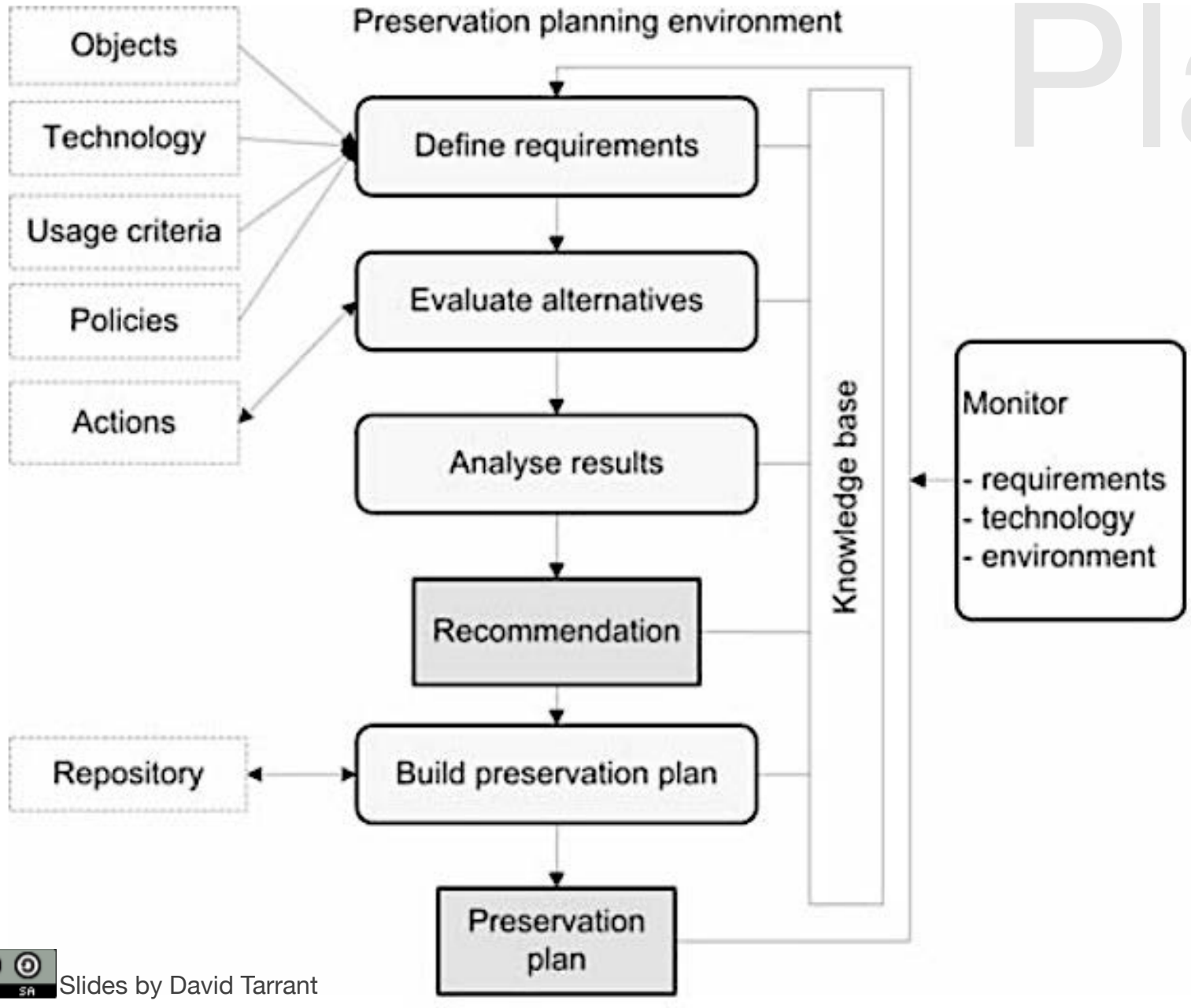
C3PO

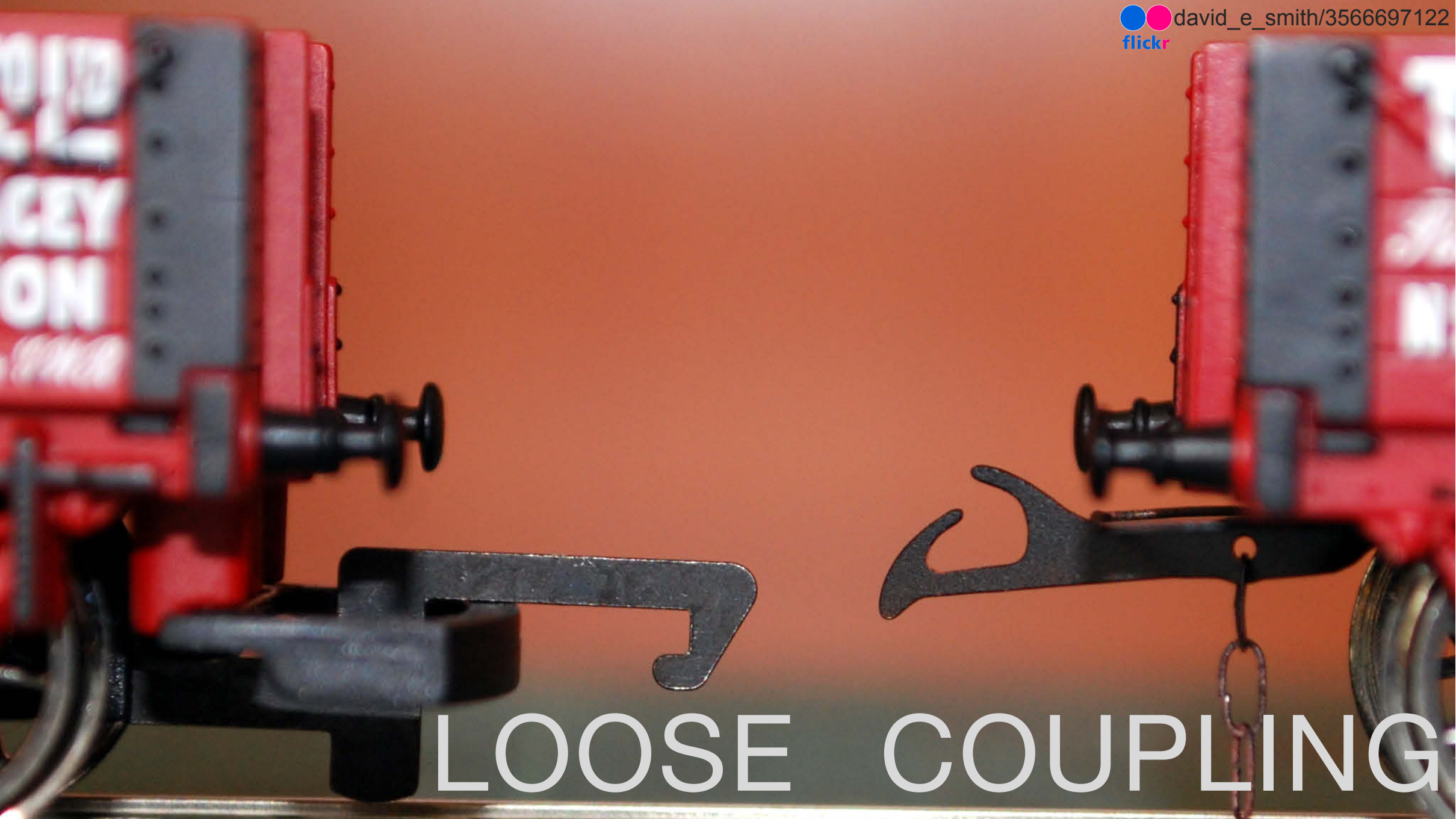
Scout

PLATO

FITS

Planning



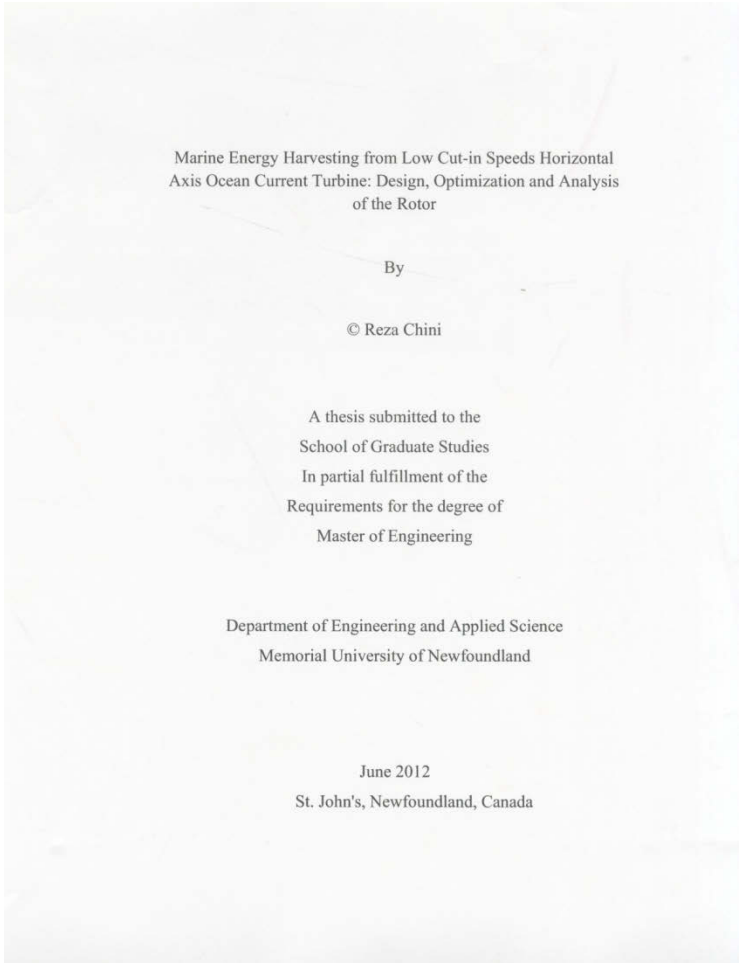

david_e_smith/3566697122
flickr
LOOSE COUPLING

Schedule

- Introduction to Digital Preservation
- Media Type Preservation Planning
- Current State of Preservation Tools
- **E-Theses Preservation**
- PDF to PDF/A Migration Workflow

  <BREAK>
- Archivematica
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- Q and A

E-Theses Preservation



Marine Energy Harvesting from Low Cut-in Speeds Horizontal Axis Ocean Current Turbine: Design, Optimization and Analysis of the Rotor

By

© Reza Chini

A thesis submitted to the
School of Graduate Studies
In partial fulfillment of the
Requirements for the degree of
Master of Engineering

Department of Engineering and Applied Science
Memorial University of Newfoundland

June 2012
St. John's, Newfoundland, Canada

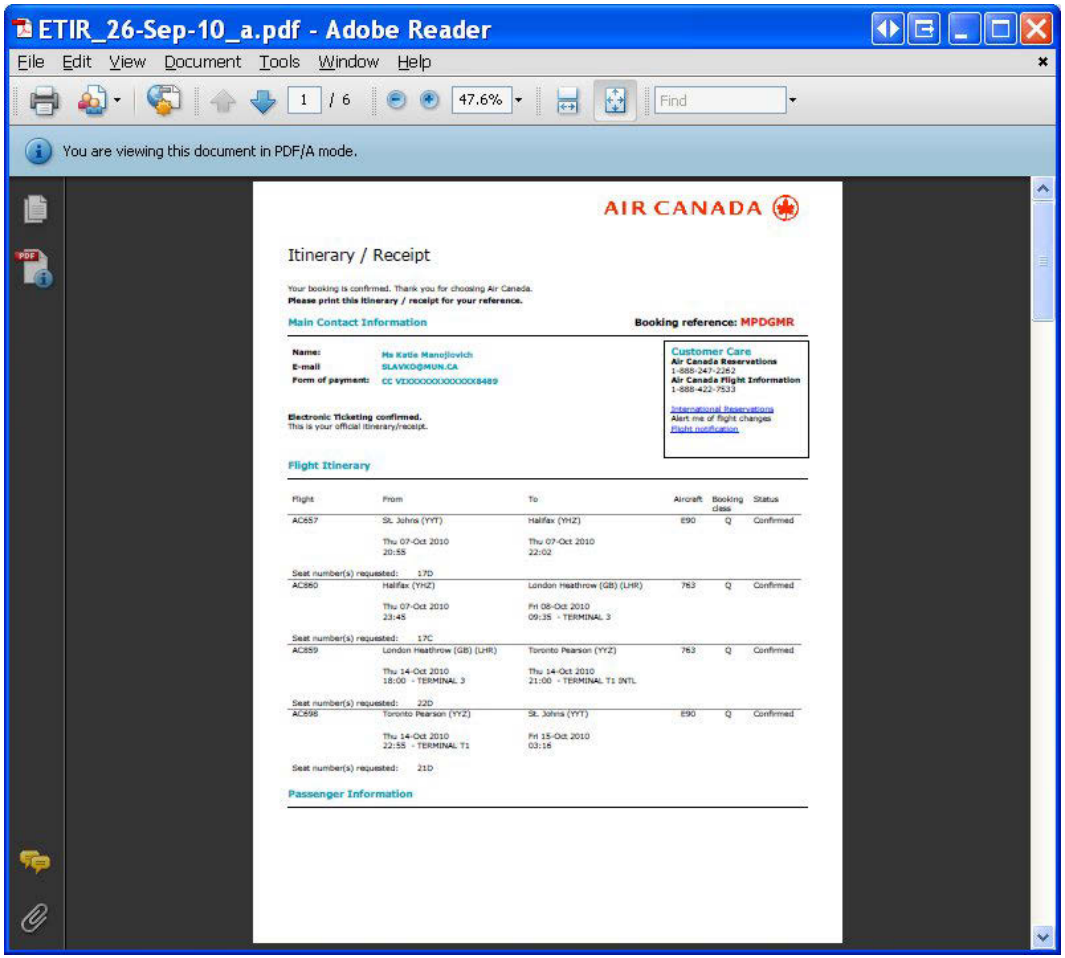
Electronic Theses (E-Theses)

- Library and Archives Canada
  "By 2014, LAC will only accept theses and dissertations from Canadian universities in electronic form…" OAI-PMH ETD-MS
- Requires
  - Change in Graduate Studies policy which will mandate electronic submission.
  - Additional instructions for e-thesis preparation.
  - Development of an e-thesis submission and processing workflow.

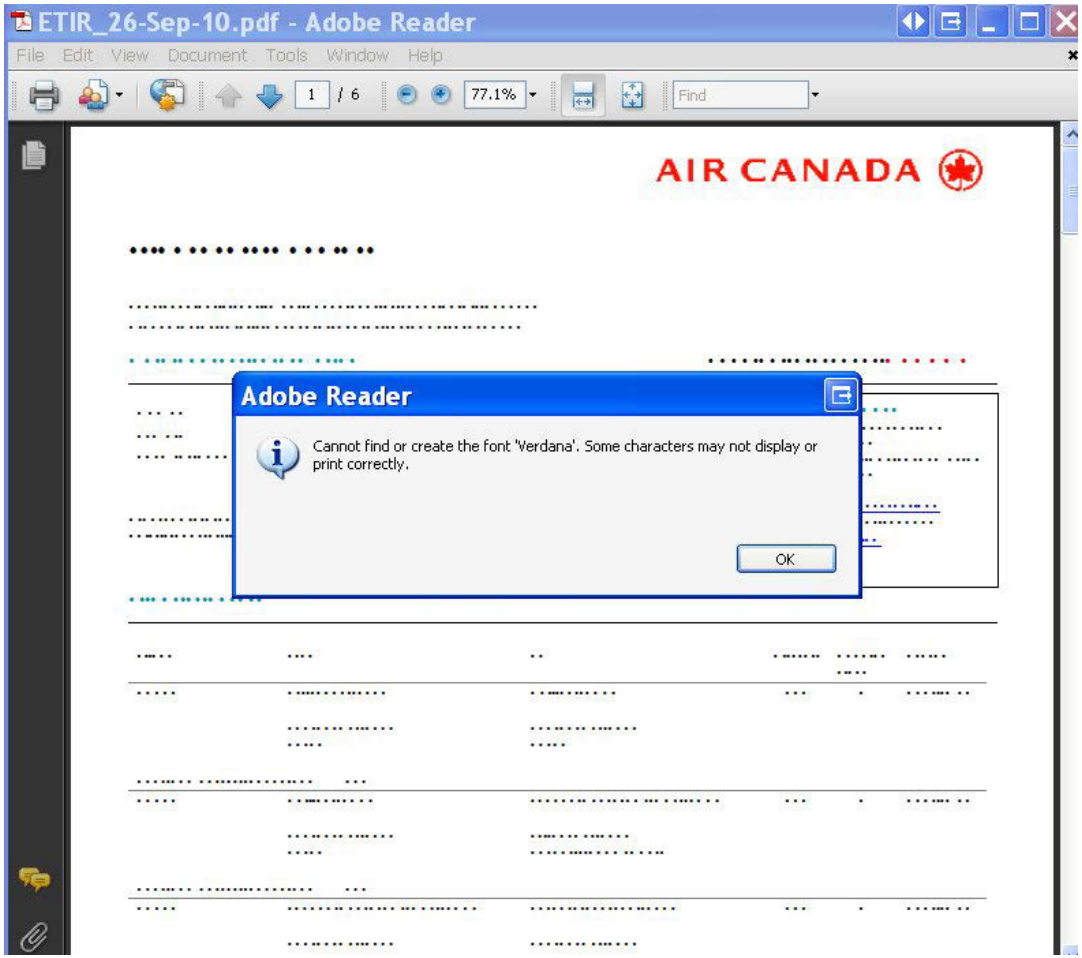
PDF/A for Preservation of and Access to Text –based Files

- PDF/A file format maximizes:
  - Device independence
  - Self-containment
  - Self-documentation
- 16 new fonts in Windows 8 plus many updated fonts.

PDF/A Preserves Look and Feel



PDF Does Not (missing fonts)



E-Theses Supplementary Files

Supplemental Files in Electronic Theses and Dissertations: Implications for Policy and Practice

Sarah L. Shreeves - sshreeve@illinois.edu - @sshreeves
University of Illinois at Urbana-Champaign
IDCC 2013 – Amsterdam, Netherlands – 14-16 January 2013

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

ETDs and Supplemental Files at UIUC

- ETDs and supplemental files (in all formats) accepted since 2010.
- Must be approved by the advisor / committee.
- Must be described in an appendix.
- No standard for description in Grad College policy.
- Currently investigating what policies should be in place, if any, for format, metadata, and other requirements for supplemental files.

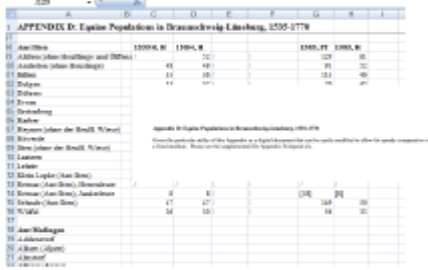
3538 ETDs deposited between 2010-05 and 2012-08

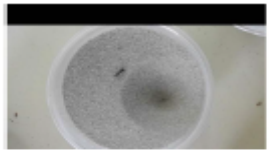
- 59% Doctoral Dissertations
- 41% Masters Theses
- 63% Immediately Open Access
- 20% Illinois only for 2 years
- 17% Closed for 2 years

Of these **77 (2%)** ETDs deposited with at least one supplemental file

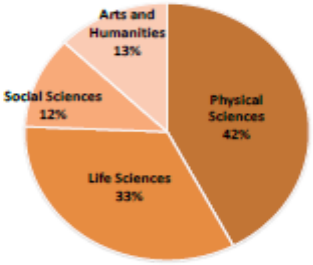
- 47% Doctoral dissertations
- 53% Masters Theses
- 73% Immediately Open Access
- 19% Illinois only for 2 years
- 8% Closed for two years

Examples of Supplemental Files and Documentation in Appendix





Supplemental File(s) by Discipline

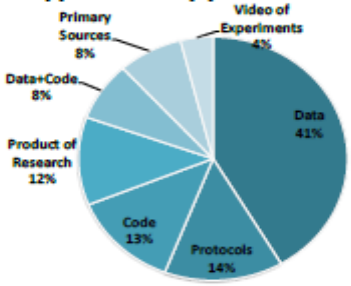


Numbers of Supplemental Files

29% - 1 file
36% - 2-5 files
16% - 6-20 files
19% - 21 or more files

One thesis had over **2000** supplemental files (data+code) included in a **zip** file

Types of Supplemental Files



Formats of Supplemental Files

Preliminary Findings

Wide variety of formats. Code is a particular concern.

Type of supplementary file also varies. Data and data+code make up about half of the files.

Basic description of what a supplementary file is generally included in the appendix. Documentation necessary for use of supplementary file varies. In most cases, this level of documentation is not in the appendix, but contained within the text of the ETD itself.

Definite need for more structured description of supplemental files that can assist both in use and long term access and preservation.

More investigation needed into when supplemental files are included and when they are not.

E-Theses Supplementary Files

- Integral part of the thesis.
- Support replication, validation and extension of experiments.
- Support reuse of digital content, for example, images.
- MUN Disciplines: Anthropology, Biology, Chemistry, Computer Science, Earth Sciences, Engineering, Environmental Science, Music, Physics.

E-Theses Supplementary Files

Gan, Gregory (2010) To our hopeless affair : a visual anthropology study about women of the Russian Intelligentsia in the post-Soviet era. Masters thesis, Memorial University of Newfoundland.

"A feature-length ethnographic film produced during the period of fieldwork in Moscow and based on participants' memories is appended to the thesis."

Here is a link to the Gregory Gan thesis and video in Memorial's Research Repository:

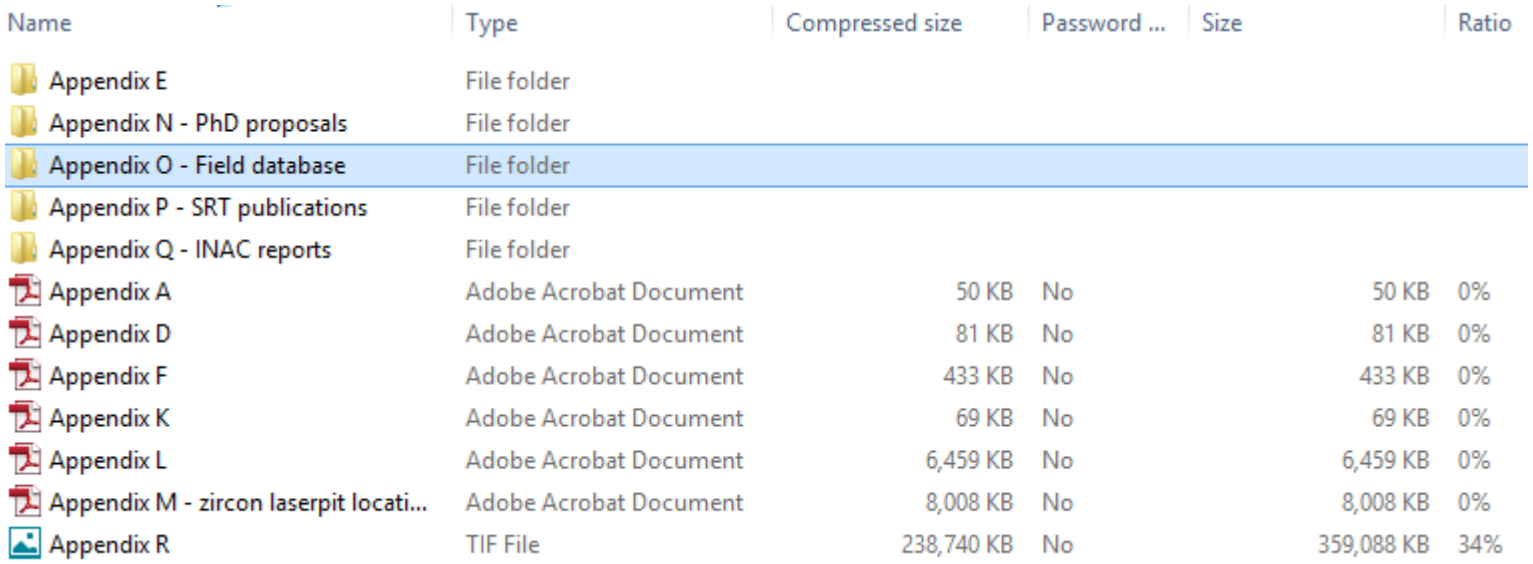
http://research.library.mun.ca/1686/

E-Theses Supplementary Files

- Types of files submitted with 20 MUN theses:
  - Video: mp4, wmv, avi
  - Image: jpg, tif
  - Spreadsheet: Microsoft Excel
  - Database: Microsoft Access
  - Software: original C#/Fortran code, open source software, links to commercial software.
  - Data: asv CAD/CAM, cif Crystallographic, etc.
  - Documentation: from good to nil.
  - Many theses have files organized in folders.

E-Theses Supplementary Files

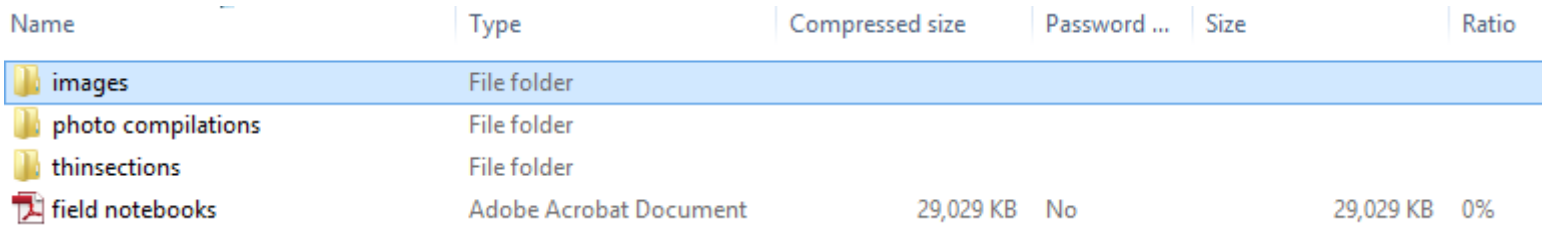
- E-Theses with supplementary files in folders



Name	Type	Compressed size	Password ...	Size	Ratio
Appendix E	File folder				
Appendix N - PhD proposals	File folder				
Appendix O - Field database	File folder				
Appendix P - SRT publications	File folder				
Appendix Q - INAC reports	File folder				
Appendix A	Adobe Acrobat Document	50 KB	No	50 KB	0%
Appendix D	Adobe Acrobat Document	81 KB	No	81 KB	0%
Appendix F	Adobe Acrobat Document	433 KB	No	433 KB	0%
Appendix K	Adobe Acrobat Document	69 KB	No	69 KB	0%
Appendix L	Adobe Acrobat Document	6,459 KB	No	6,459 KB	0%
Appendix M - zircon laserpit locati...	Adobe Acrobat Document	8,008 KB	No	8,008 KB	0%
Appendix R	TIF File	238,740 KB	No	359,088 KB	34%

E-Theses Supplementary Files

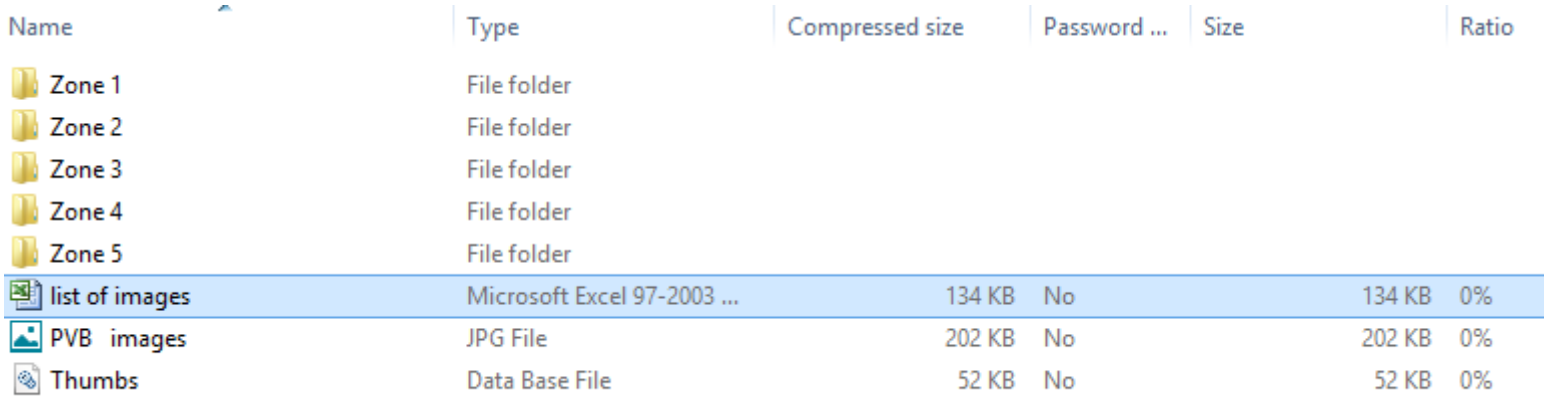
- E-Theses with supplementary files in folders



Name	Type	Compressed size	Password ...	Size	Ratio
images	File folder				
photo compilations	File folder				
thinsections	File folder				
field notebooks	Adobe Acrobat Document	29,029 KB	No	29,029 KB	0%

E-Theses Supplementary Files

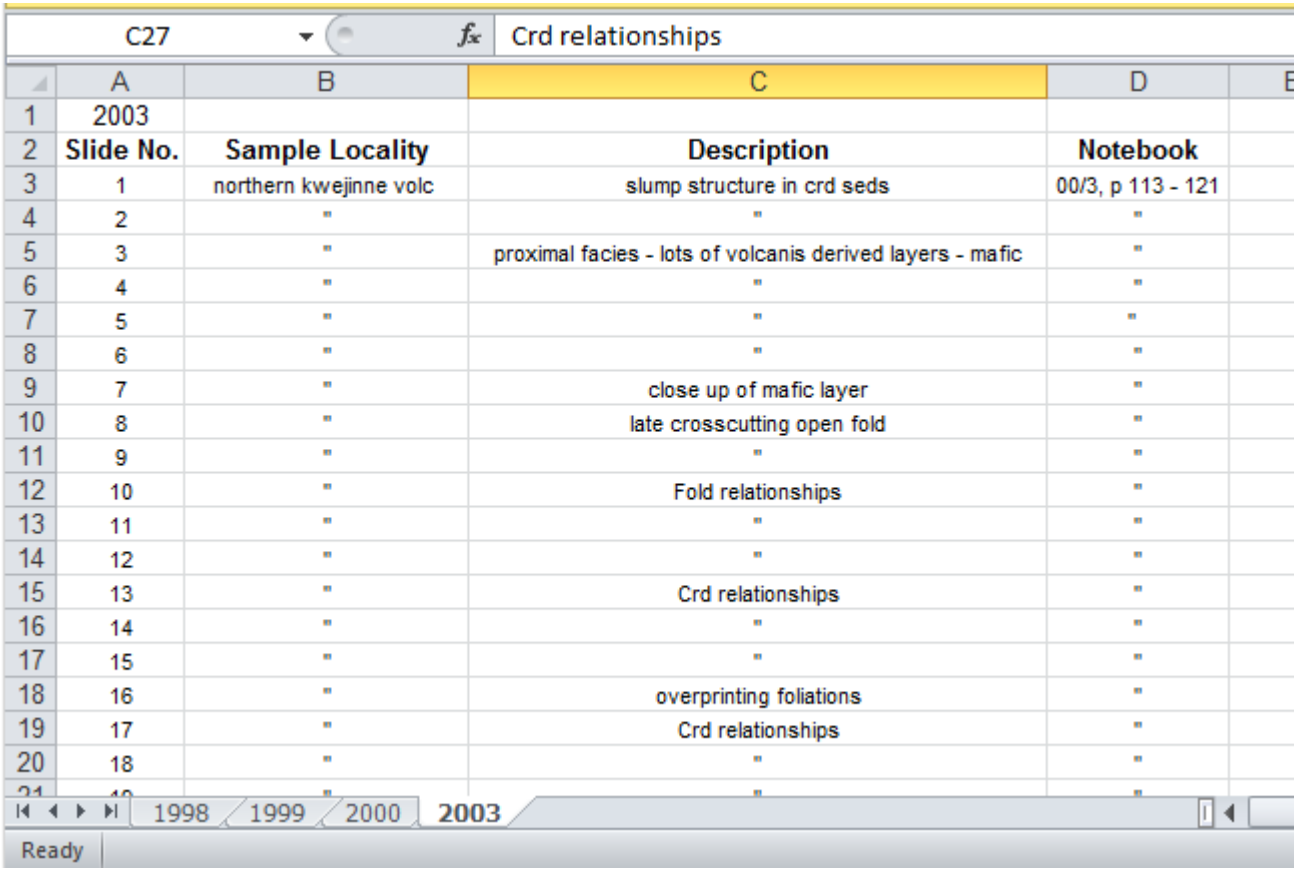
- E-Theses with supplementary files in folders



Name	Type	Compressed size	Password ...	Size	Ratio
Zone 1	File folder				
Zone 2	File folder				
Zone 3	File folder				
Zone 4	File folder				
Zone 5	File folder				
list of images	Microsoft Excel 97-2003 ...	134 KB	No	134 KB	0%
PVB images	JPG File	202 KB	No	202 KB	0%
Thumbs	Data Base File	52 KB	No	52 KB	0%

E-Theses Supplementary Files

- E-Theses with supplementary files in folders



C27 | Crd relationships

	A	B	C	D
1	2003			
2	**Slide No.**	**Sample Locality**	**Description**	**Notebook**
3	1	northern kwejinne volc	slump structure in crd seds	00/3, p 113 - 121
4	2	"	"	"
5	3	"	proximal facies - lots of volcanis derived layers - mafic	"
6	4	"	"	"
7	5	"	"	"
8	6	"	"	"
9	7	"	close up of mafic layer	"
10	8	"	late crosscutting open fold	"
11	9	"	"	"
12	10	"	Fold relationships	"
13	11	"	"	"
14	12	"	"	"
15	13	"	Crd relationships	"
16	14	"	"	"
17	15	"	"	"
18	16	"	overprinting foliations	"
19	17	"	Crd relationships	"
20	18	"	"	"

1998 | 1999 | 2000 | **2003**

Ready

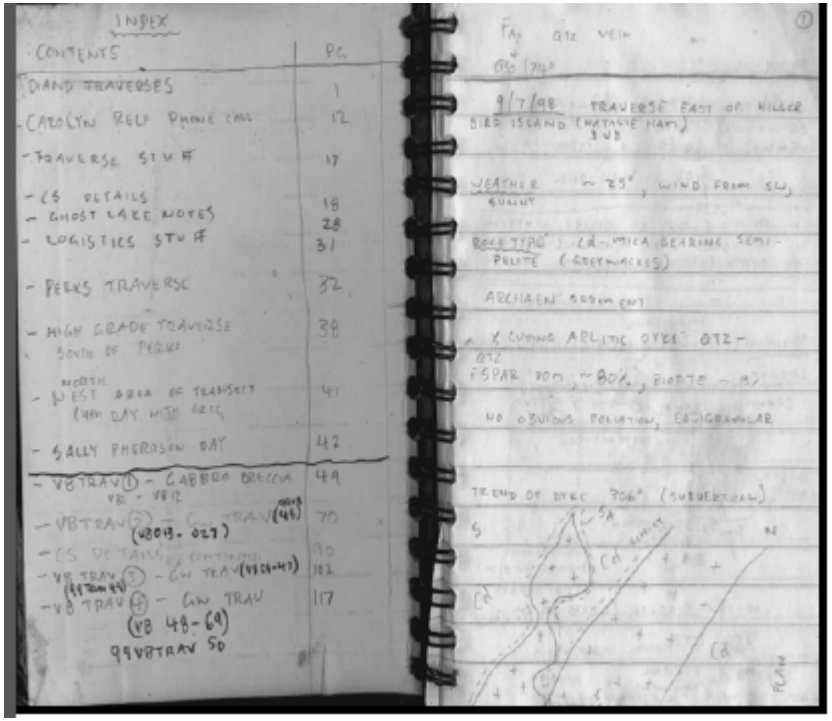
E-Theses Supplementary Files

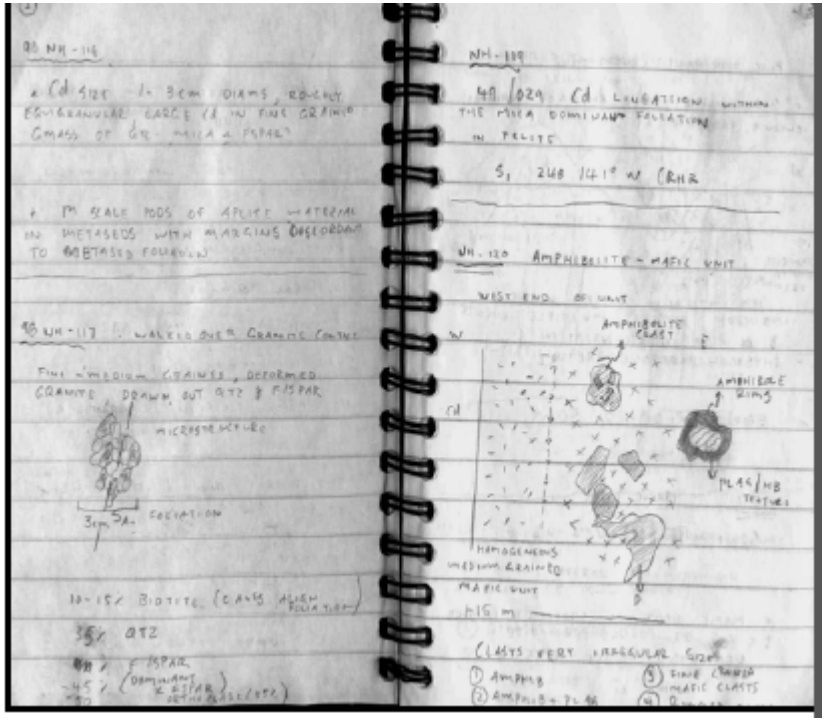
- E-Theses with supplementary files in folders

VB6

E-Theses Supplementary Files

- E-Theses with supplementary files in folders





E-Theses Supplementary Files

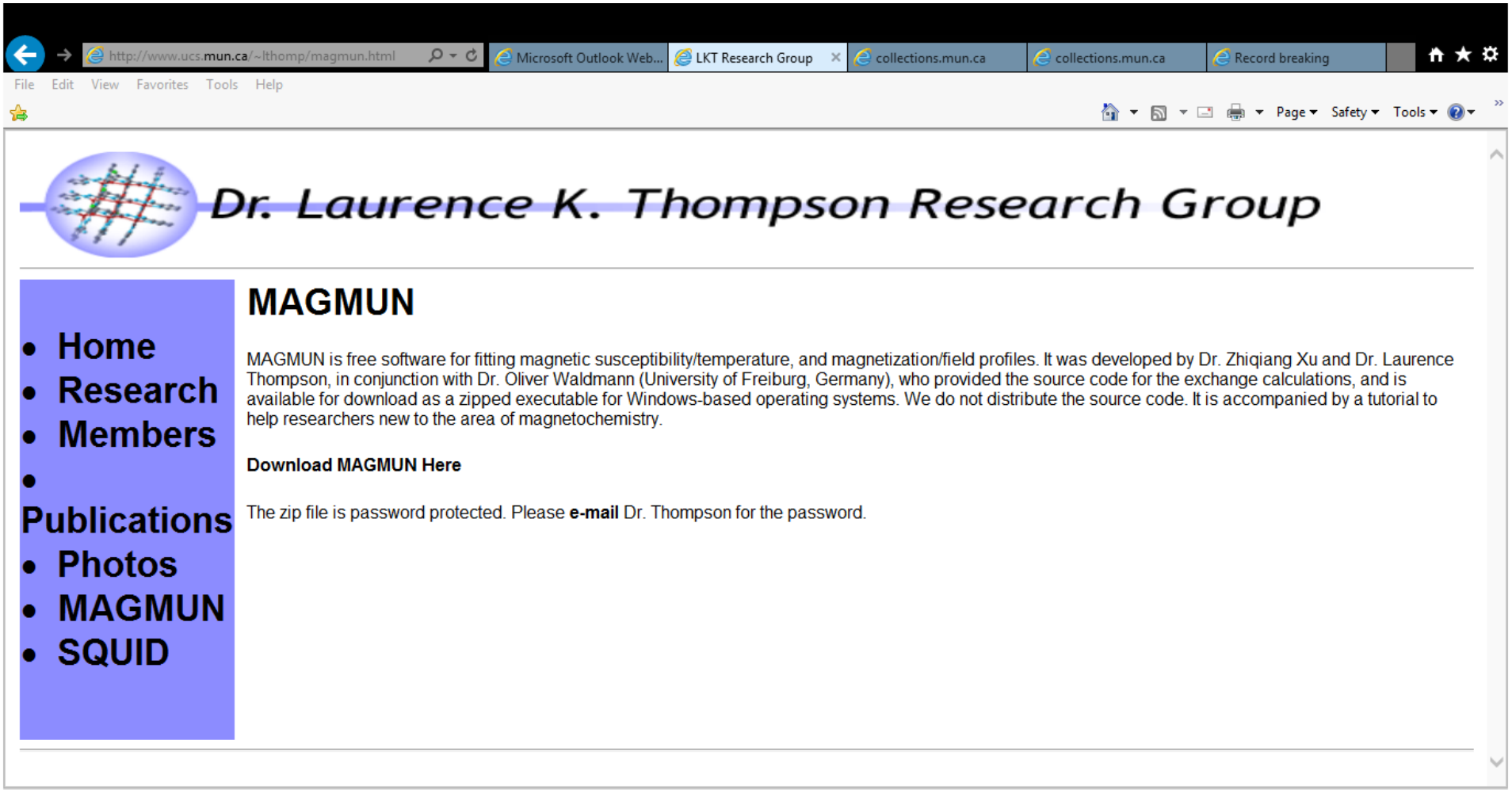
- Convert known files for long term access / preservation based on the MTTP:
  - Video: mp4, wmv, avi
  - Image: jpg, tif
  - Spreadsheet: Microsoft Excel
  - Database: Microsoft Access
  - Software: original C#/Fortran code, open source software, commercial software.
  - Data: asv CAD/CAM, cif Crystallographic, etc.
  - Documentation: from good to nil (readme files).

E-Theses Supplementary Files

- Download, test, preserve open source software, compilers, emulators, etc.:
  - Video: mp4, wmv, avi
  - Image: jpg, tif
  - Spreadsheet: Microsoft Excel
  - Database: Microsoft Access
  - Software: original C#/Fortran code, open source software, commercial software.
  - Data: asv CAD/CAM, cif Crystallographic, etc.
  - Documentation: from good to nil (readme files).

E-Theses Supplementary Files

Download and preserve the associated software, if possible.

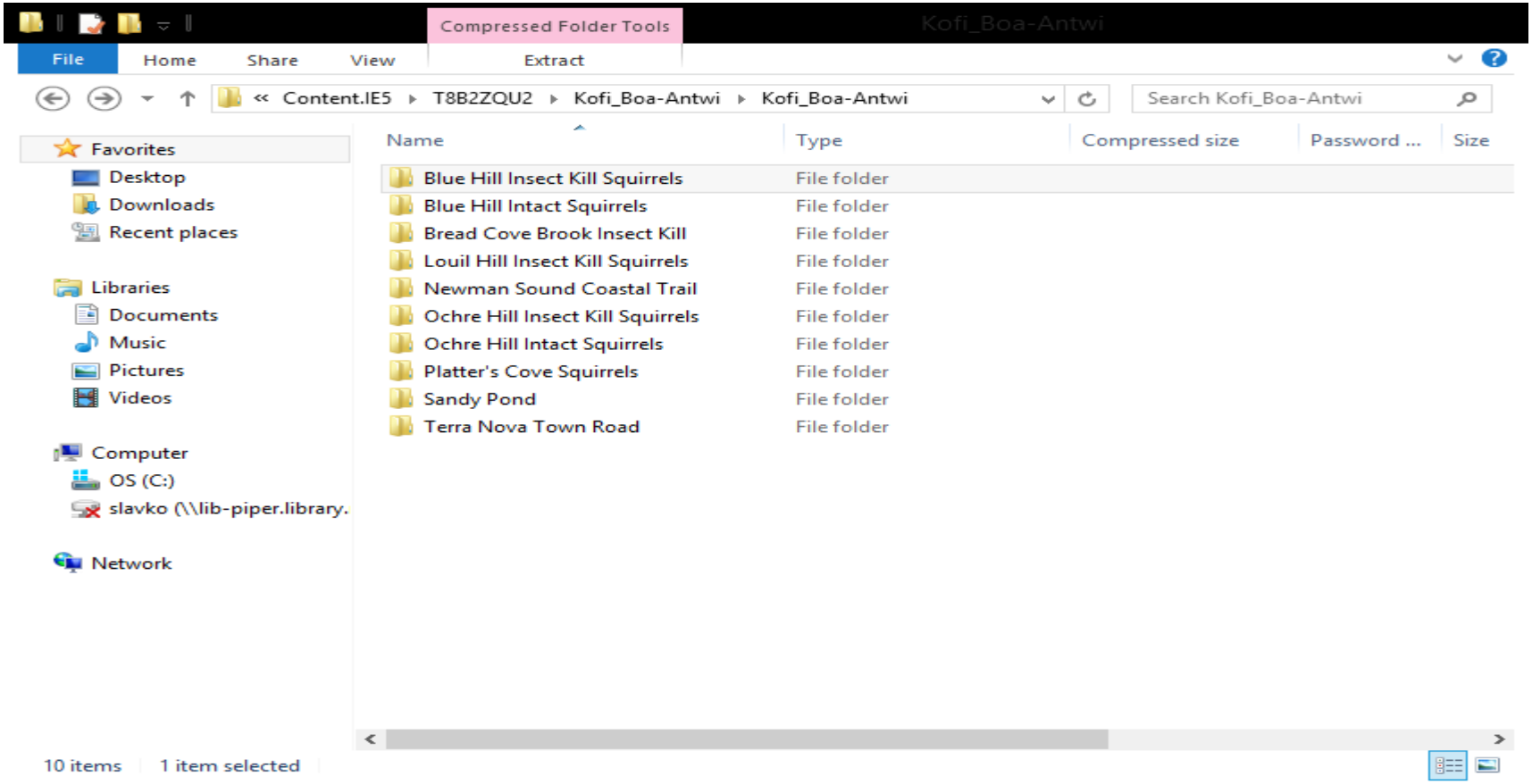


E-Theses Supplementary Files

- Determine sustainability of remaining formats:
  - Video: mp4, wmv, avi
  - Image: jpg, tif
  - Spreadsheet: Microsoft Excel
  - Database: Microsoft Access
  - Software: original C#/Fortran code, open source software, commercial software.
  - Data: asv CAD/CAM, cif Crystallographic, etc.
  - Documentation: from good to nil.

E-Theses Supplementary Files

- Use zip uncompressed to preserve the folder structure until metadata is created



E-Theses Supplementary Files

- Metadata
  - ETD-MS Dublin core profile is inadequate.
  - Available METS Profiles:
    - UC San Diego Electronic Theses and Dissertations Profile

    The primary document, be it a thesis or dissertation, is contained in a PDF file. Associated, archival quality files may also be included in a METS record for electronic theses and dissertations…. Structural relationships between the primary and associated files are not required in this profile.

E-Theses Supplementary Files

- Metadata (continued)
  - Available METS Profiles:
    - UC San Diego Complex Object Profile

    Hierarchical structural relationships should be detailed as nested elements according to the METS schema and rules because it is richer than that provided as PREMIS semantic units. If the scope of exchange objects is preservation, implementers should also use the PREMIS relationship elements in the Object schema for structural relationships. PREMIS relationship elements should always be used for derivative types of relationships.

E-Theses Supplementary Files

- Metadata (continued)
  - Next Steps
    - Add the structural relationship components of the UC San Diego METS Complex Object Profile to the Electronic Theses and Dissertations Profile.
    - The result will be a METS profile supporting:
      - Descriptive Metadata: Dublin Core, ETD-MS and MODS
      - Technical File Metadata: METS
      - Preservation and Rights Metadata: PREMIS and METS
      - Structural Metadata: METS and PREMIS
    - Package the metadata and associated files using Bagit and/or ZIP uncompressed.

E-Theses Supplementary Files

- Determine the skill sets required to staff the E-Thesis processing workflow:
  - Metadata specialist: METS, Premis, Dublin Core, Mods.
  - Digital Preservation specialist: file identification, validation, access and preservation.
  - Systems Analyst: file migration, cross platform/browser testing, packaging.
- Important question: Will LAC/BAC be able to accommodate E-Theses with supplementary files?

Schedule

- Introduction to Digital Preservation
- Media Type Preservation Planning
- Current State of Preservation Tools
- E-Theses Preservation
- **PDF to PDF/A Migration Workflow**

  <BREAK>
- Archivematica
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- Q and A

PDF to PDF/A Migration Workflow

- Preservation Goals
  - Why PDF/A, not PDF?
- General workflow considerations
  - Automation
  - Format identification
  - Format migration
  - Tool selection
- Workflow stages
- Implementation

* Preservation Goals * (Rationale for PDF/A vs PDF)

- General Usability
  - Platform independence
  - Self containment
  - Metadata inclusion
  - Searchable text
    - Indexable
    - Also facilitates reuse
- Long Term Access

* Workflow Considerations *

- Degree of Automation
  - Identification / Validation
  - Migration
- Format identification
  - Source
  - Target
- Format migration
- Tool selection

Automation

- Degree of automation
  - Validation Only?
  - Conversion?
  - On ingest?
- Dependencies:
  - Resources available for preservation actions
  - Fidelity required in migrated materials
- In our case, limited resources suggest need for as much automation as practicable

Format identification

- Need to identify PDF files as potential migration targets
  - Sub-task: PDF files which are not already PDF/A
- Need to validate migration output as compliant

Format identification (source)

- PDF-type format identification generally successful at a coarse level
- Since PDF is analogous to a container format, within workflow, a first pass doesn't particularly need to version the pdf
- Except for mis-named files, file extension testing may be sufficient
  - Mime type label an alternative where available

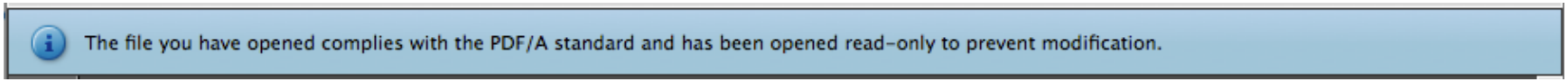
Format identification (target): "Which" PDF/A?

- PDF/A as a format is more nuanced
  - Currently many PDF/A "flavours"
    - 1 – Initial standard, layers, actions forbidden
    - 2 – Transparency / layers permitted, JPEG2000
    - 3 – Embedding of non-managed document formats
    - + Conformance level
      - A – accessible
      - B – basic
      - U – unicode
  - Changed as recently as Oct 2012 (PDF/A-3)

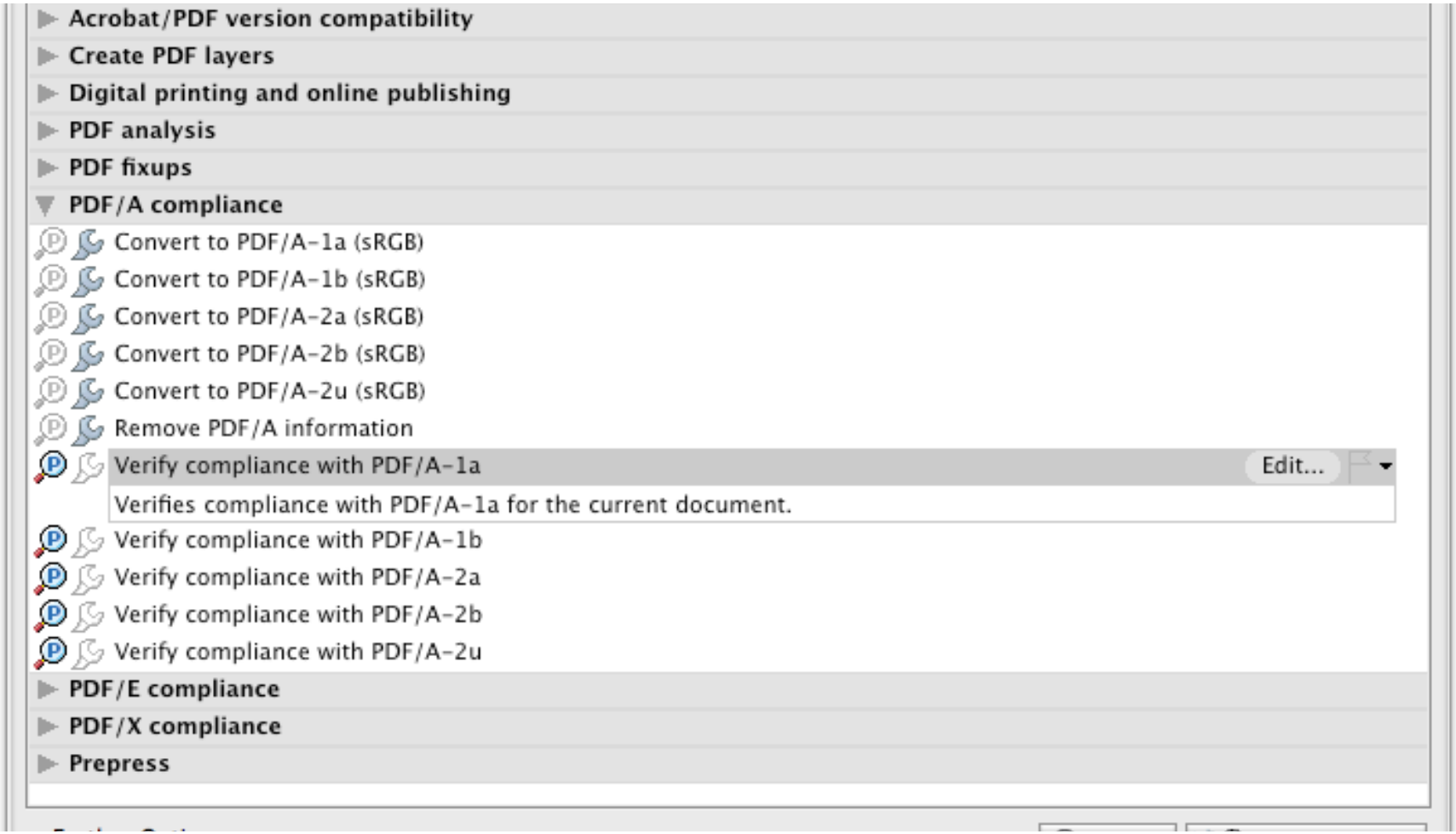
Format identification (target): "Which" PDF/A? (reprise)

- Further complication – scope of validation testing:
  - Pronom (via DROID, JHOVE)
  - PDFBox (Preflight)
  - Adobe (Preflight; Reader 'PDF/A' notation)
  - PDF/A Manager
- Implies tool selection also a critical stage
  - Because of variability of results, compliance testing within conversion tool may be more stable, will bias to single tool's "interpretation"

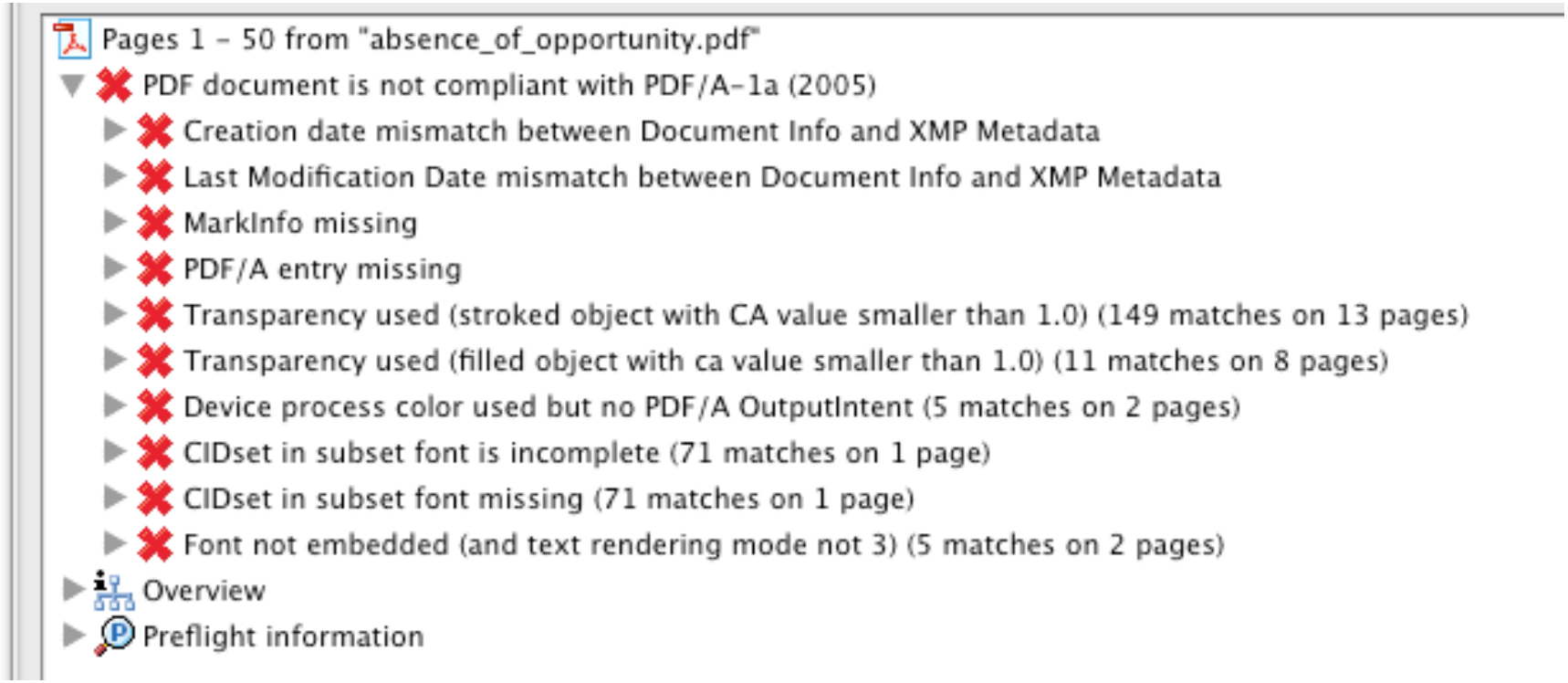
Adobe Reader “Blue bar”



Adobe Preflight



Adobe Preflight Results



PDF/A Manager

```
system@puppy:~/or2013samples$ pdfa --level 1A ./absence_of_opportunity.pdf --noxml
Processing...
Processing...
/home/system/or2013samples/./absence_of_opportunity.pdf
VLD-[FAIL]: absence_of_opportunity.pdf
        - e_PDFA14: Contains compressed object streams
          Obj Refs:498, 499, 500, 501, 502, 503, 504, 505, 514
        - e_PDFA15: Contains cross-reference streams
        - e_PDFA45: Transparency used ('CA' value is not 1.0)
          Obj Refs:263, 274, 278, 281, 291, 296, 298, 313, 319, 320
        - e_PDFA46: Transparency used ('ca' value is not 1.0)
          Obj Refs:201, 263, 264, 274, 281, 296, 298, 319
        - e_PDFA341: The font is not embedded
          Obj Refs:5
        - e_PDFA354: The font descriptor dictionary does not include a CIDSet stream for CIDFont subset
          Obj Refs:306
        - e_PDFA361: Widths in embedded font are inconsistent with /Widths entry in the font dictionary
          Obj Refs:5
        - e_PDFA723: The XMP Metadata stream is not valid
          Obj Refs:2
        - e_PDFA737: Document information entry 'CreationDate' not synchronized with XMP
          Obj Refs:2
        - e_PDFA738: Document information entry 'ModDate' not synchronized with XMP
          Obj Refs:2
        - e_PDFA822: The PDF is not marked as Tagged PDF
        - e_PDFA2331: Device-specific color space used, but no GTS_PDFA1 OutputIntent
          Obj Refs:419, 494, 495, 512, 513
        - e_PDFA8332: Each structure element dictionary in the structure hierarchy must have a Type entry wi
system@puppy:~/or2013samples$
```

Which PDF/A to use?

- Depends on:
  - End goals
  - Also:
    - Source material
    - Original PDF (/A) creation method
- Guide by PDF Association (PDFA in a Nutshell):
  - http://www.pdfa.org/wp-content/uploads/2013/04/PDFA_in_a_Nutshell_21.pdf

Our current target: PDF/A-2B

- Allows transparency & layers
  - Noted in many of our source materials
- Profile appears to "agree" with scanned documents better than higher levels
- PDF/A-1 files can be embedded, however, not a key concern for our needs
- PDF/A-3 appears to introduce contradictory complication: adding non-managed, embedded documents

Original

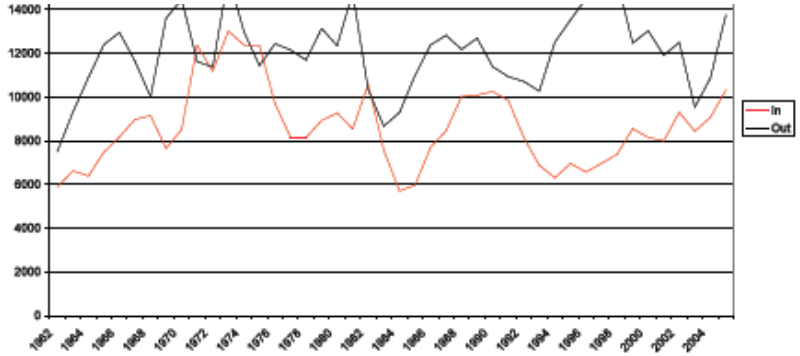
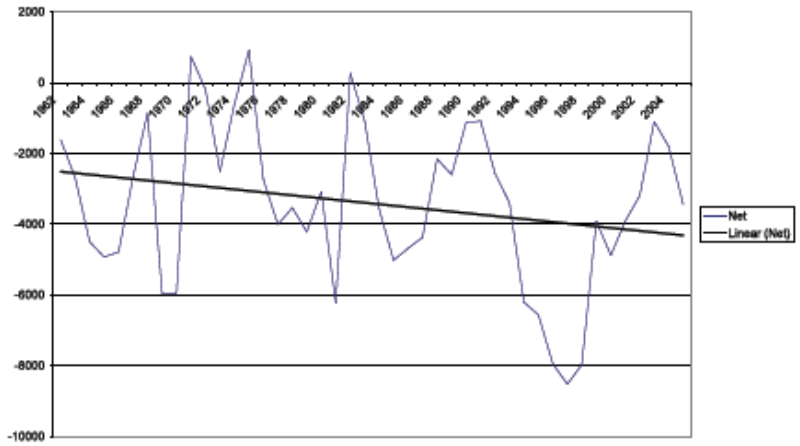


Figure 2
Net Out-Migration Newfoundland and Labrador
1962-2005



It is important to understand where people are moving to and where they are coming from. Figures 3 through 6, illustrate for selected years, the distribution of migration flows by

PDF/A-1A

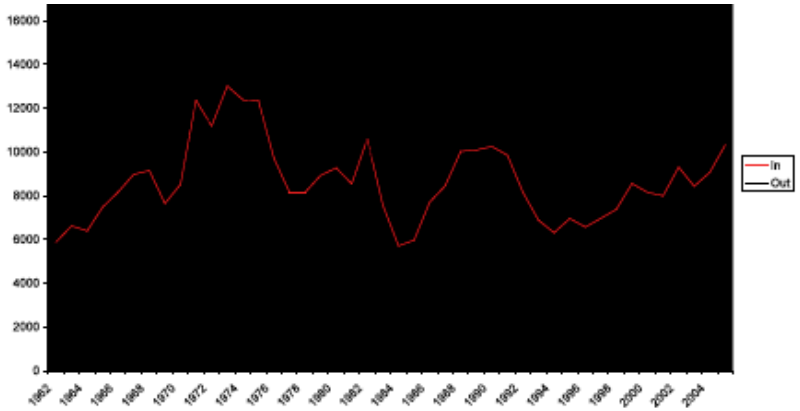
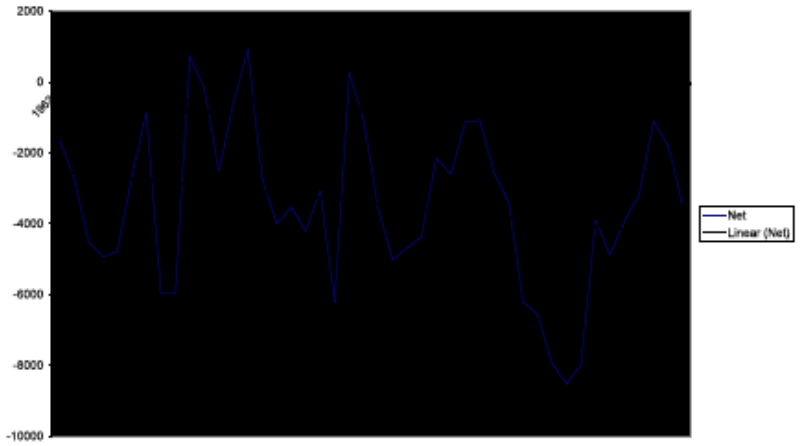


Figure 2
Net Out-Migration Newfoundland and Labrador
1962-2005



PDF/A-2B

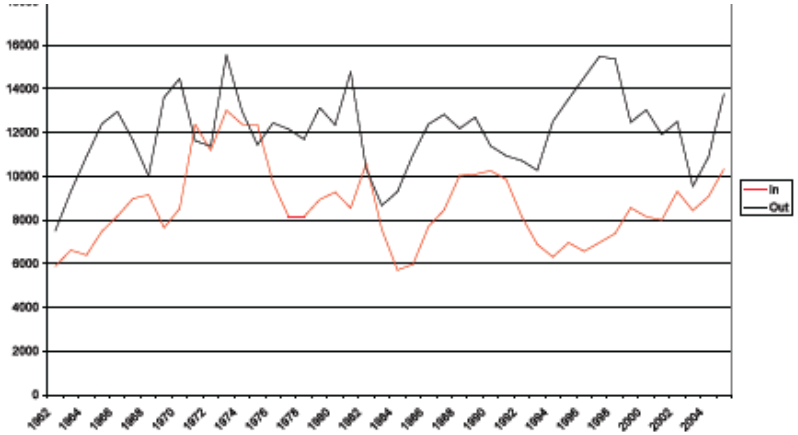
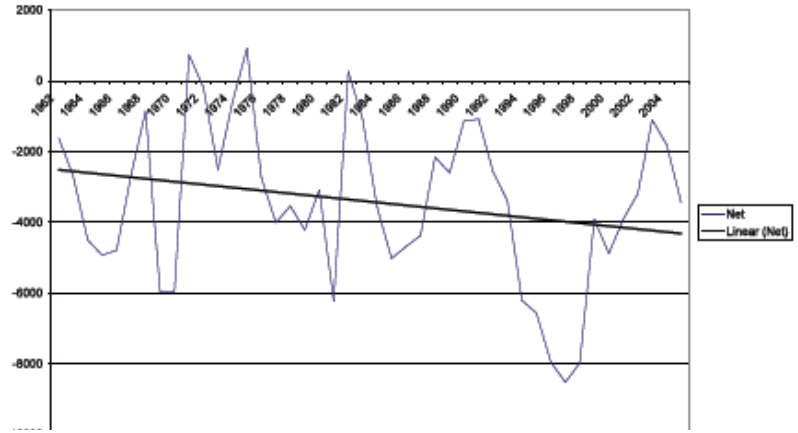


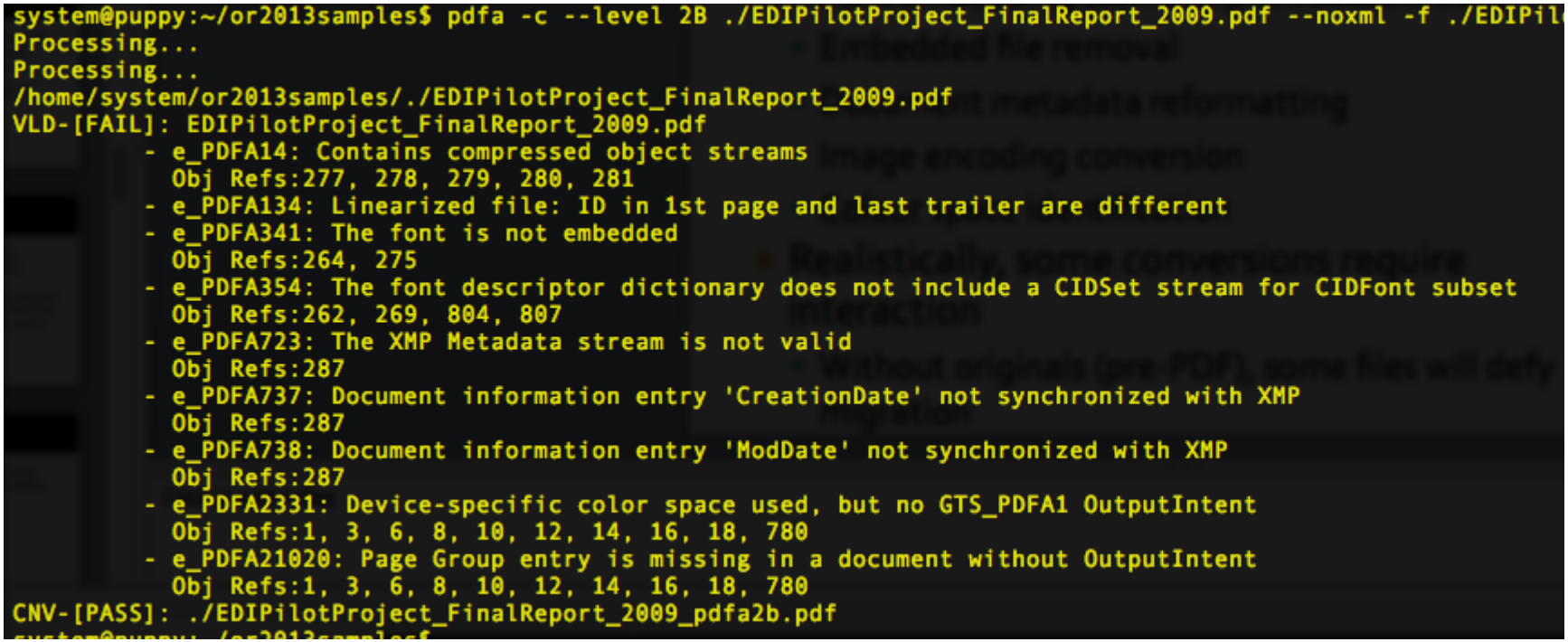
Figure 2
Net Out-Migration Newfoundland and Labrador
1962-2005



Format Migration

- Ideally, migration is a non-interactive step
  - Font embedding (implicit substitution)
  - Embedded file removal
  - Document metadata reformatting
  - Image encoding conversion
  - Colour space identification
- Realistically, some conversions require interaction
  - Without originals (pre-PDF), some files will defy migration to some compliance levels

PDF/A Manager Migration



Format Migration (Failed migration)

- Best we can hope for is some feedback on how to resolve issues when they arise
  - Notes on specific compliance failures, e.g.:
    - Conflicting metadata (usually dates)
    - Missing data (colour space)
- Prefer tools which can provide this info

PDF/A Manager Failed Migration

```
system@puppy:~/or2013samples$ pdfa -c --level 2B ./absence_of_opportunity.pdf --noxml -f ./absence_of_oppor
Processing...
Processing...
/home/system/or2013samples/./absence_of_opportunity.pdf
VLD-[FAIL]: absence_of_opportunity.pdf
        - e_PDFA14: Contains compressed object streams
          Obj Refs:498, 499, 500, 501, 502, 503, 504, 505, 514
        - e_PDFA15: Contains cross-reference streams
        - e_PDFA341: The font is not embedded
          Obj Refs:5
        - e_PDFA354: The font descriptor dictionary does not include a CIDSet stream for CIDFont subset
          Obj Refs:306
        - e_PDFA723: The XMP Metadata stream is not valid
          Obj Refs:2
        - e_PDFA737: Document information entry 'CreationDate' not synchronized with XMP
          Obj Refs:2
        - e_PDFA738: Document information entry 'ModDate' not synchronized with XMP
          Obj Refs:2
        - e_PDFA2331: Device-specific color space used, but no GTS_PDFA1 OutputIntent
          Obj Refs:419, 494, 495, 512, 513
        - e_PDFA21020: Page Group entry is missing in a document without OutputIntent
          Obj Refs:317, 318, 319, 320, 334, 338, 339, 342, 343, 344
        - e_PDFA24222: tintTransform is different in Separations with the same colorant name
          Obj Refs:365, 392, 407
CNV-[FAIL]: ./absence_of_opportunity_pdfa2b.pdf
        - e_PDFA24222: tintTransform is different in Separations with the same colorant name
          Obj Refs:365, 392, 407
```

Tool selection criteria

- Ability to perform level of validation identified as necessary to meet goals
- "Automate-ability"
  - GUI only tools usually suggest manual interaction (scripting can be an option, resource depending)
  - Tools with API or CLI access generally easier to automate

Our current tool: PDFTron PDF/A Manager

- Performs validation and migration steps
- Has command line interface, API available
- Windows / Linux distributions

* Workflow *

- Desire automation as much as practicable
  - Limited staff for manual identification, conversion, validation
- Integration with repository
- Necessarily some manual QA expected

Workflow Stages

- Automated
  - Rename incoming files (format independant)
  - Identify candidate PDF items
    - Reject previously converted
  - Check for PDF/A Compliance
    - Attempt migration of non-compliant items
  - Store details on (un*) successful migrations
- Manual
  - Vet/review results

File Naming Conventions

- Drop / replace characters which have platform dependent meaning
  - E.g. Mac filenames are exceptionally permissive, Windows less so (see '?')
- Log changes made to facilitate later audit
- In our repository, performed on ingest

Identify candidate PDFs

- Select all PDF files as base target pool
- Check for prior preservation (existing migrated, related files)
  - If yes, abort
- Perform PDF/A validation testing to selected compliance level (2B) using selected tool
  - Parse response, if valid, abort (**)

Perform Migration Action

- Attempt PDF/A conversion to selected compliance level (2B) using selected tool
- Perform validation on resulting file
  - Success:
    - Commit conversion
    - Rename to indicate pdfa conversion
  - Failure:
    - Delete attempt
    - Note as failed conversion, feedback, logging (**)

Store Results

- Annotate changes in item history
- Item automatically entered into backup / archive cycle
- Item automatically subject to checksumming within backing store

Room for improvement

- (*) Automate reporting of failed migrations
- (**) Add metadata notation to indicate previously reviewed (but not migrated) files
- Re-write to implement FOSS solutions as practicable
- Reduce manual vetting required
- Adapt for direct to PDF/A from alternate formats
  - Could permit higher level of compliance owing to greater metadata availability in source

* Implementation *

- Eprints 3.3.7
- Ubuntu Linux 10.04
- Perl script using Eprints API
  - Eprint_PDF_Preservation.pl
  - Heavily borrows from Eprints Preservation package circa 3.3.7 (Dave Tarrant; Tim Brody)
- Cron ("regularly scheduled task"), weekly

Eprint_PDF_Preservation.pl (Format Identification)

- Iterates over document objects within each eprint to locate PDF (mimetype 'application/PDF') documents
- Tests for isMigratedVersionOf relation, skips if PDF format (presumes pdfa – only action on PDFs)
- Checks for validation with pdf/a manager
  - Success, skips

Eprint_PDF_Preservation.pl (Preservation Action)

- Attempts conversion under name with suffix _pdfa
- Adds document relations to source (isMigratedVersionOf/hasMigratedVersion)
- Copies document metadata
  - Appends descriptive text ("Migrated (PDFA Conversion) from original format: ")

Post-Preservation Action QA

- Periodic visual review
- Originals left intact, accessible as fall back in case of any undetected issue

Eprints Repository Display

- Preserved document made primary
- Original document referenced as secondary file in case of unanticipated loss in migration
  - Default behavour inherited from original Eprints plugin

Sample Display

The Absence of Opportunity: Understanding the Dynamics of Out-Migration in Newfoundland and Labrador

Lynch, Scott (2007) *The Absence of Opportunity: Understanding the Dynamics of Out-Migration in Newfoundland and Labrador.* Project Report. The Harris Centre.

[English] PDF (Migrated (PDF/A Conversion) from original format: (application/pdf)) - Published Version
Download (4Mb)

[English] PDF - Published Version (Original Version)

Official URL: http://www.mun.ca/harriscentre/reports/arf/2006/Fi...

Abstract

According to the 2006 Census, the population in Canada population increased by 5.4 percent between the years 2001 to 2006. During the same period, the population of Newfoundland and Labrador decreased by 1.5 percent or 7,461 people. The natural component of population growth in Newfoundland and Labrador turned negative in 2005, implying that the death rates per one thousand exceeded the birth rate per one thousand. If present patterns of interprovincial migration continue, the population of Newfoundland and Labrador will continue to decrease. This does not bode well for Newfoundland and Labrador since many government programs are funded on a per

PDF to PDF/A Migration Workflow

- Preservation Goals
  - Why PDF/A, not PDF?
- General workflow considerations
  - Automation
  - Format identification
  - Format migration
  - Tool selection
- Workflow stages
- Implementation

Schedule

- Introduction to Digital Preservation
- Media Type Preservation Planning
- Current State of Preservation Tools
- E-Theses Preservation
- PDF to PDF/A Migration Workflow

  <BREAK>
- Archivematica
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- Q and A

Schedule

- Introduction to Digital Preservation
- Media Type Preservation Planning
- Current State of Preservation Tools
- E-Theses Preservation
- PDF to PDF/A Migration Workflow

  <BREAK>
- **Archivematica**
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- Q and A

Archivematica 0.10-beta

archivematica®

artefactual systems inc.

What is Archivematica?

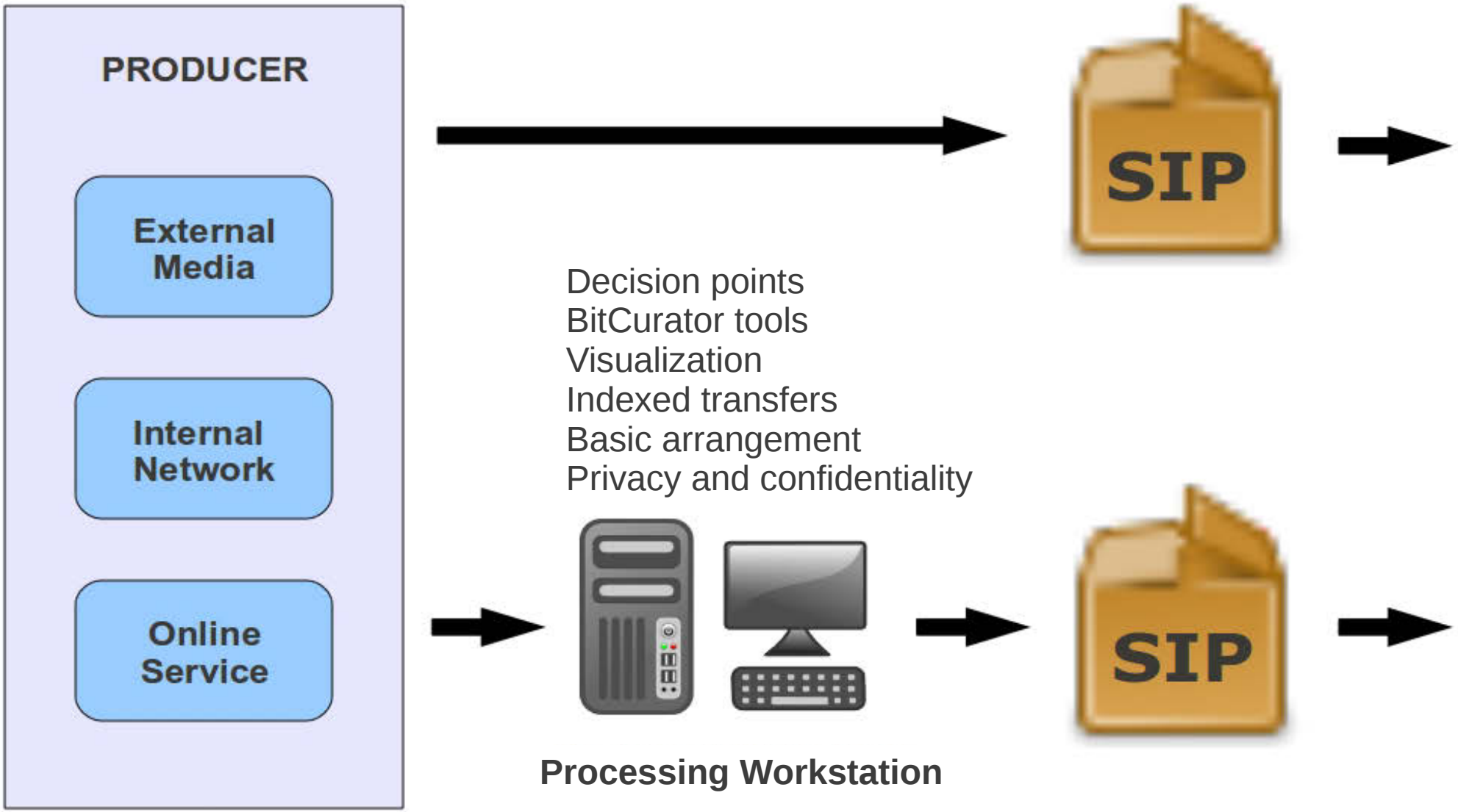
- free and open-source digital preservation system (AGPLv3)
- designed to maintain standards-based, long-term access to collections of digital objects

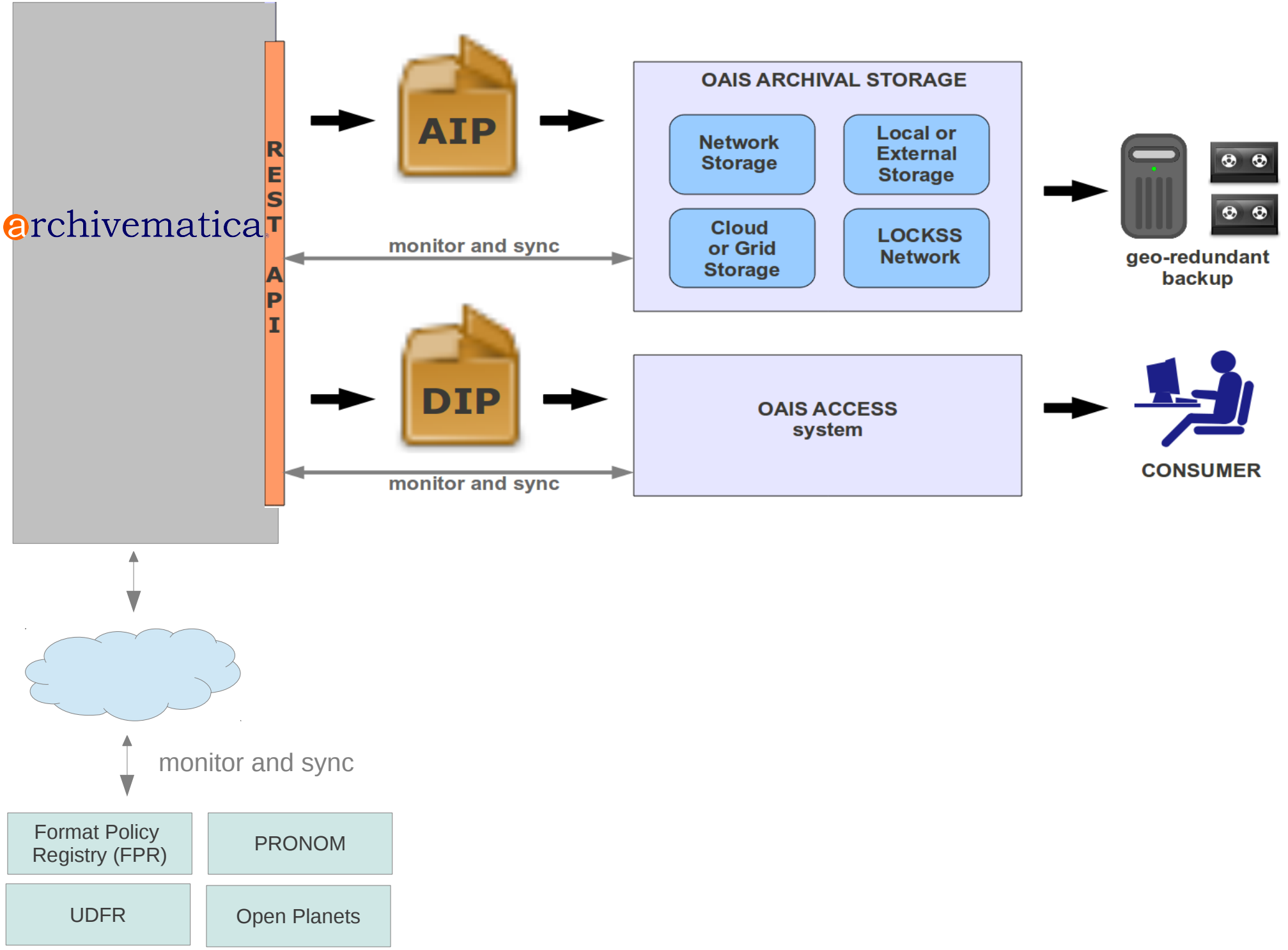
What is Archivematica?

- allows users to process digital objects from ingest to access in conformance with the ISO-OAIS functional model
- Archivematica implements format normalization upon ingest and preserves originals to support emulation and migration strategies

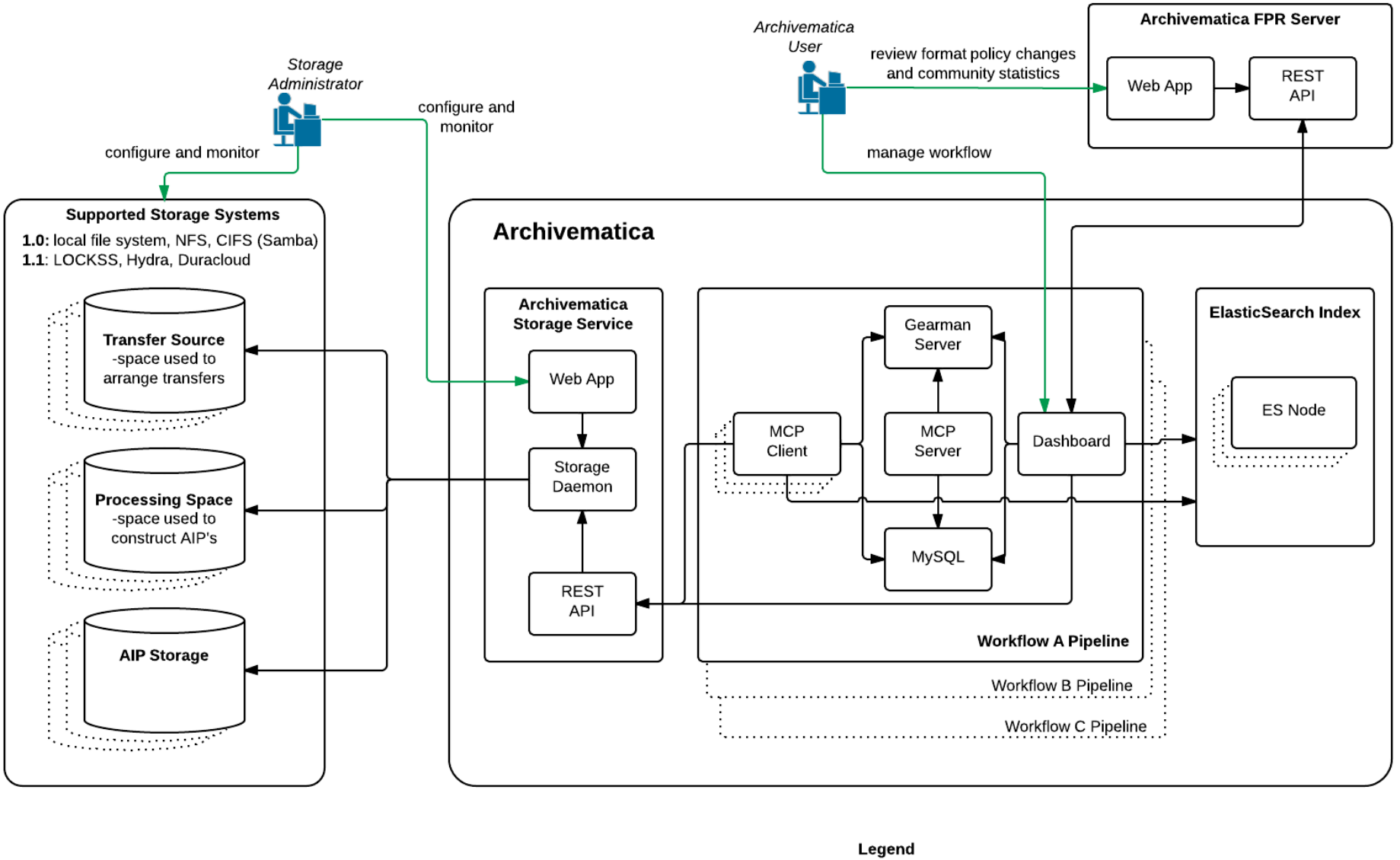
What is Archivematica?

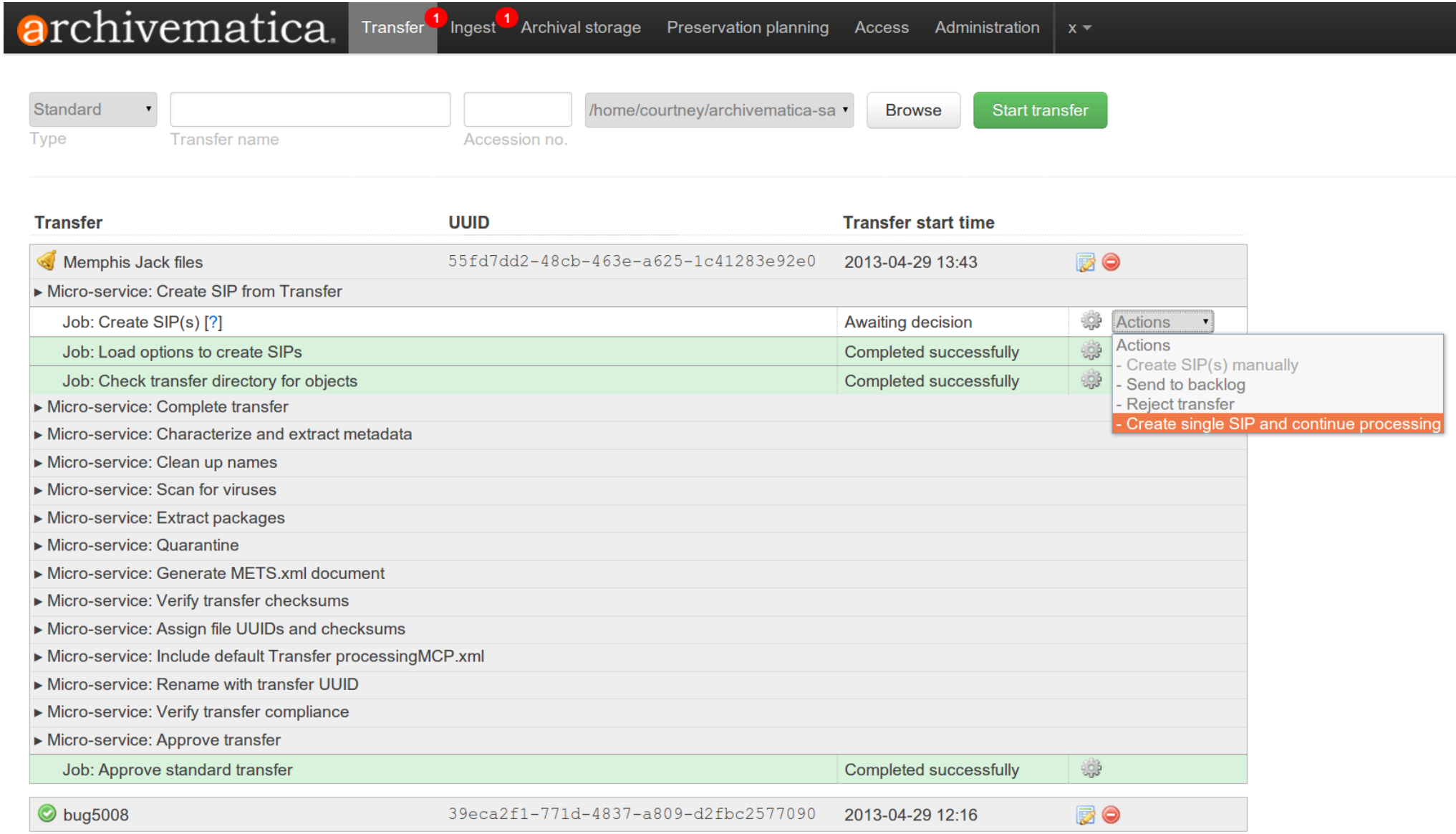
- Archivematica is a processing pipeline
- Archivematica is designed to output high-quality, standards-compliant Archival Information Packages
  - Bagit, METS, PREMIS


PRODUCER
External Media
Internal Network
Online Service
SIP
Decision points
BitCurator tools
Visualization
Indexed transfers
Basic arrangement
Privacy and confidentiality
SIP
Processing Workstation


archivematica
REST API
AIP
DIP
monitor and sync
monitor and sync
OAIS ARCHIVAL STORAGE
Network Storage
Local or External Storage
Cloud or Grid Storage
LOCKSS Network
geo-redundant backup
OAIS ACCESS system
CONSUMER
monitor and sync
Format Policy Registry (FPR)
PRONOM
UDFR
Open Planets

1.0 Architecture




archivematica
Transfer 1
Ingest 1
Archival storage
Preservation planning
Access
Administration
x
Standard
Type
Transfer name
Accession no.
/home/courtney/archivematica-sa
Browse
Start transfer
Transfer
UUID
Transfer start time
Memphis Jack files
55fd7dd2-48cb-463e-a625-1c41283e92e0
2013-04-29 13:43
Micro-service: Create SIP from Transfer
Job: Create SIP(s) [?]
Awaiting decision
Actions
Actions
- Create SIP(s) manually
- Send to backlog
- Reject transfer
- Create single SIP and continue processing
Job: Load options to create SIPs
Completed successfully
Job: Check transfer directory for objects
Completed successfully
Micro-service: Complete transfer
Micro-service: Characterize and extract metadata
Micro-service: Clean up names
Micro-service: Scan for viruses
Micro-service: Extract packages
Micro-service: Quarantine
Micro-service: Generate METS.xml document
Micro-service: Verify transfer checksums
Micro-service: Assign file UUIDs and checksums
Micro-service: Include default Transfer processingMCP.xml
Micro-service: Rename with transfer UUID
Micro-service: Verify transfer compliance
Micro-service: Approve transfer
Job: Approve standard transfer
Completed successfully
bug5008
39eca2f1-771d-4837-a809-d2fbc2577090
2013-04-29 12:16

atom
access to memory

archivematica®

artefactual systems inc.

The METS file

<dmdSec> (descriptive metadata)

Dublin Core XML

<amdSec> (administrative metadata)

<techMD>

PREMIS: object

<digiProvMD>

PREMIS: events

PREMIS: agents

<rightsMD>

PREMIS: rights

<fileSec> (a list of the files and their roles and relationships)

<structMap> (a representation of the physical structure of the AIP)

Preservation planning

- A two-pronged approach:
  - **Normalization** on ingest
  - Preservation of the original file to support future strategies such as **migration** and **emulation**
- Normalization relies on *format policies* based on an analysis of the significant characteristics of file formats
    - A format policy indicates the actions, tools and settings to apply to a file of a particular file format (e.g. normalization to preservation and/or access format)

Media type	File formats	Preservation format(s)	Access format(s)	Normalization tool
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PST	MBOX	MBOX	readpst
Email	Maildir**	Original format	MBOX	md2mb.py
Office Open XML	DOCX, PPTX, XLSX	Original format	PDF for PPTX	OpenOffice
Plain text	TXT	Original format	Original format	None
Portable Document Format	PDF	PDF/A	Original format	Ghostscript
Presentation files	PPT	Original format	PDF	OpenOffice
Raster images	BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA	Uncompressed TIFF	JPEG	ImageMagick
Raw camera files/Digital Negative format**	3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F	Original format	JPEG	ImageMagick/UFRaw
Spreadsheets	XLS	Original format	Original format	None
Vector images	AI, EPS, SVG	SVG	PDF	Inkscape
Video	AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV	FFV1/LPCM in MKV	MPEG-1	FFmpeg
Word processing files	DOC, WPD, RTF	• ODF (WPD and RTF) • Original format (DOC)	PDF	OpenOffice

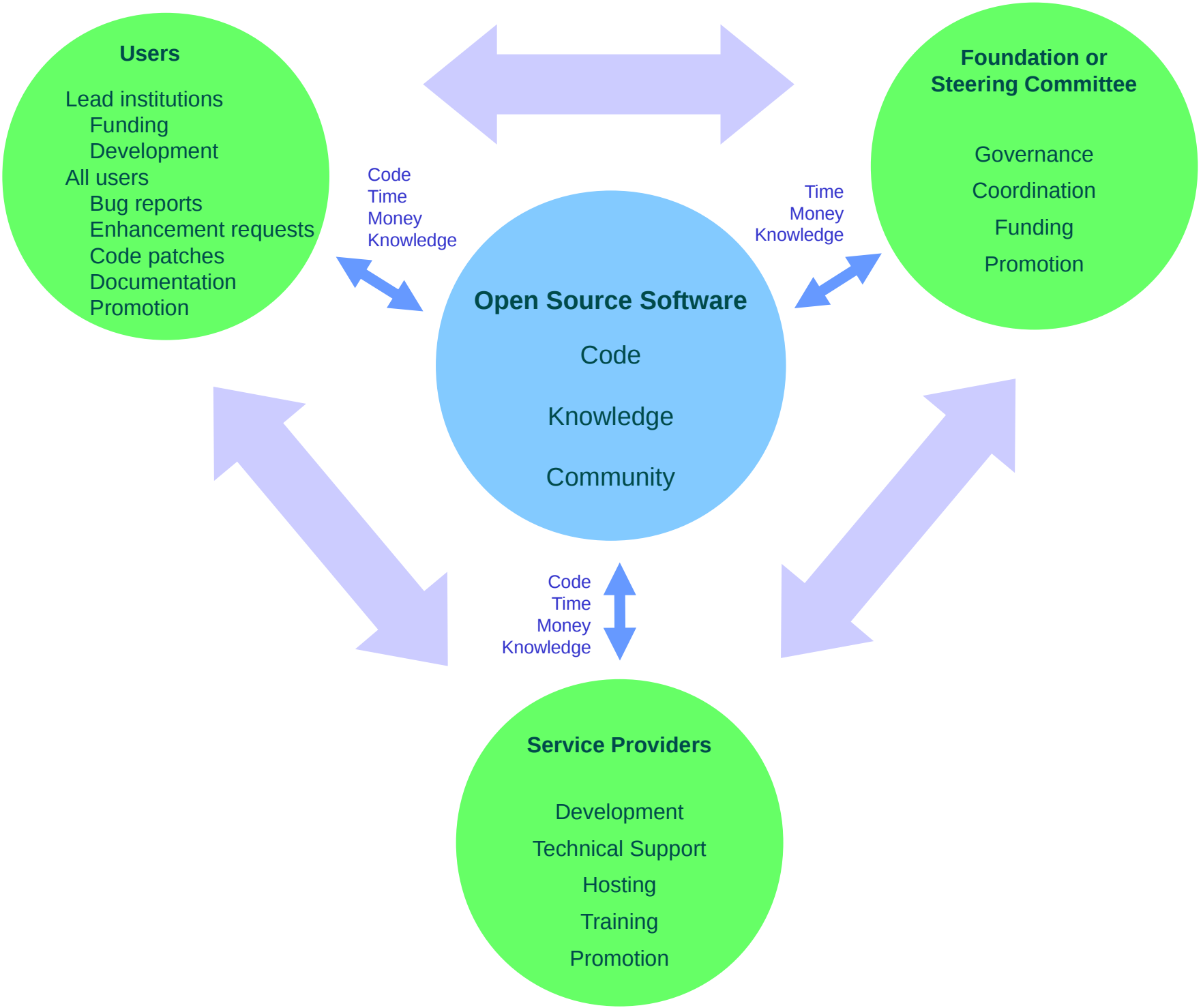
https://www.archivematica.org/preservation

Archivematica format policies

- Criteria for selecting default formats:
  - Non-proprietary
  - Freely available specifications
  - Widely used/endorsed by major repositories
  - No compression/lossless compression
  - Tools available to write and render the format
- Format policies will change as community standards, practices and tools evolve.

Agile development model

- Release early, release often
  - Feb 2009: Release 0.1-alpha
  - June 2011: Release 0.7.1-alpha
  - January 2012: Release 0.8-alpha
  - August 2012: Release 0.9-beta
  - April 2013: Release 0.10-beta
- Each iteration leads to updated and improved:
  - Requirements
  - Software
  - Documentation
  - Development resources


Users
Lead institutions
Funding
Development
All users
Bug reports
Enhancement requests
Code patches
Documentation
Promotion
Code
Time
Money
Knowledge
Open Source Software
Code
Knowledge
Community
Time
Money
Knowledge
Foundation or
Steering Committee
Governance
Coordination
Funding
Promotion
Code
Time
Money
Knowledge
Service Providers
Development
Technical Support
Hosting
Training
Promotion

Public wiki – archivematica.org

archivematica®

184.69.130.182 Talk for this IP address Log in / create account

Navigation

Main page
Recent changes
Random page

Toolbox

What links here
Related changes
Special pages
Printable version
Permanent link

Page Discussion | Read View source View history | Go Search

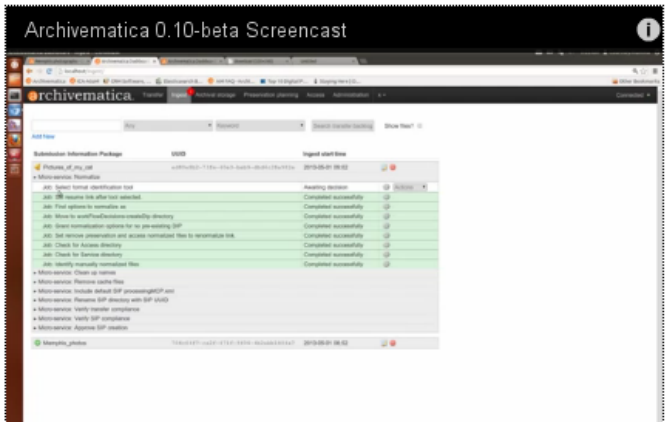
Main Page

What is Archivematica?

Archivematica is a free and open-source digital preservation system that is designed to maintain standards-based, long-term access to collections of digital objects.

Archivematica uses a micro-services design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. Users monitor and control the micro-services via a web-based dashboard. Archivematica uses METS, PREMIS, Dublin Core and other best practice metadata standards. Archivematica implements format policies based on an analysis of the significant characteristics of file formats.

The overview section provides a detailed description of Archivematica's functionality and technical architecture. This Archivematica 0.10-beta screencast gives a demo of the core features in the current release.



Download
Release 0.10-beta

- Release notes

Documentation

- User Manual

Developer resources

- Development roadmap
- Issues list

Community support

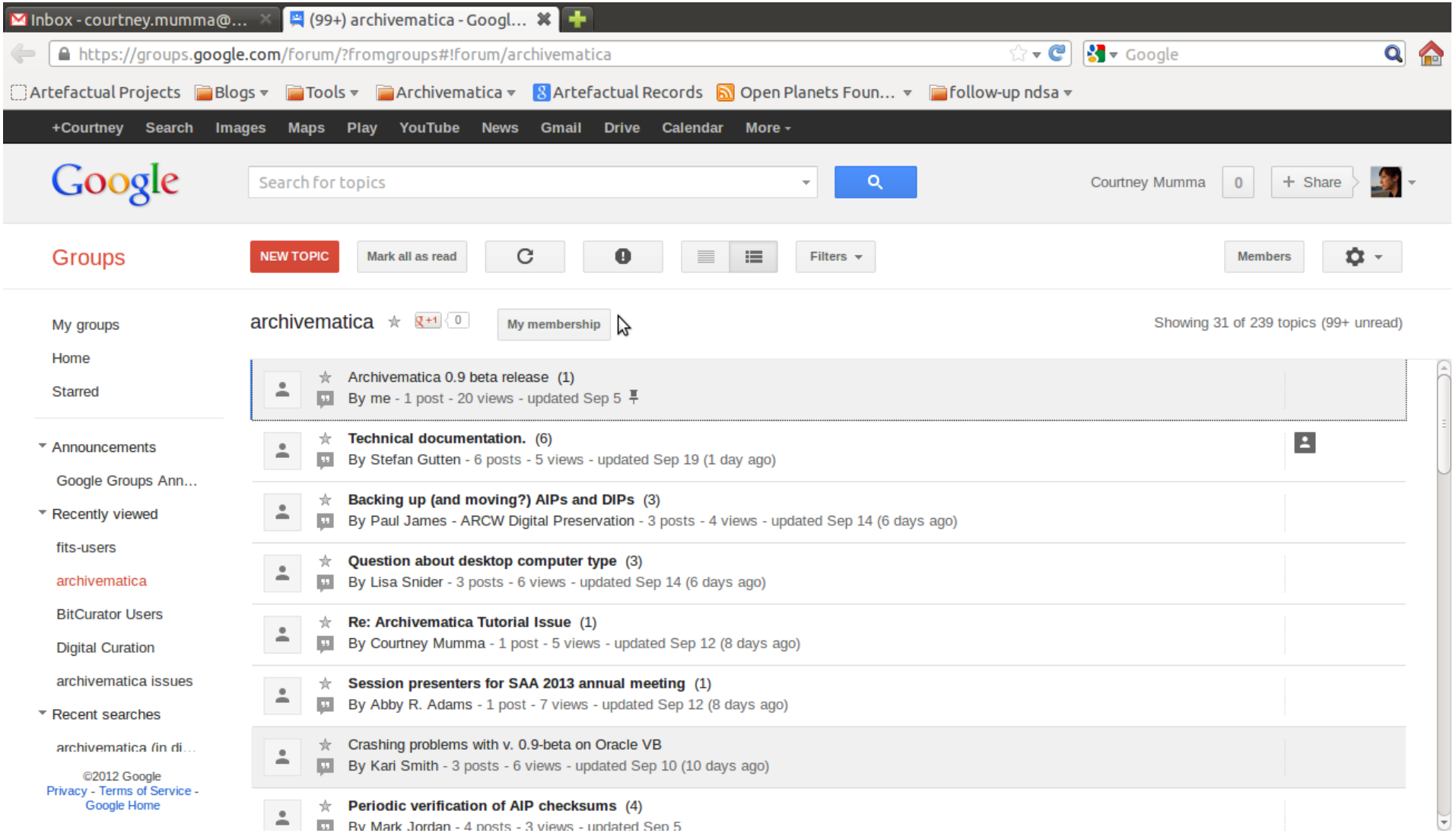
- Discussion list
- Community

Twitter

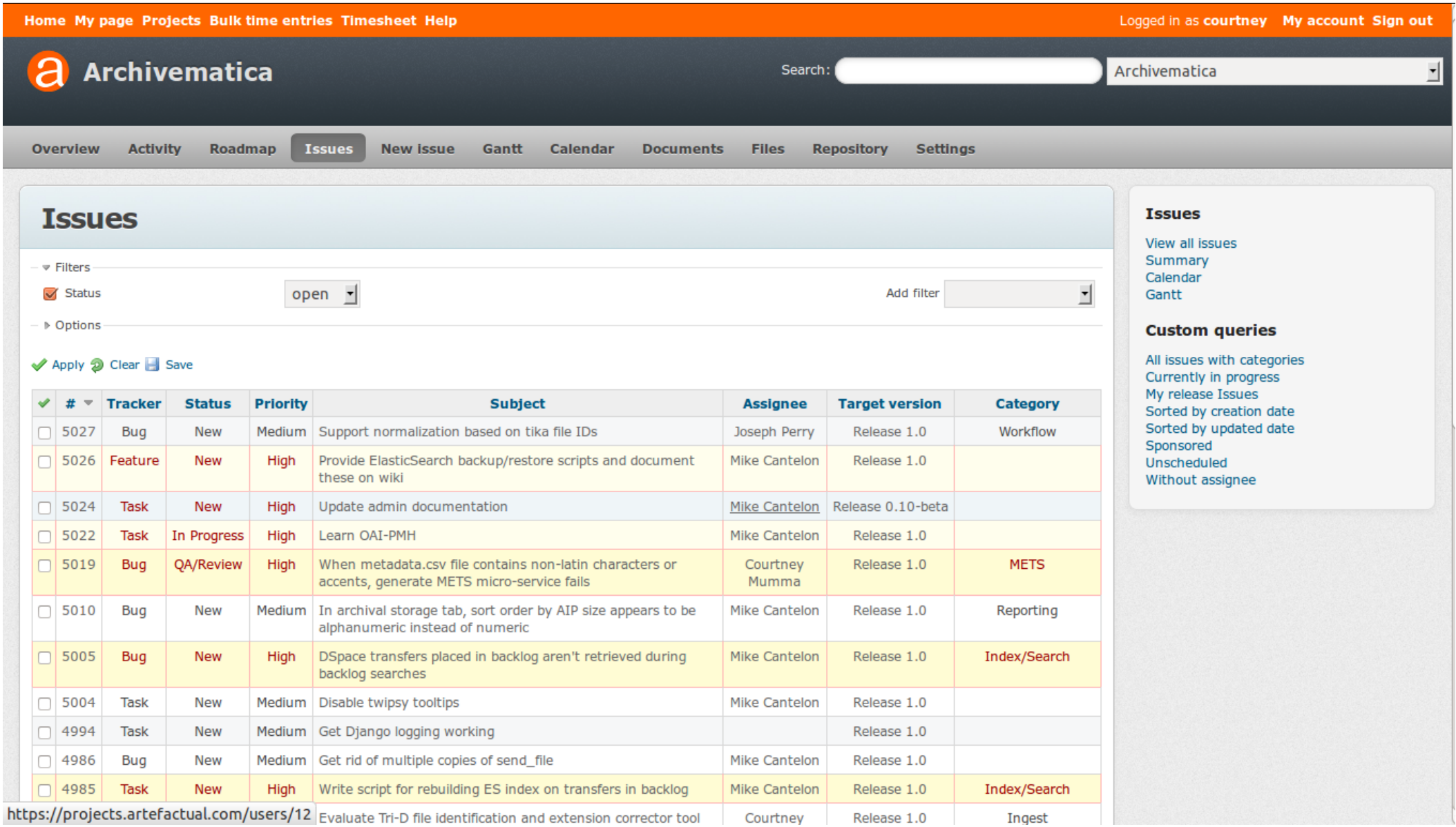
What people are saying about
Archivematica

pjvangarderen RT @archivematica: 1st production AIP stored at dev partners @ubclibrary! Thanks for helping make digital preservation open & accessible
4 minutes ago · reply · retweet · favorite

User list


Inbox - courtney.mumma@...
(99+) archivematica - Googl...
https://groups.google.com/forum/?fromgroups#!forum/archivematica
Google
Artefactual Projects
Blogs
Tools
Archivematica
Artefactual Records
Open Planets Foun...
follow-up ndsa
+Courtney
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
Search for topics
Courtney Mumma
0
+ Share
Groups
NEW TOPIC
Mark all as read
Filters
Members
My groups
Home
Starred
Announcements
Google Groups Ann...
Recently viewed
fits-users
archivematica
BitCurator Users
Digital Curation
archivematica issues
Recent searches
archivematica (in di...
©2012 Google
Privacy - Terms of Service - Google Home
archivematica
My membership
Showing 31 of 239 topics (99+ unread)
Archivematica 0.9 beta release (1)
By me - 1 post - 20 views - updated Sep 5
Technical documentation. (6)
By Stefan Gutten - 6 posts - 5 views - updated Sep 19 (1 day ago)
Backing up (and moving?) AIPs and DIPs (3)
By Paul James - ARCW Digital Preservation - 3 posts - 4 views - updated Sep 14 (6 days ago)
Question about desktop computer type (3)
By Lisa Snider - 3 posts - 6 views - updated Sep 14 (6 days ago)
Re: Archivematica Tutorial Issue (1)
By Courtney Mumma - 1 post - 5 views - updated Sep 12 (8 days ago)
Session presenters for SAA 2013 annual meeting (1)
By Abby R. Adams - 1 post - 7 views - updated Sep 12 (8 days ago)
Crashing problems with v. 0.9-beta on Oracle VB
By Kari Smith - 3 posts - 6 views - updated Sep 10 (10 days ago)
Periodic verification of AIP checksums (4)
By Mark Jordan - 4 posts - 3 views - updated Sep 5

Issue list


Home My page Projects Bulk time entries Timesheet Help
Logged in as courtney My account Sign out
Archivematica
Search:
Archivematica
Overview Activity Roadmap Issues New issue Gantt Calendar Documents Files Repository Settings
Issues
Filters
Status
open
Add filter
Options
Apply Clear Save
Tracker Status Priority Subject Assignee Target version Category
5027 Bug New Medium Support normalization based on tika file IDs Joseph Perry Release 1.0 Workflow
5026 Feature New High Provide ElasticSearch backup/restore scripts and document these on wiki Mike Cantelon Release 1.0
5024 Task New High Update admin documentation Mike Cantelon Release 0.10-beta
5022 Task In Progress High Learn OAI-PMH Mike Cantelon Release 1.0
5019 Bug QA/Review High When metadata.csv file contains non-latin characters or accents, generate METS micro-service fails Courtney Mumma Release 1.0 METS
5010 Bug New Medium In archival storage tab, sort order by AIP size appears to be alphanumeric instead of numeric Mike Cantelon Release 1.0 Reporting
5005 Bug New High DSpace transfers placed in backlog aren't retrieved during backlog searches Mike Cantelon Release 1.0 Index/Search
5004 Task New Medium Disable twipsy tooltips Mike Cantelon Release 1.0
4994 Task New Medium Get Django logging working Release 1.0
4986 Bug New Medium Get rid of multiple copies of send_file Mike Cantelon Release 1.0
4985 Task New High Write script for rebuilding ES index on transfers in backlog Mike Cantelon Release 1.0 Index/Search
Evaluate Tri-D file identification and extension corrector tool Courtney Release 1.0 Ingest
Issues
View all issues
Summary
Calendar
Gantt
Custom queries
All issues with categories
Currently in progress
My release Issues
Sorted by creation date
Sorted by updated date
Sponsored
Unscheduled
Without assignee
https://projects.artefactual.com/users/12

http://archivematica.org

archivematica®

Demonstration time!

Schedule

- Introduction to Digital Preservation
- Media Type Preservation Planning
- Current State of Preservation Tools
- E-Theses Preservation
- PDF to PDF/A Migration Workflow

  <BREAK>
- Archivematica
- **The Trappist Method for the Dissemination and Preservation of Digital Objects**
- Q and A

LEARNING RESOURCES

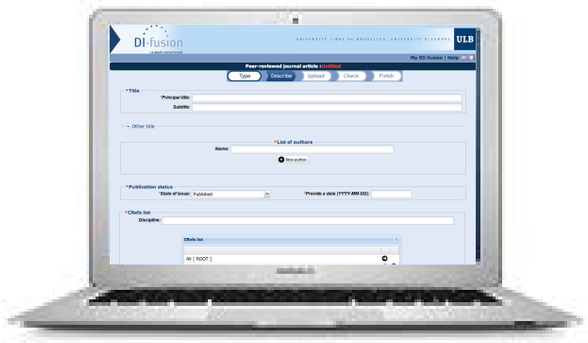
ARCHIVES & BIBLIOTHÈQUES

Dissemination and Preservation of Digital Objects

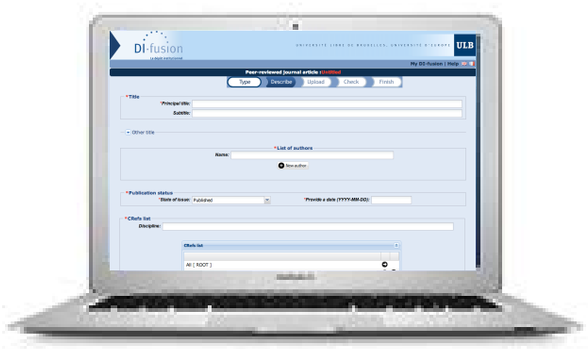
Case study of digitized Ph.D. theses

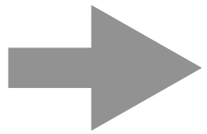
Benoît Pauwels
Anthony Leroy

What our **submission** strategy of digital objects used to be...



What our **submission** strategy of digital objects used to be...

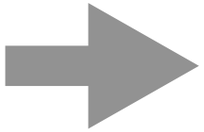




DSPACE

Institutional Repository (dipot)

What our **submission** strategy of digital objects used to be...



DSPACE

Institutional Repository
(dipot)

item {handle}

bundle "Original"

bitstream 1: Abstract

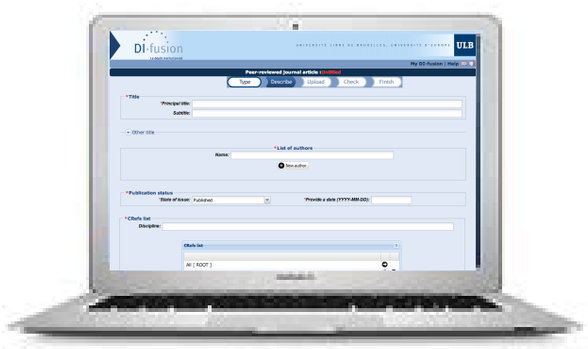
bitstream 2: Chapter 1-3

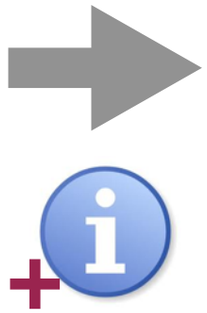
bitstream 3: Chapter 4-7

bitstream 4: Chapter 8-10

bitstream 5: Slides of the PhD Defense

What our **submission** strategy of digital objects used to be...





Institutional Repository (dipot)

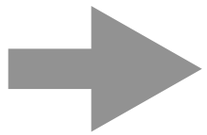
item {handle}

bundle "Original"

- bitstream 1: Abstract
- bitstream 2: Chapter 1-3
- bitstream 3: Chapter 4-7
- bitstream 4: Chapter 8-10
- bitstream 5: Slides of the PhD Defense

Flat objects without file structure or relationship

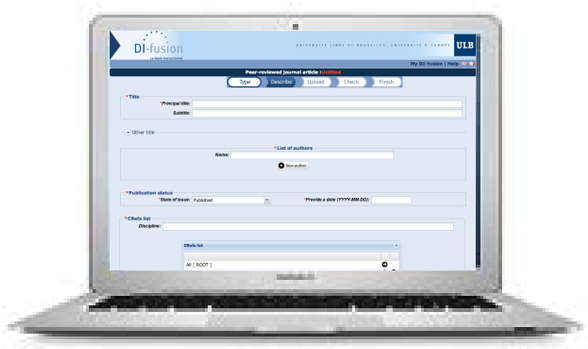
What our **dissemination** strategy of digital objects used to be...

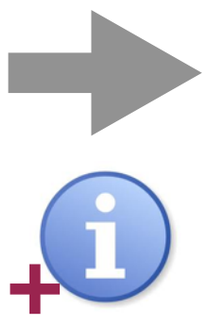


DSPACE

Institutional Repository (dipot)

What our **dissemination** strategy of digital objects used to be...





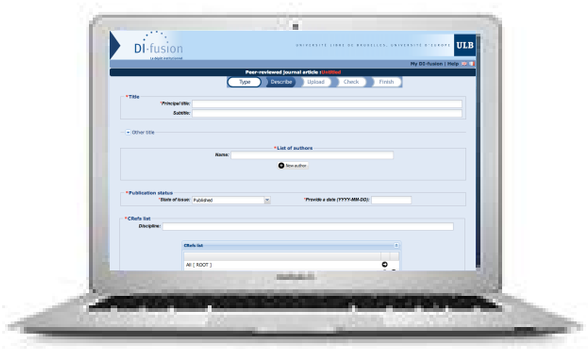
DSPACE

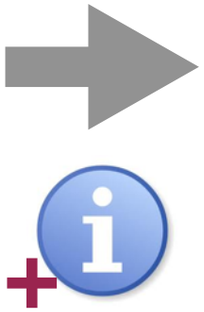
Institutional Repository (dipot)

OAI-PMH

DIDL/MODS

What our **dissemination** strategy of digital objects used to be...





D SPACE

Institutional Repository (dipot)

OAI-PMH

DIDL/MODS

DI-fusion

Dissemination Interface

LEARNING RESOURCES

What our **dissemination** strategy of digital objects used to be...

DI·fusion

UMONS ULB ACADÉMIE UNIVERSITAIRE WALLONIE-BRUXELLES

Recherche avancée Historique de recherche Mon DI-fusion ULB | Mon DI UMONS | Aide |

Passe-partout Recherche

« Retourner aux résultats de recherche Mettre en favoris Citer ShareThis

La culture du vin dans la littérature italienne du Moyen Âge tardif au début des Temps Modernes. Critères de qualité, systèmes de représentation et identités.

par Grappe, Yann
Promoteur Devroey, Jean-Pierre, Montanari, Massimo
Publication Publié, 2009-10-19
Thèse de doctorat

ACCÈS EN LIGNE DÉTAILS CONTENU

Fichier	Version du fichier	Droits d'accès au fichier
YannGrappeIntroductionConclusion.pdf	Pubprint	Public
YannGrappeChap1et2.pdf	Pubprint	Bloqué
YannGrappeChap3.pdf	Pubprint	Bloqué
YannGrappeTableDesMatieres.pdf	Pubprint	Public
YannGrappeBibliographie.pdf	Pubprint	Bloqué

Documents en relation

DI-fusion

Merci intangibili e patrimonio culturale. La costruzione del turismo enogastronomico a Montepulciano. (provincia di Siena, regione Toscana, Italia) Intangibles merchandises and cultural heritage. The construction of the gastronomic tourism at Montepulciano (Siena, Tuscany, Italy)
par Fiorillo, Alessia
Publication Brussels, Universite Libre de Bruxelles, 2010-06-28

"Bon mangeur, mauvais mangeur. Pratiques alimentaires et critique sociale dans l'œuvre de Sidoine Apollinaire et de ses contemporains"
par Raga, Emmanuelle

What our **preservation** strategy of digital objects used to be...

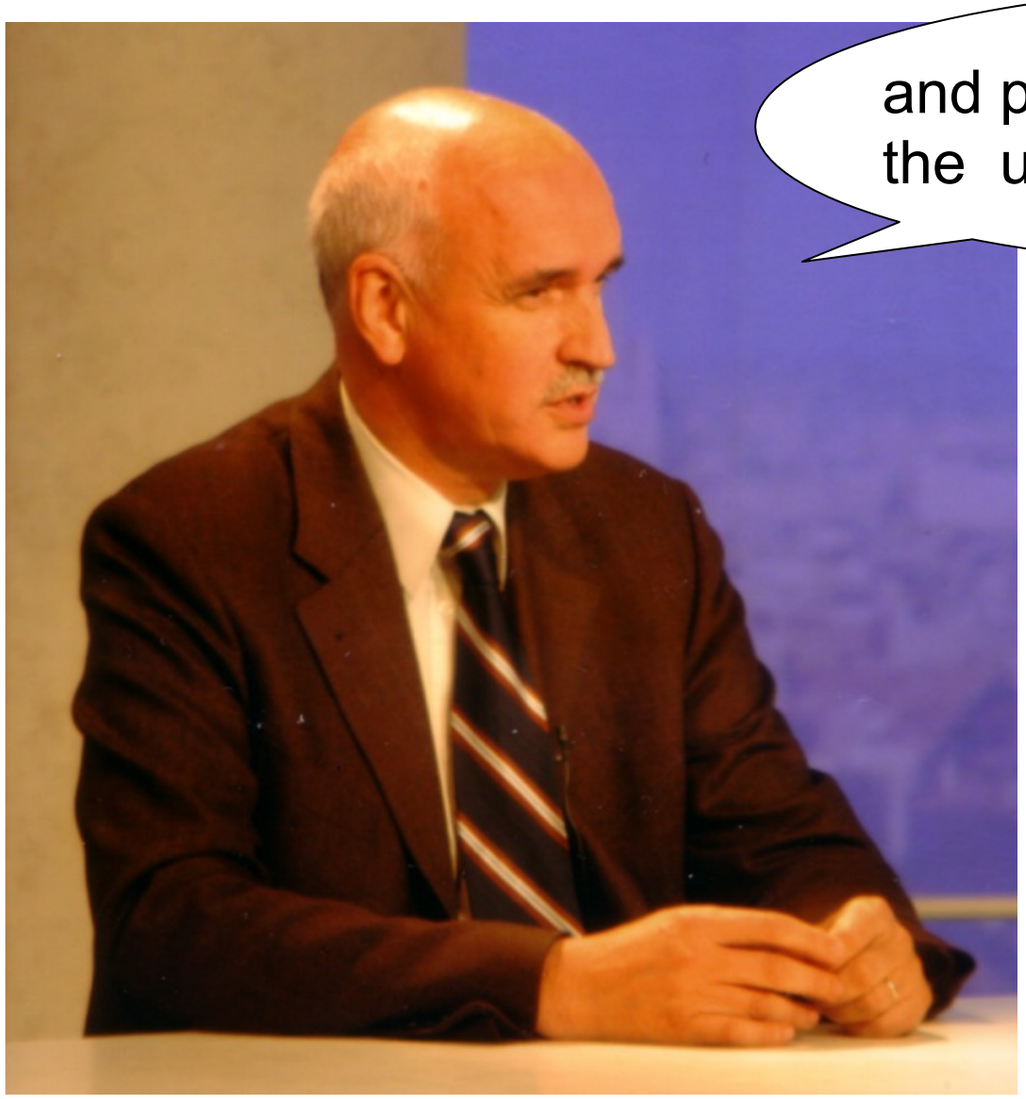
What our **preservation** strategy of digital objects used to be...

What our **preservation** strategy of digital objects used to be...

backup strategy but no preservation plan

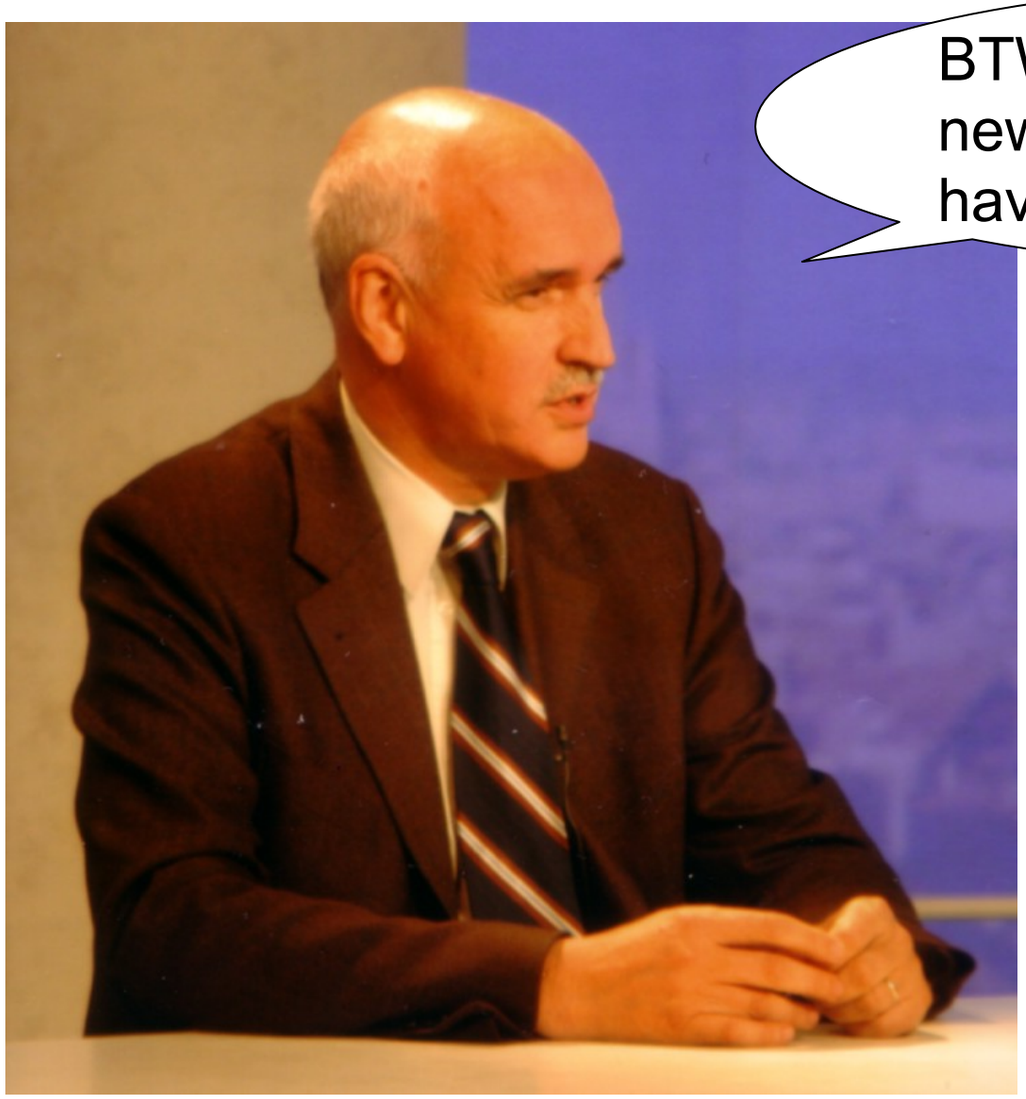
Until one day...

Until one day...

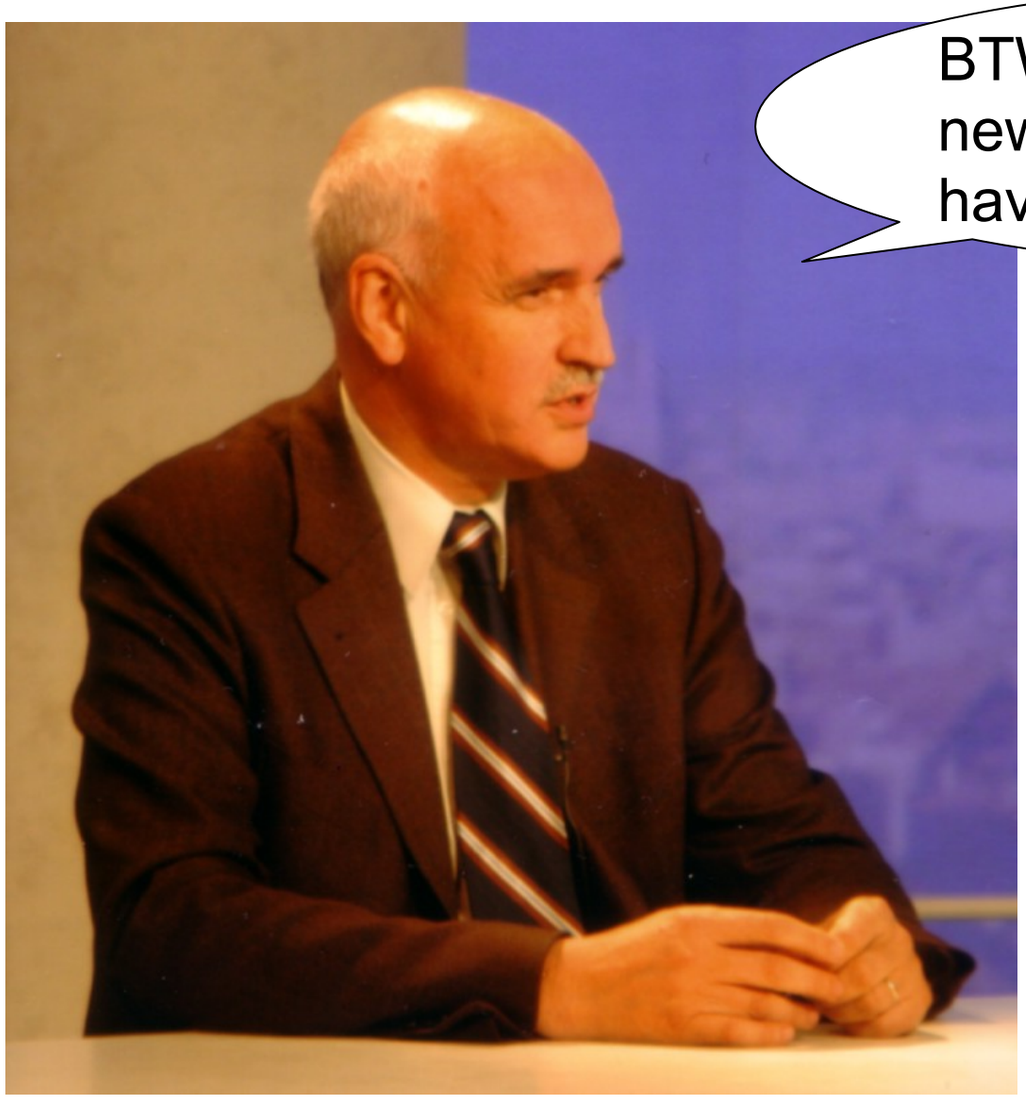


Until one day...

Until one day...



Until one day...



BTW, we need space for the new Learning Center so we have to get rid of the paper

Ok... Perhaps we have to change a few things than...

These changes are not limited to theses...

Digitized Books

45k pages
(per year)

These changes are not limited to theses...

Digitized Books

45k pages
(per year)

Master & PhD Theses & Syllabus

3M pages
(one-shot)

These changes are not limited to theses...

Digitized Books

45k pages
(per year)

Master & PhD Theses & Syllabus

3M pages
(one-shot)

Archives & Varia

??

These changes are not limited to theses...

Digitized Books	**Master & PhD Theses & Syllabus**	**Archives & Varia**	**Born dig**
45k pages (per year)	3M pages (one-shot)	??	20 000 existing objects (+ 15%/year)

These changes are not limited to theses...

Digitized Books	Master & PhD Theses & Syllabus	Archives & Varia	Born dig
45k pages (per year)	3M pages (one-shot)	??	20 000 existing objects (+ 15%/year)

Integrated submission, dissemination and preservation workflow

What has to be changed?

New high-volume digitization workflow

What has to be changed?

New high-volume digitization workflow

New description model for all digital objects

What has to be changed?

New high-volume digitization workflow

New description model for all digital objects

A preservation strategy for all digital objects

What has to be changed?

New high-volume digitization workflow

New description model for all digital objects

A preservation strategy for all digital objects

For the theses project, we need to digitize +10.000 volumes, over 3M pages

Shouldn’t we outsource?
Can we do this ourselves?

Shouldn’t we outsource?
Can we do this ourselves?

- total control of the production
- less expensive
- once in place, the workflow
can be used for other projects

Shouldn’t we outsource?
Can we do this ourselves?

- total control of the production
- less expensive
- once in place, the workflow can be used for other projects

- extra human resources needed
- longer time to start
- taking all responsability

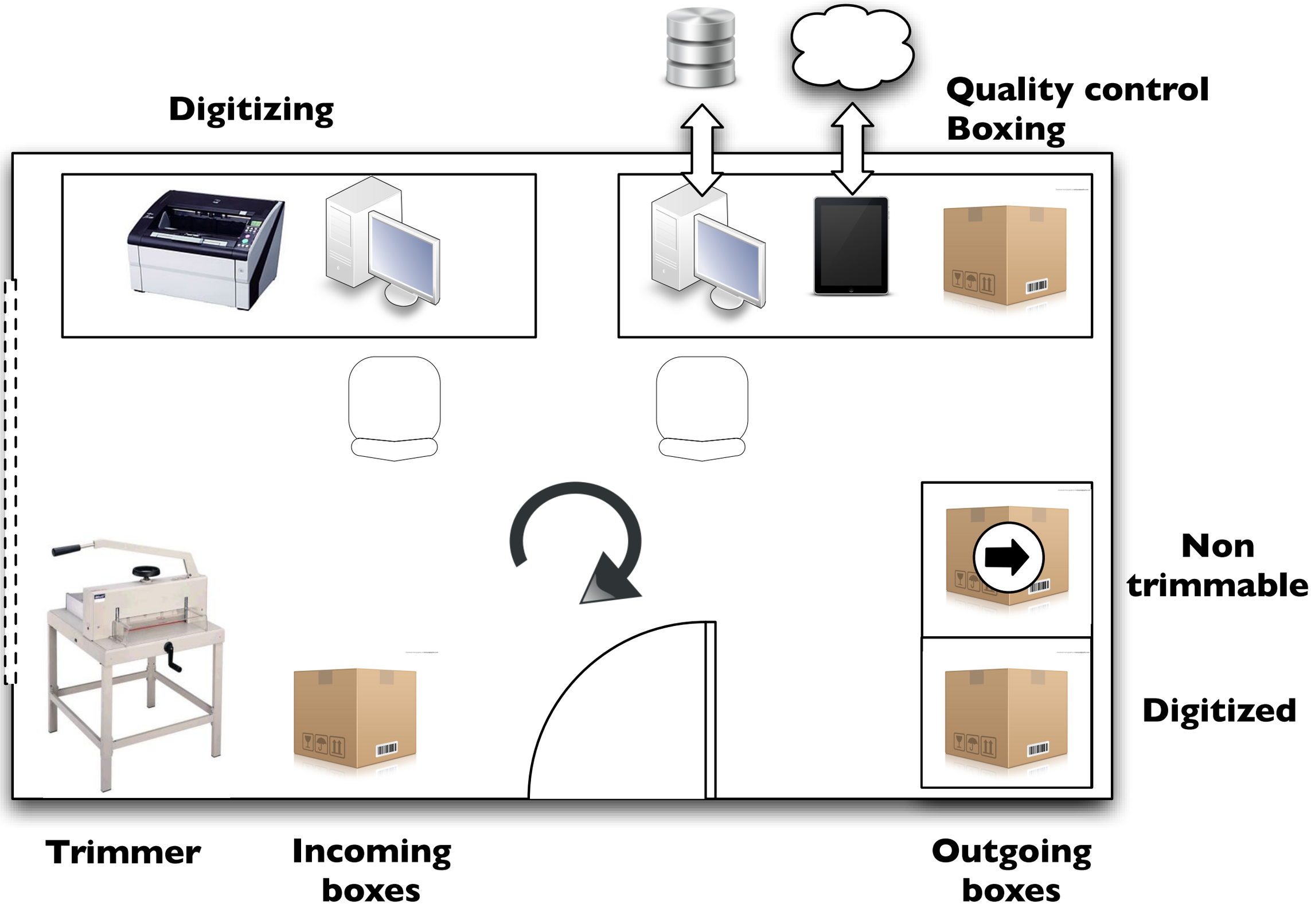
Shouldn’t we outsource?
Can we do this ourselves?

- total control of the production
- less expensive
- once in place, the workflow can be used for other projects

- extra human resources needed
- longer time to start
- taking all responsability

Yes, we can!

A simple digitization workflow



A simple digitization workflow

Our workflow targets zero error tolerance

automatic filenaming with QR code generated from our catalog

Our workflow targets zero error tolerance

automatic filenaming with QR code generated from our catalog

duplex scanning even for simplex print

Our workflow targets zero error tolerance

automatic filenaming with QR code generated from our catalog

duplex scanning even for simplex print

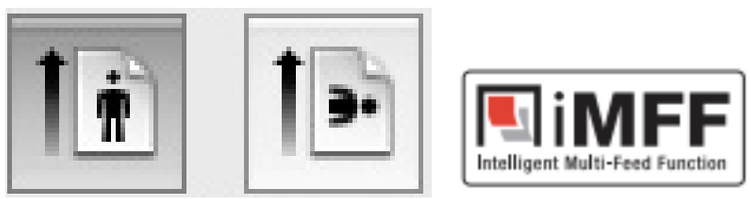
both raw and automatic quality improved images

Our workflow targets zero error tolerance

automatic filenaming with QR code generated from our catalog

duplex scanning even for simplex print

both raw and automatic quality improved images



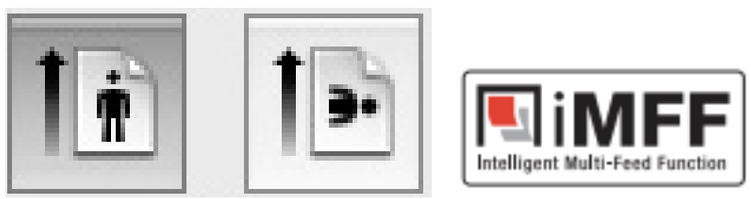
double scan with iMFF: portrait and landscape

Our workflow targets zero error tolerance

automatic filenaming with QR code generated from our catalog

duplex scanning even for simplex print

both raw and automatic quality improved images



double scan with iMFF: portrait and landscape

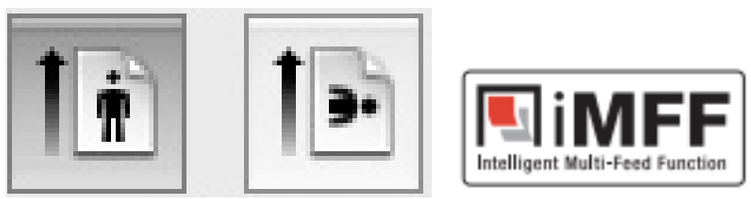
4 page count verifications

Our workflow targets zero error tolerance

automatic filenaming with QR code generated from our catalog

duplex scanning even for simplex print

both raw and automatic quality improved images



double scan with iMFF: portrait and landscape

4 page count verifications

software-assisted quality control process

In this workflow, our high-volume scanner digitizes 250 thesis/week

In this workflow, our high-volume scanner digitizes 250 thesis/week

~ 100 GB of data produced every day

The main problem of raw digitizing is file size

The main problem of raw digitizing is file size

Digitized Ph.D. and Master theses : **60 TB**

The main problem of raw digitizing is file size

Digitized Ph.D. and Master theses : **60 TB**

Cloud storage: ~ 60 k€ /year

Upload (@10Mbps): 1.5 years

The main problem of raw digitizing is file size

Digitized Ph.D. and Master theses : **60 TB**

Cloud storage: ~ 60 k€ /year

Upload (@10Mbps): 1.5 years

Preservation is unaffordable

The main problem of raw digitizing is file size

Digitized Ph.D. and Master theses : **60 TB**

Cloud storage: ~ 60 k€ /year

Upload (@10Mbps): 1.5 years

Preservation is unaffordable

Solution: Migration to JPEG2000

JPEG2000 **visually lossless** image compression

Original

JPEG 2000
(1:10)

JPEG2000 **visually lossless** image compression

Original

JPEG 2000
(1:10)

Quality metrics (PSNR...)

JPEG2000 **visually lossless** image compression

Original

JPEG 2000
(1:10)

Quality metrics (PSNR...)

Jpylizer

Compression detail losses are insignificant compared to artefacts produced while digitizing

Compression detail losses are insignificant compared to artefacts produced while digitizing

spatial resolution

Compression detail losses are insignificant compared to artefacts produced while digitizing

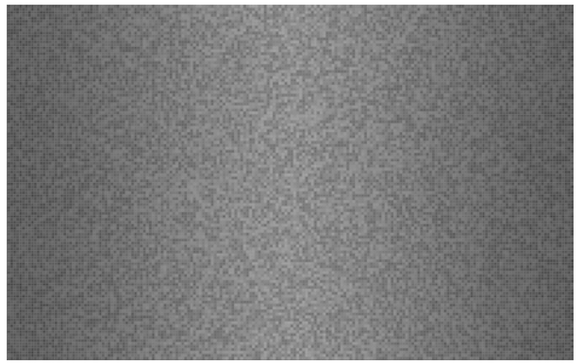
spatial resolution

color reproduction

Compression detail losses are insignificant compared to artefacts produced while digitizing

spatial resolution

color reproduction

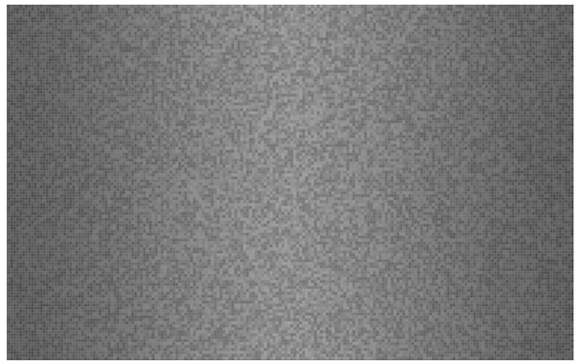


noise

Compression detail losses are insignificant compared to artefacts produced while digitizing

spatial resolution

color reproduction



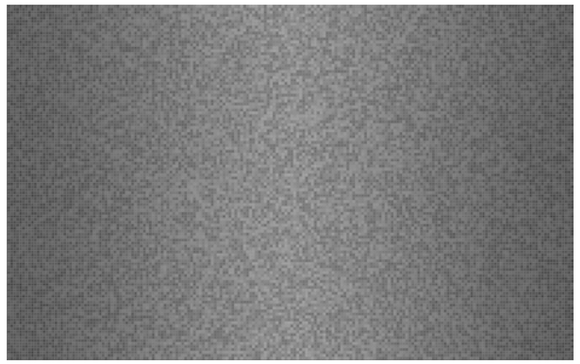
noise

dust

Compression detail losses are insignificant compared to artefacts produced while digitizing

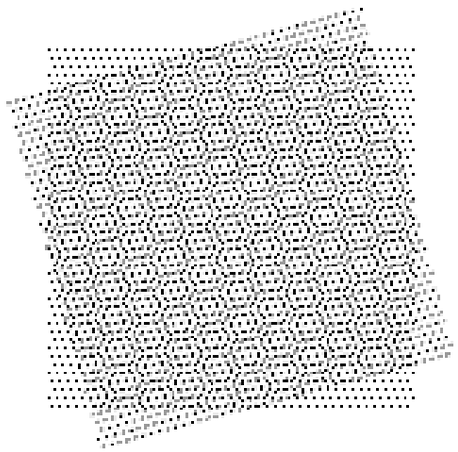
spatial resolution

color reproduction



noise

dust

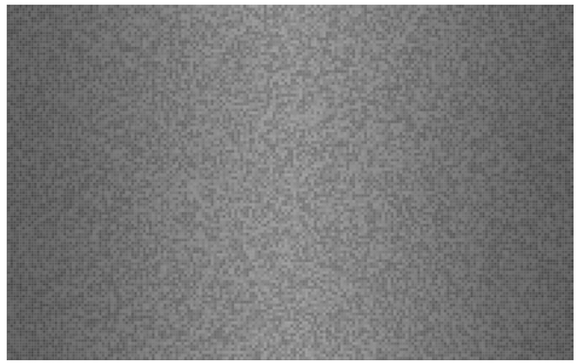


moiré

Compression detail losses are insignificant compared to artefacts produced while digitizing

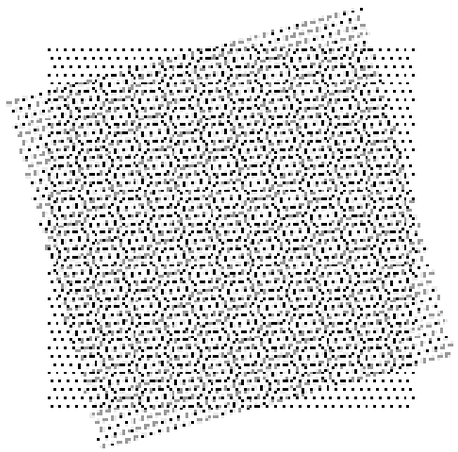
spatial resolution

color reproduction



noise

dust



moiré

geometric distorsion

JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

JPEG 2000 offers many advantages

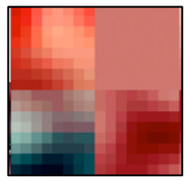
JPEG2000 part 1
is a standard

Visually loss-less high
image compression

JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

Visually loss-less high
image compression



No blocks effect

JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

Visually loss-less high
image compression

No blocks effect

High color resolution
max 48-bits

JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

Visually loss-less high
image compression

No blocks effect

High color resolution
max 48-bits

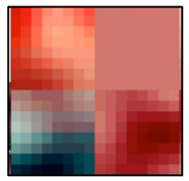
Support for very
large images

JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

Support for metadata

Visually loss-less high
image compression



No blocks effect

High color resolution
max 48-bits

Support for very
large images

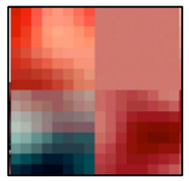
JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

Support for metadata

Visually loss-less high
image compression

Parametrizable resolution
for region of interest



No blocks effect

High color resolution
max 48-bits

Support for very
large images

JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

Support for metadata

Visually loss-less high
image compression

Parametrizable resolution
for region of interest

No blocks effect

One unique file for thumbnails
dissemination, preservation

High color resolution
max 48-bits

Support for very
large images

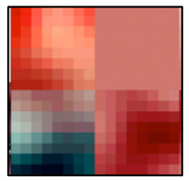
JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

Support for metadata

Visually loss-less high
image compression

Parametrizable resolution
for region of interest



No blocks effect

One unique file for thumbnails
dissemination, preservation

High color resolution
max 48-bits

IP rights management with
watermarking, degradation

Support for very
large images

JPEG 2000 offers many advantages

JPEG2000 part 1
is a standard

Support for metadata

Visually loss-less high
image compression

Parametrizable resolution
for region of interest

No blocks effect

One unique file for thumbnails
dissemination, preservation

High color resolution
max 48-bits

IP rights management with
watermarking, degradation

Support for very
large images

Robust (structure at bit
and paquet-level)

Using JPEG2000 allows us to drastically reduce the storage requirements for raw images

Books
Maps
Pictures

PDF HQ
lossless JPEG2000 (50%)

1.5TB (3TB)

Very high quality

PhD Theses and publications

PDF HQ
visually lossless JPEG2000 (10%)

3.1 TB (31TB)

High quality

Master Theses

PDF MQ

104 GB (28TB)

Medium quality

Using JPEG2000 allows us to drastically reduce the storage requirements for raw images

Books
Maps
Pictures

PDF HQ
lossless JPEG2000 (50%)

1.5TB (3TB)

Very high quality

PhD Theses and publications

PDF HQ
visually lossless JPEG2000 (10%)

3.1 TB (31TB)

High quality

Master Theses

PDF MQ

104 GB (28TB)

Medium quality

Total: 4.5 TB (60TB)

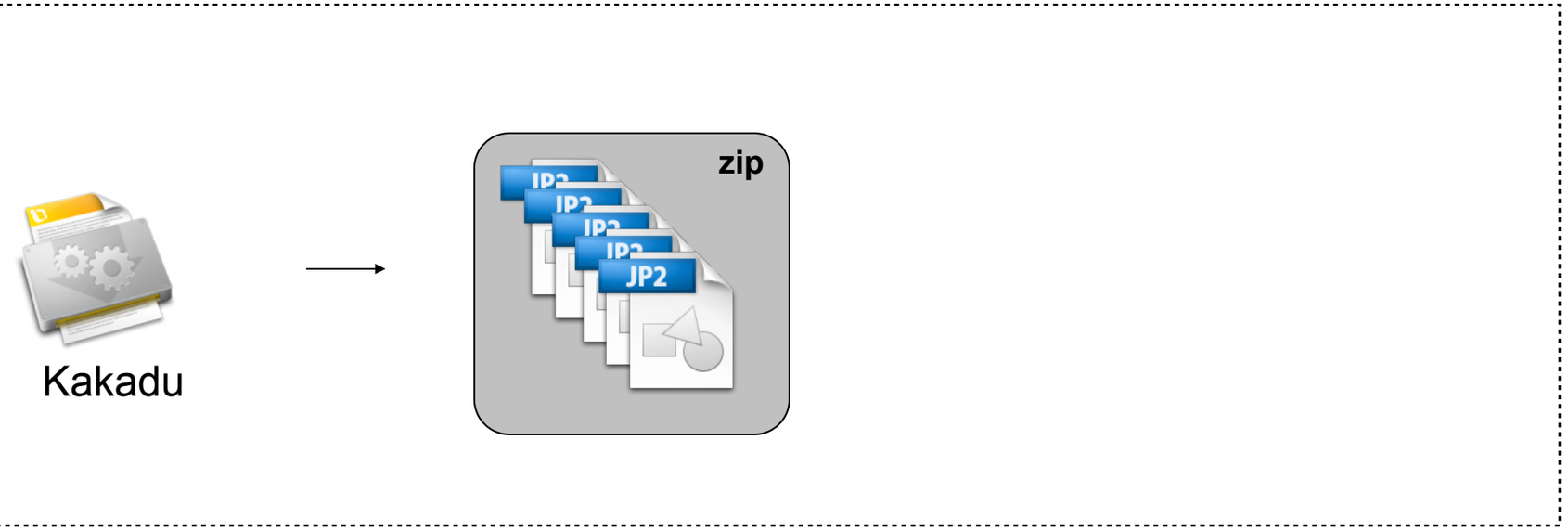
The digitizing workflow produces several files later used to create dissemination and archival packages

Raw images

Quality-improved images

The digitizing workflow produces several files later used to create dissemination and archival packages

Raw images



Quality-improved images

The digitizing workflow produces several files later used to create dissemination and archival packages

Raw images

TIFF
XML
<xml />
tech MD

Kakadu

zip
JP2

Quality-improved images

JPEG
XML
<xml />
tech MD

Finereader Engine

zip
JPEG

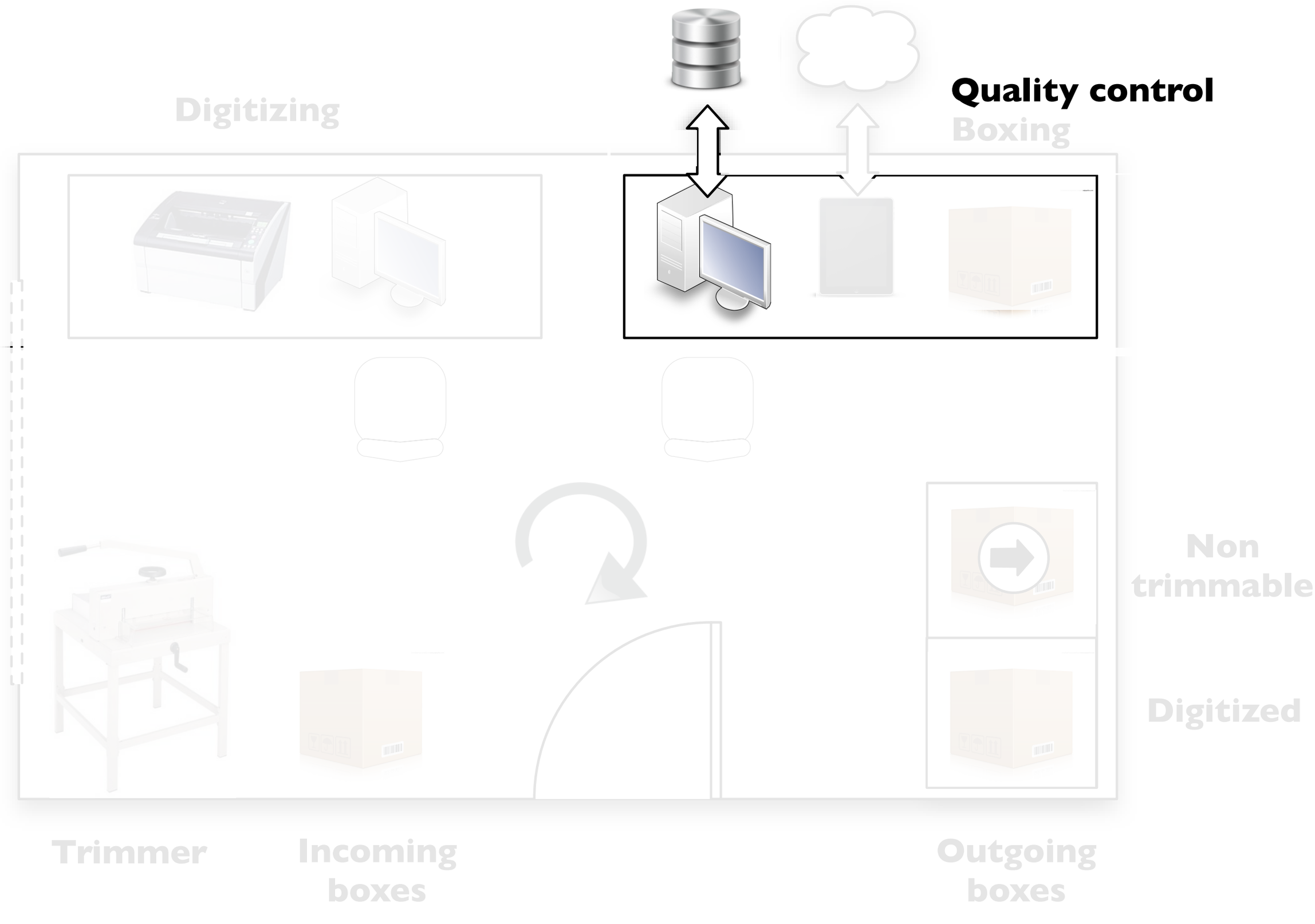
PDF/A

text under image

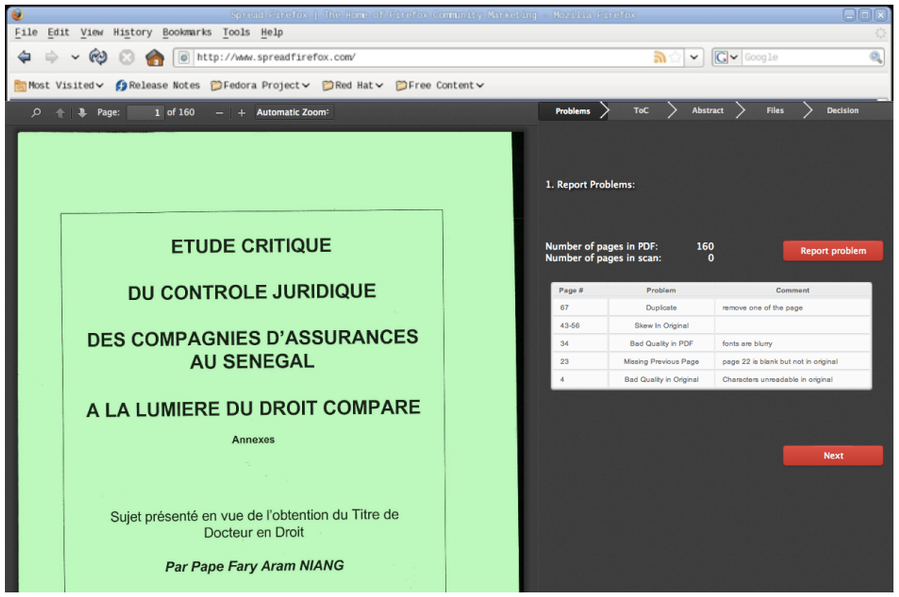
XML
<xml />
ALTO

TXT

A simple digitization workflow



We designed a client-server Quality Control application



client

server

HTML

CSS

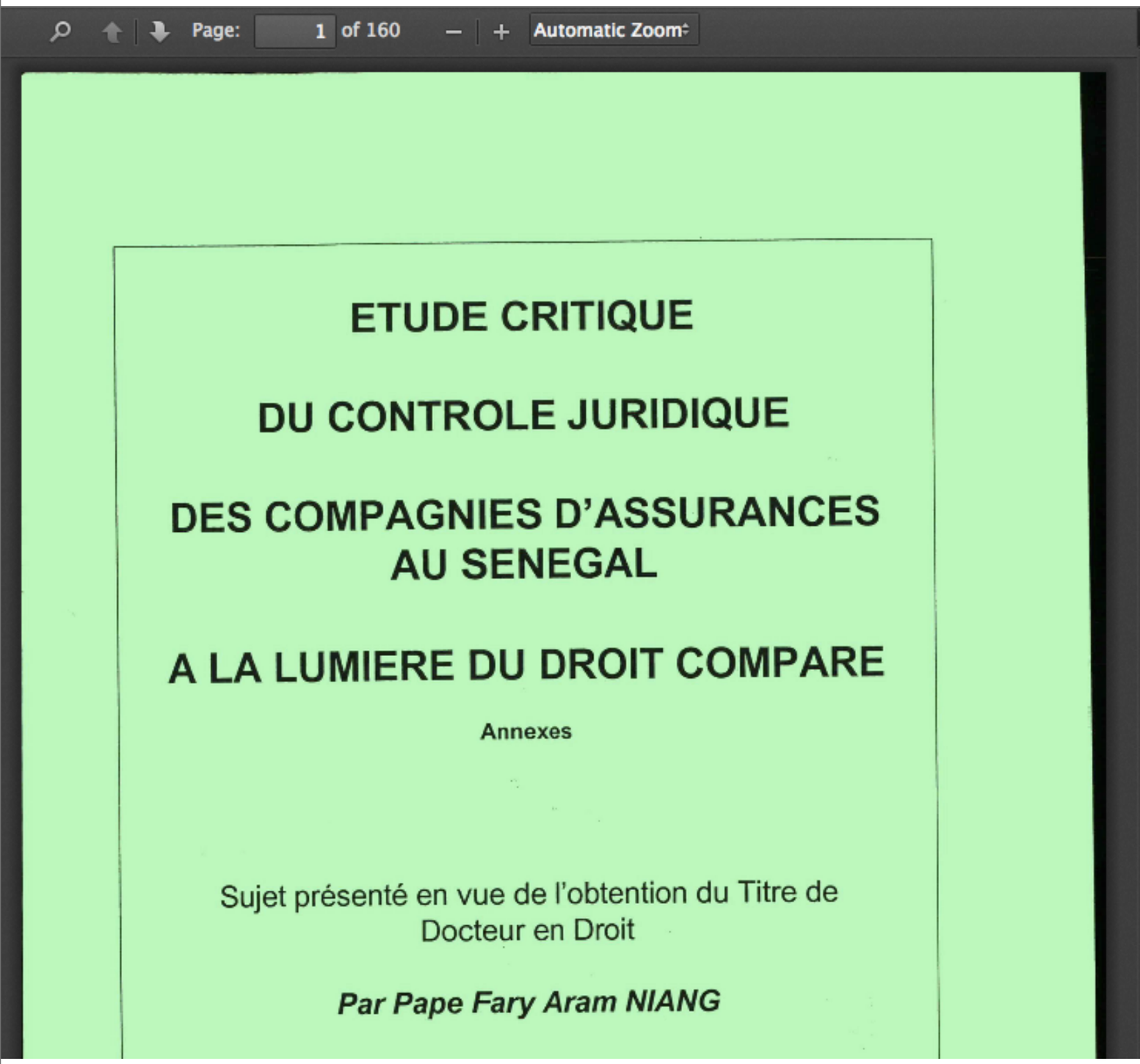
PDF JS

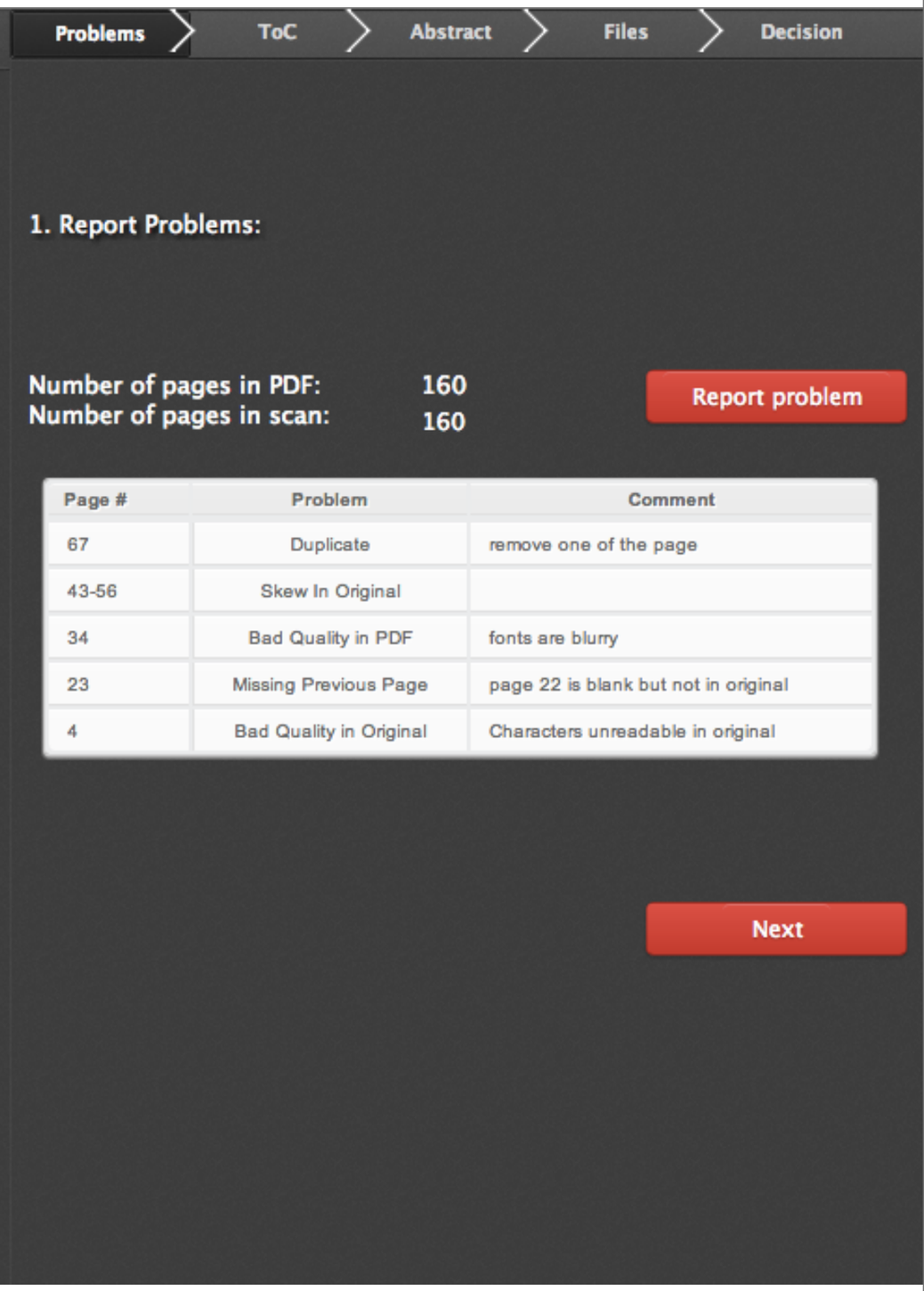
JavaScript

node JS

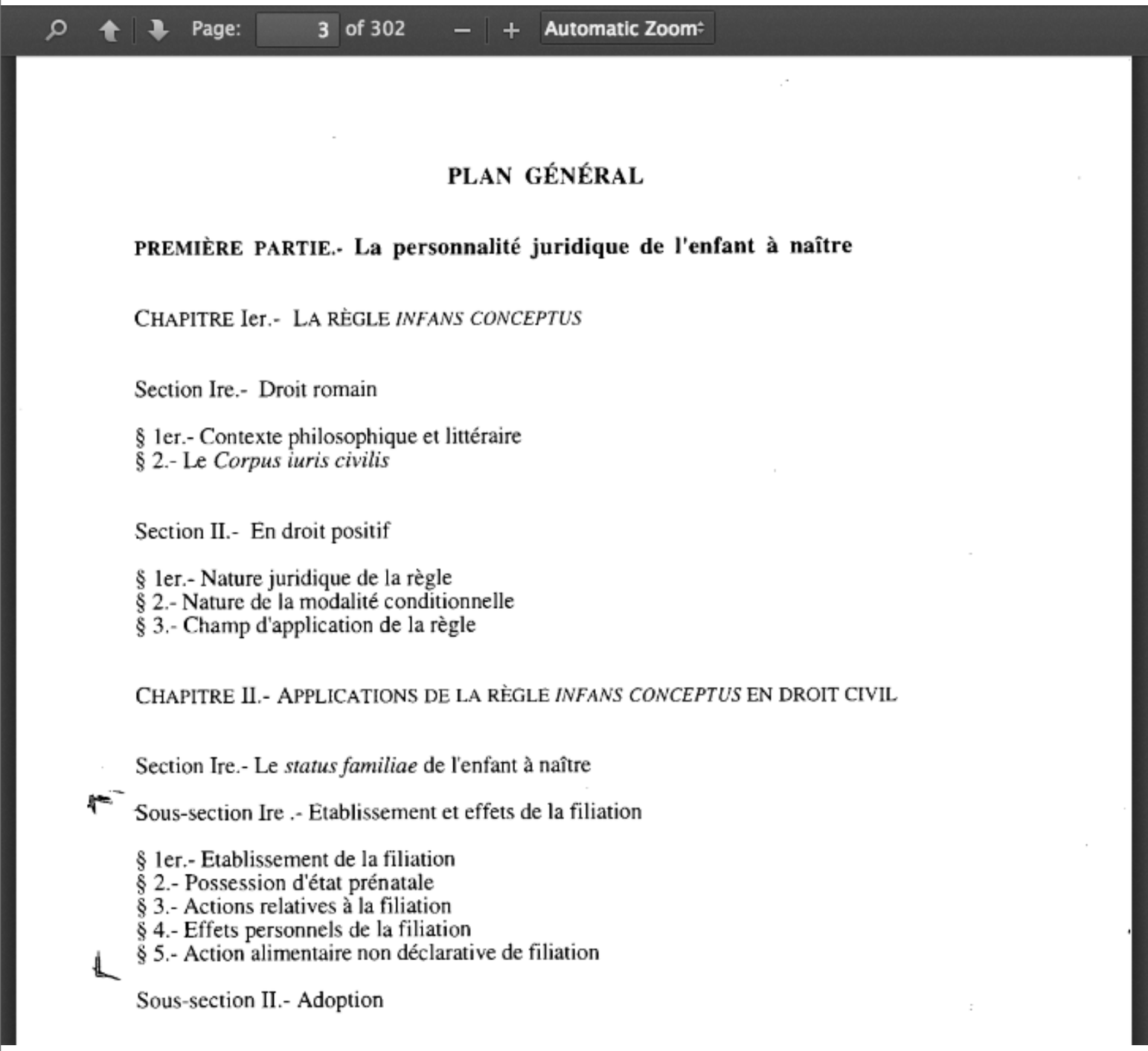
Quality control is performed with the original in hand

The application allows the operator to report problems in PDF or original

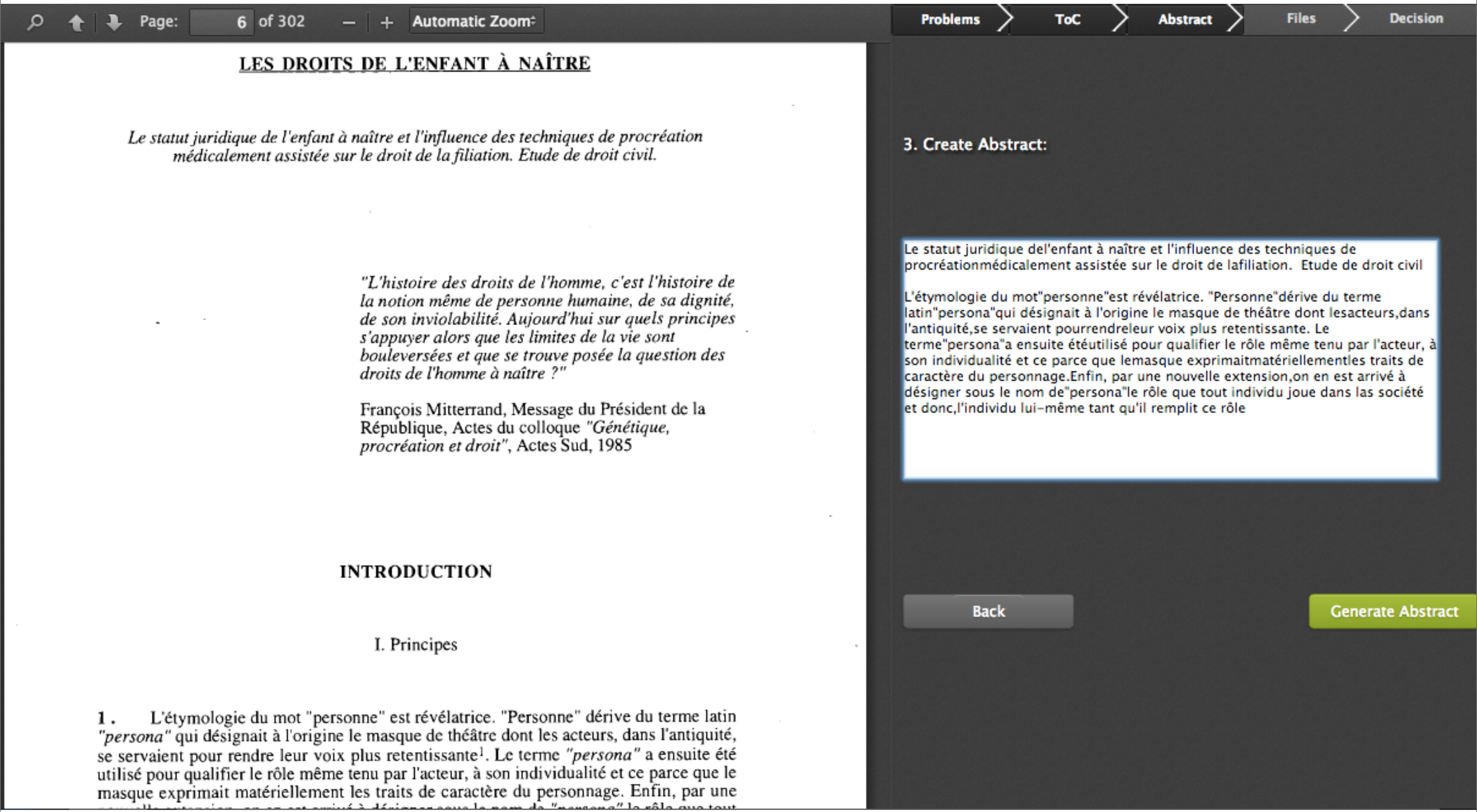




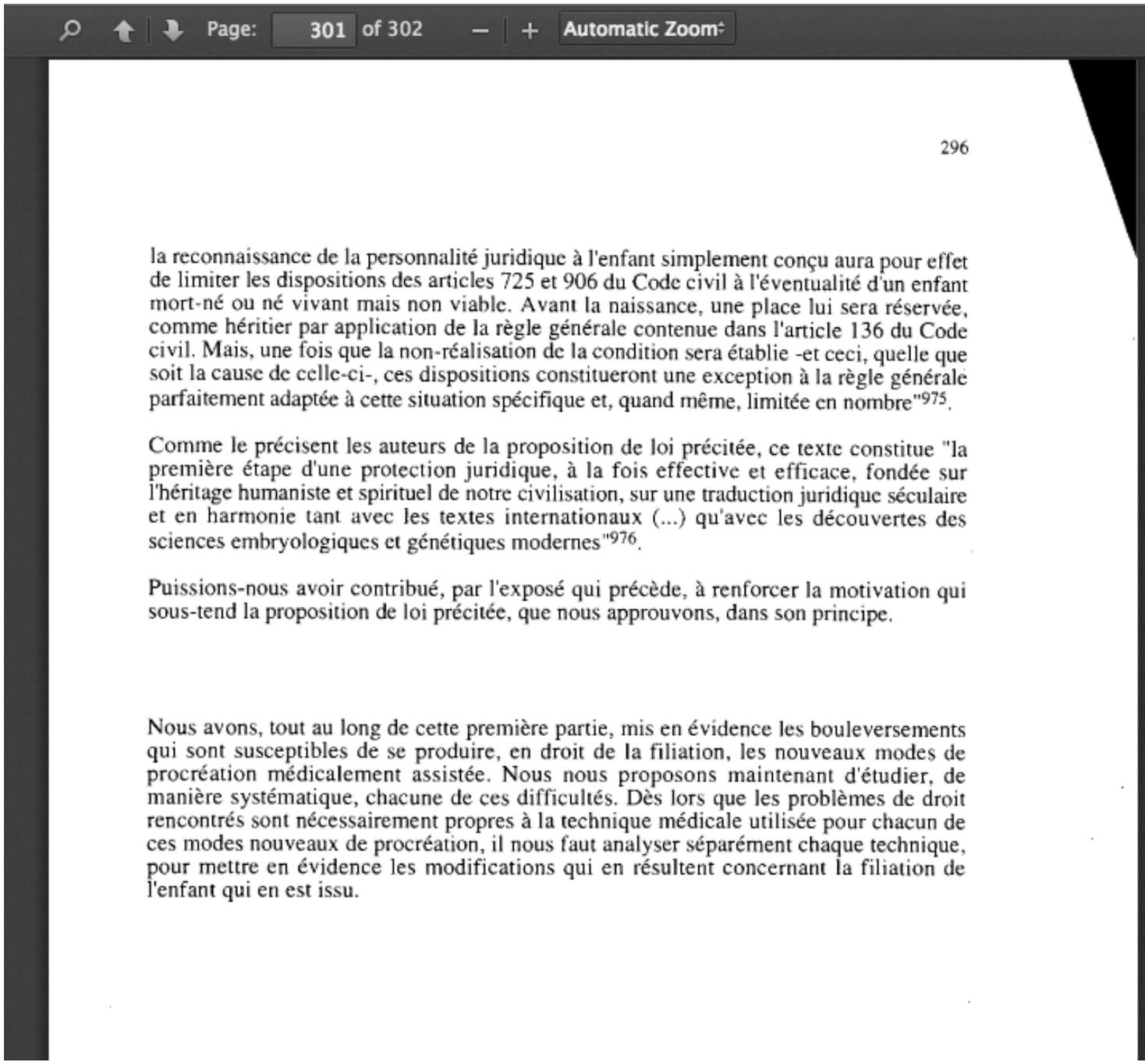
The application allows the operator to extract a Table of Content from the PDF



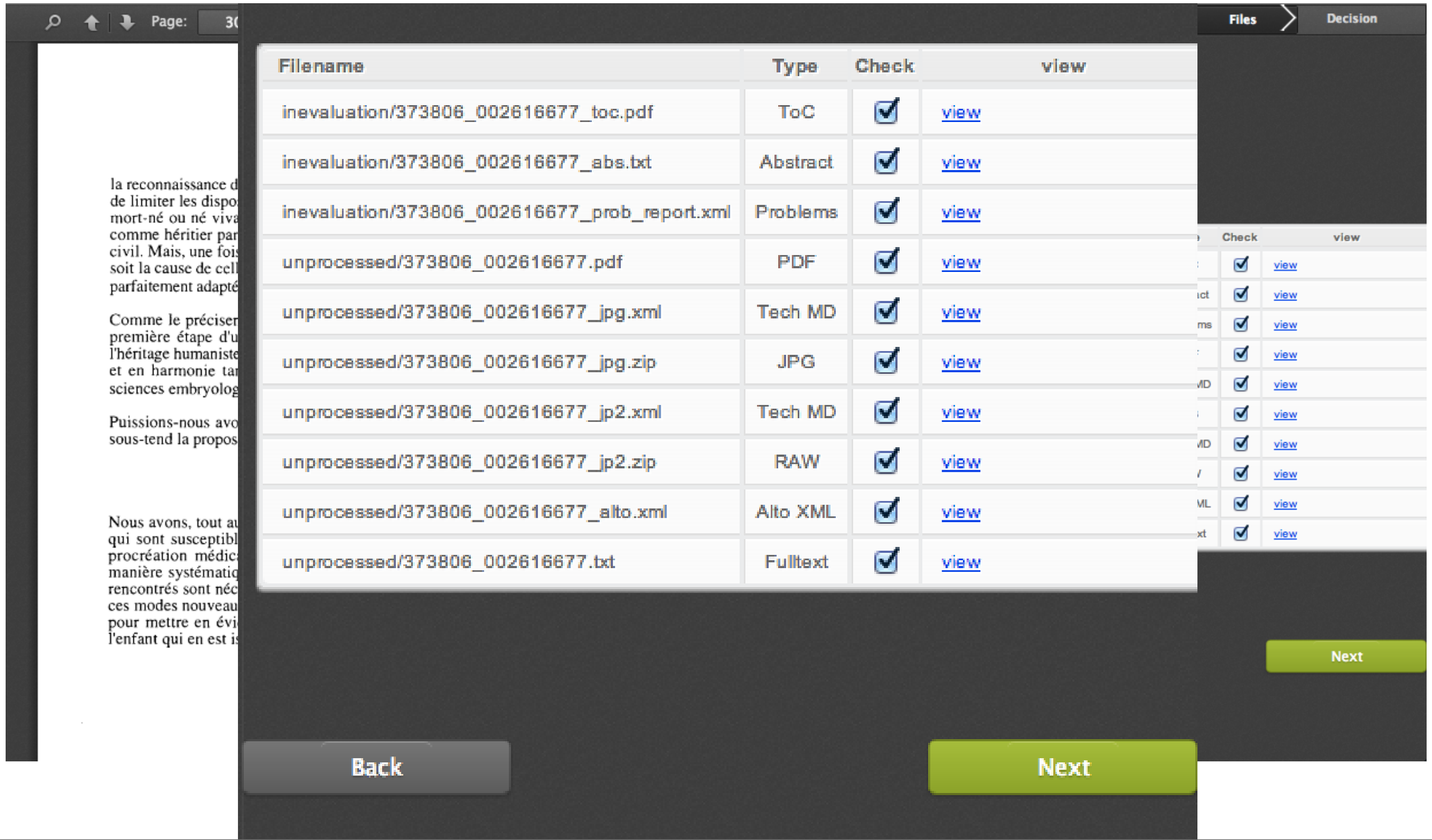
The application allows the operator to extract an abstract from the PDF



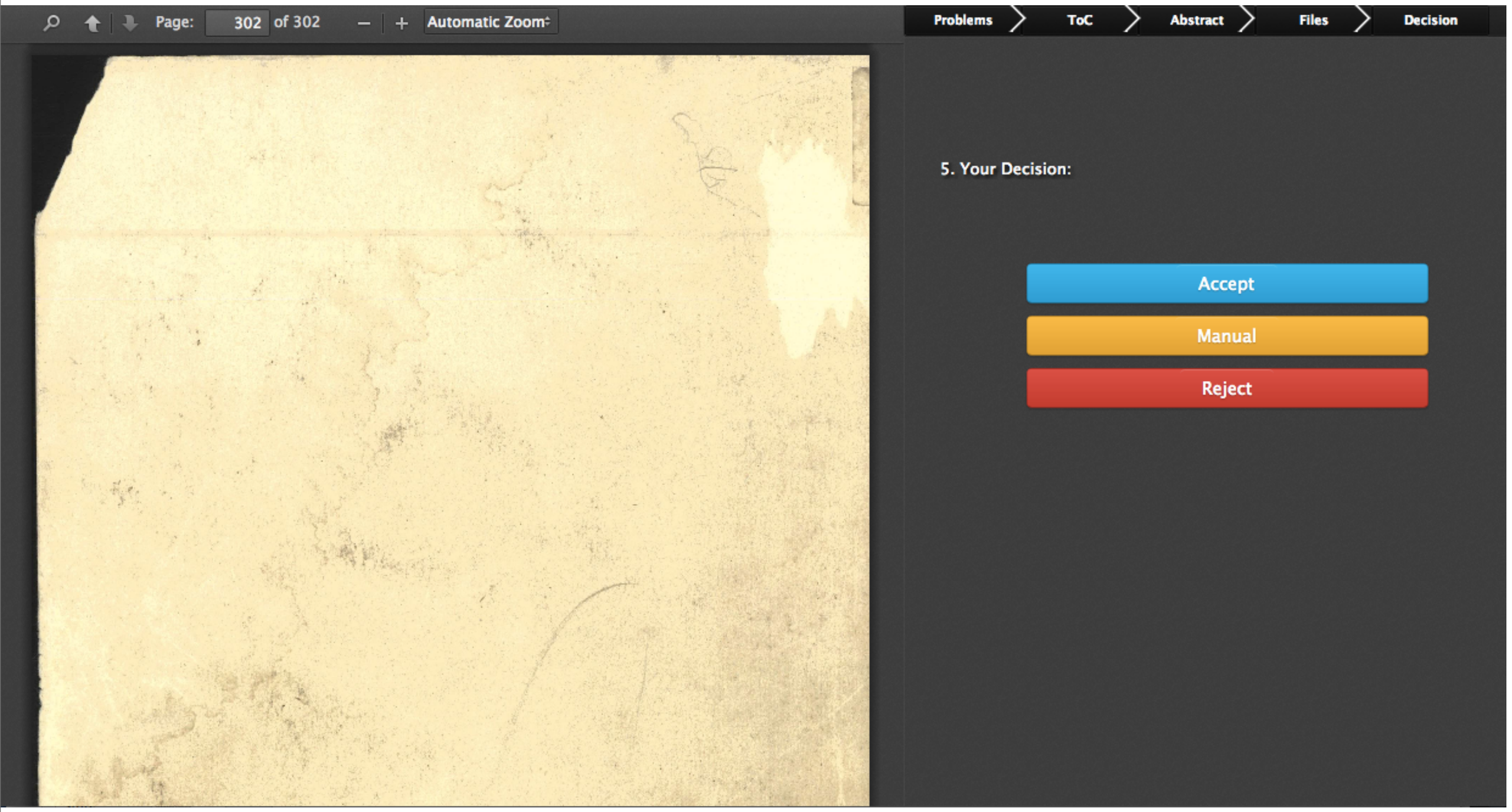
The application allows the operator to check all generated files



The application allows the operator to check all generated files



The user can then accept or reject the thesis or decide that it requires manual editing



The result of quality control...

Manual

Accepted

Rejected

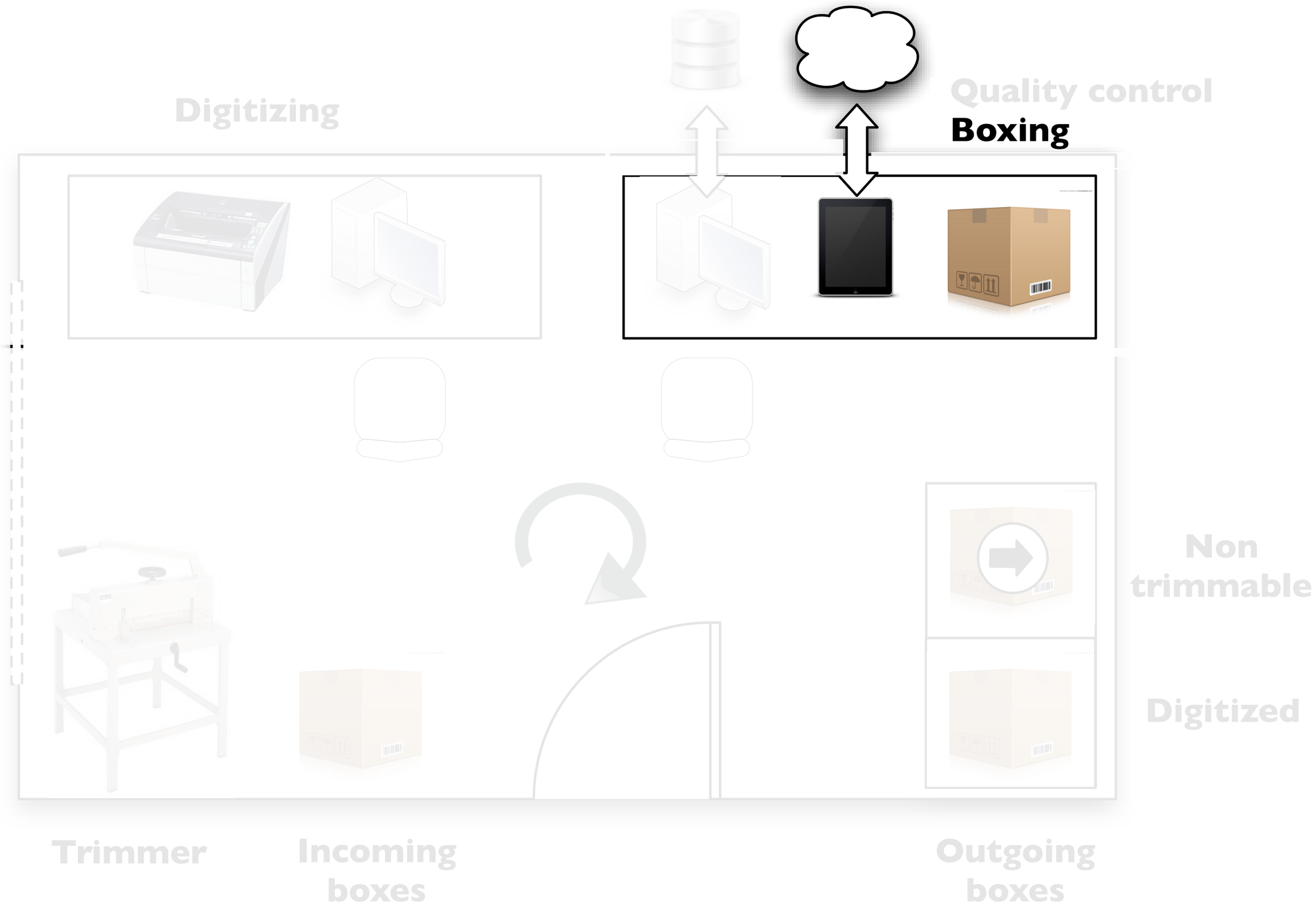
The result of quality control...

Manual

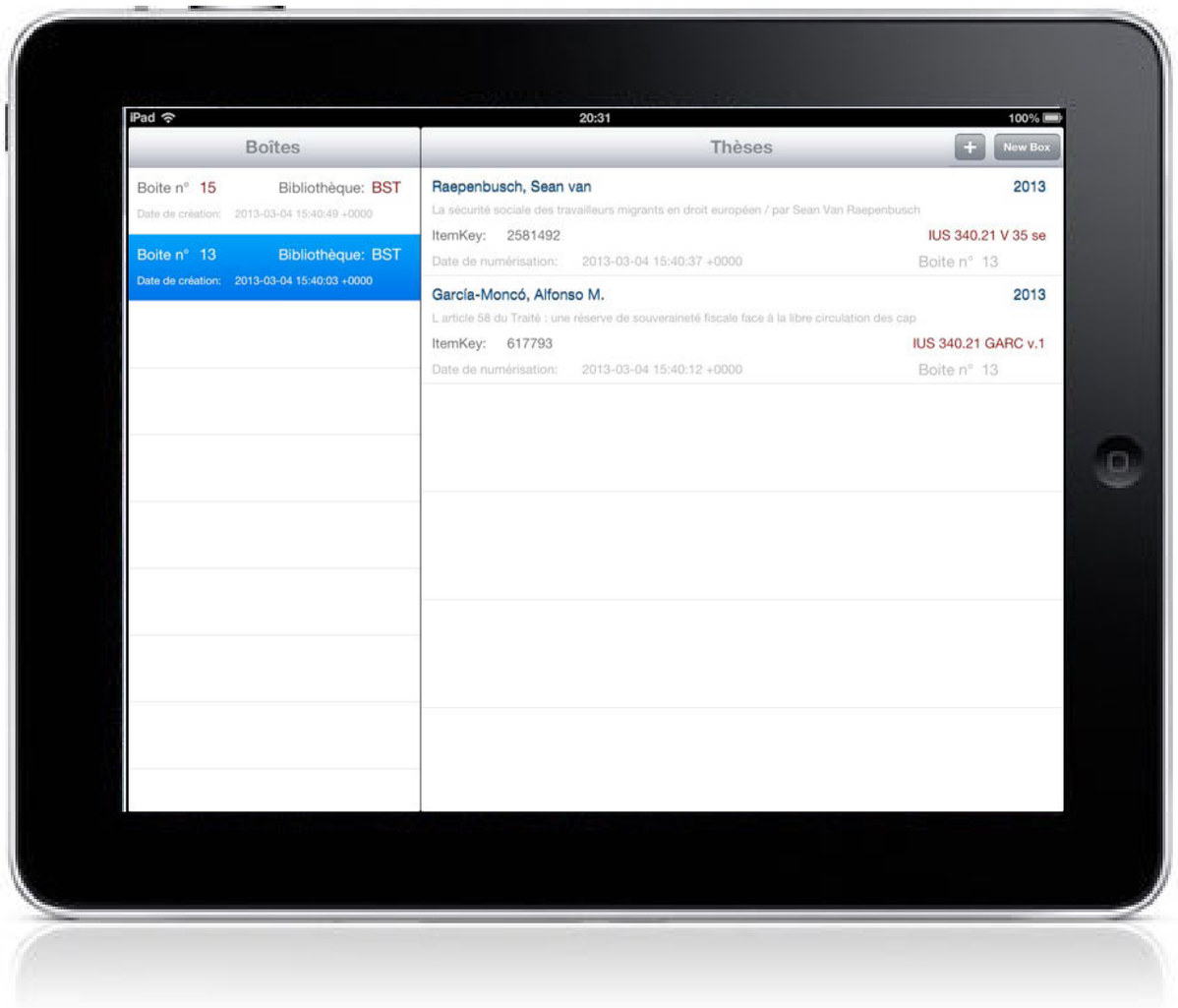
Accepted

Rejected

A simple digitization workflow



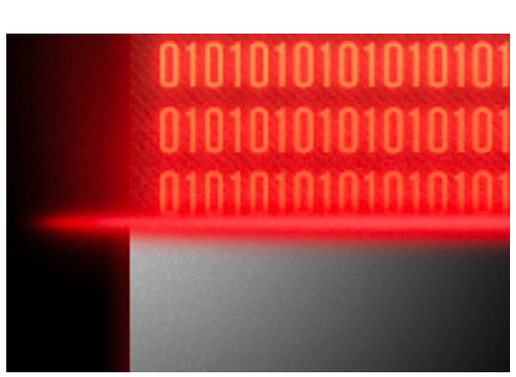
After acceptance by QC, the catalog location of paper is automatically updated with box coordinates with an iPad application



simperium

Update of location
in our catalog

Why did we produce all those files? For dissemination and preservation

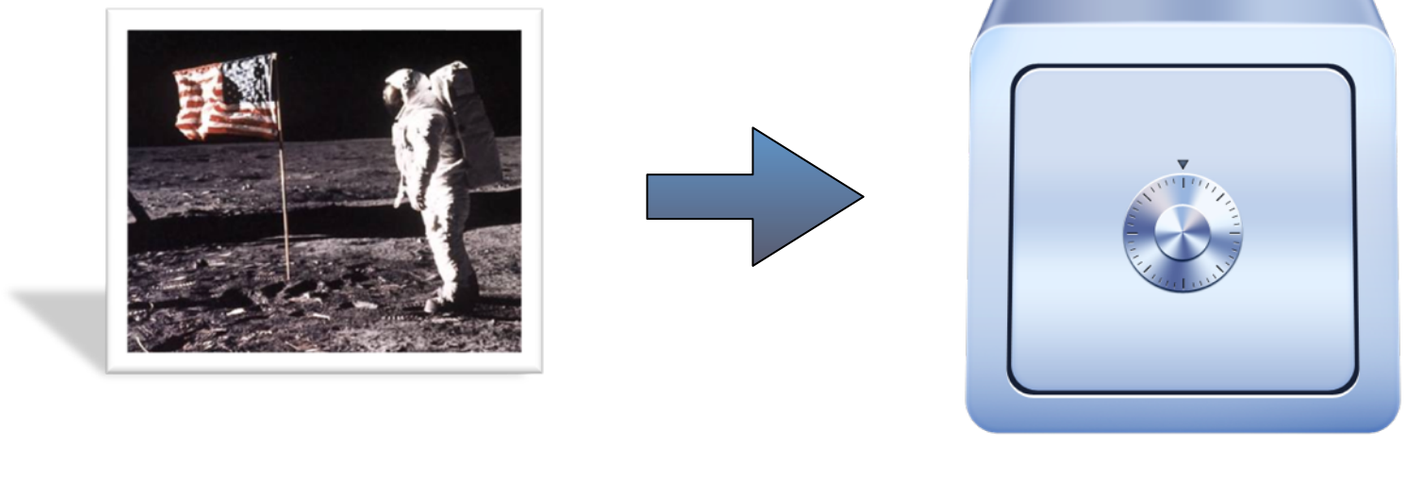


Digitization

Dissemination

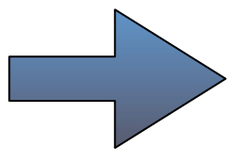
Preservation

On one hand, we want to preserve digital objects themselves for posterity



2013

We want digital objects to remain identical at the bit-level and still understandable in the long-term future



2113

LEARNING RESOURCES

On the other hand, we want to provide the users with state-of-the-art interfaces to offer the best quality of experience to view digital objects

Document
download
medium quality
connection-less
easy to copy
built-in support
PDF

Book viewer
streaming
high-quality
web based
difficult to copy
jQuery
JPEG2000

IIPImage
streaming
high-quality
web based
difficult to copy
js & flash
JPEG2000

On the other hand, we want to provide the users with state-of-the-art interfaces to offer the best quality of experience to view digital objects

Document
download
medium quality
connection-less
easy to copy
built-in support
PDF

Book viewer
streaming
high-quality
web based
difficult to copy
jQuery
JPEG2000

IIPImage
streaming
high-quality
web based
difficult to copy
js & flash
JPEG2000

We want user interfaces to *evolve* in time

We want to tackle simulatenously two opposite problems:

Preserving
data

vs

Improving
user interfaces

We need an integrated Preservation & Dissemination workflow

We want to tackle simulatenously two opposite problems:

Preserving
data

AIP

vs

Improving
user interfaces

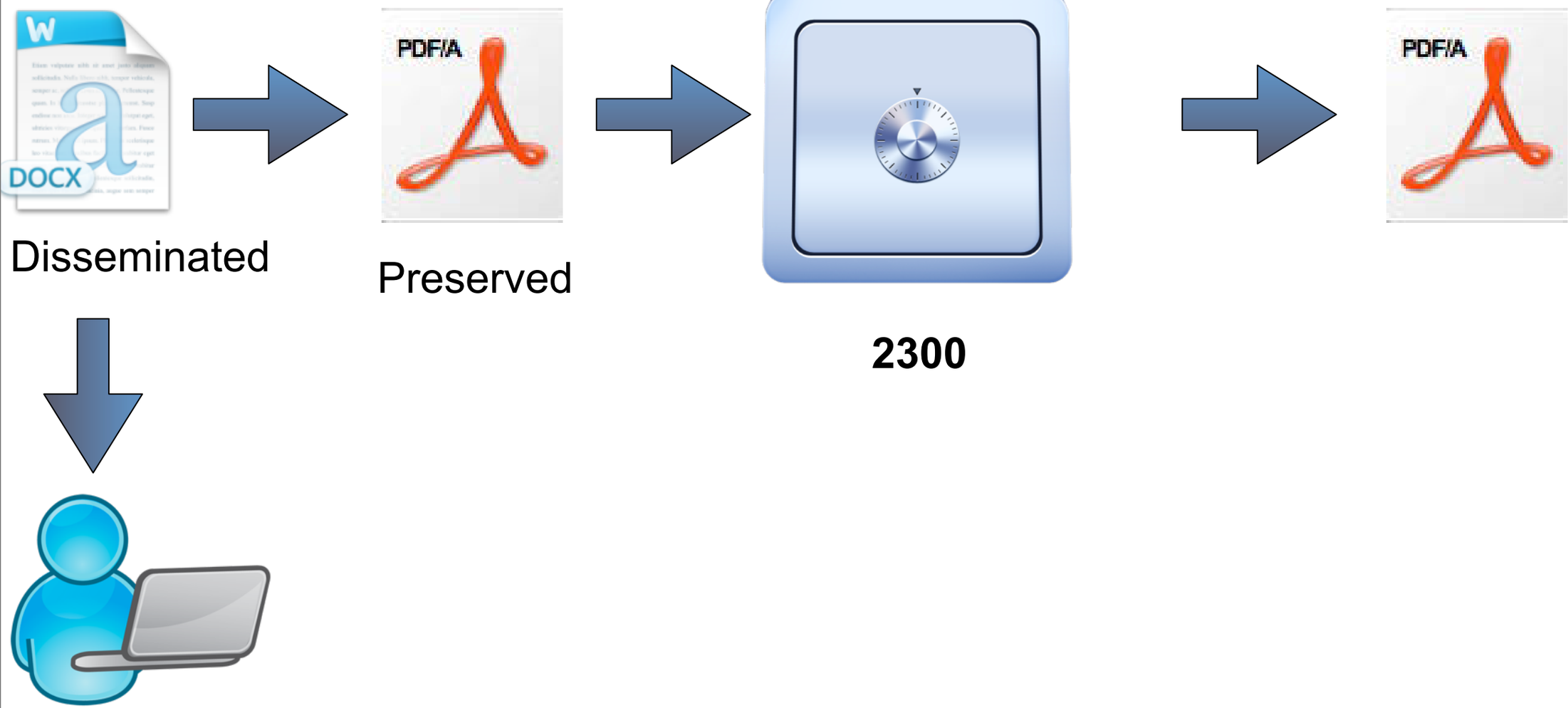
DIP

We need an integrated Preservation & Dissemination workflow

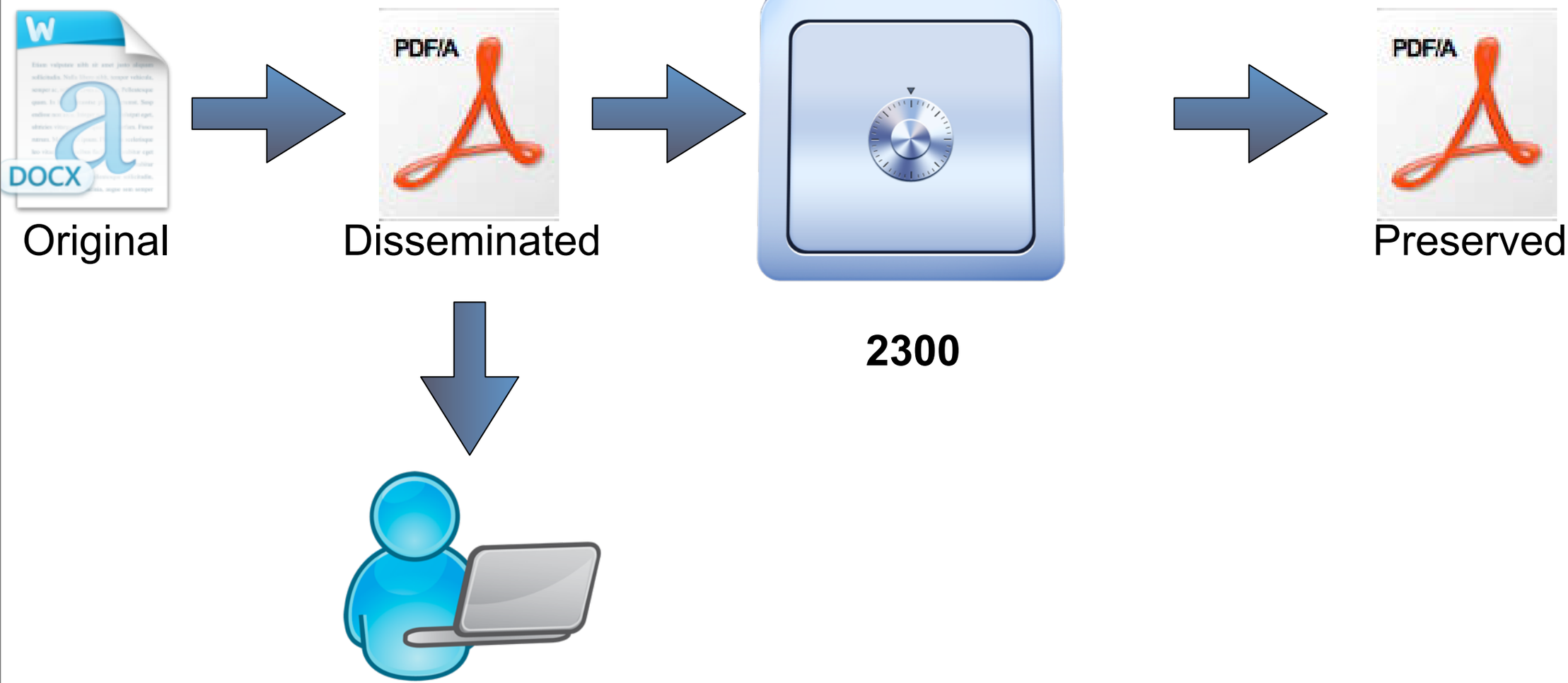
Proprietary closed formats are prone to obsolescence

2300

To ensure understandibility on the long-term, objects need to be migrated in long-lasting open formats

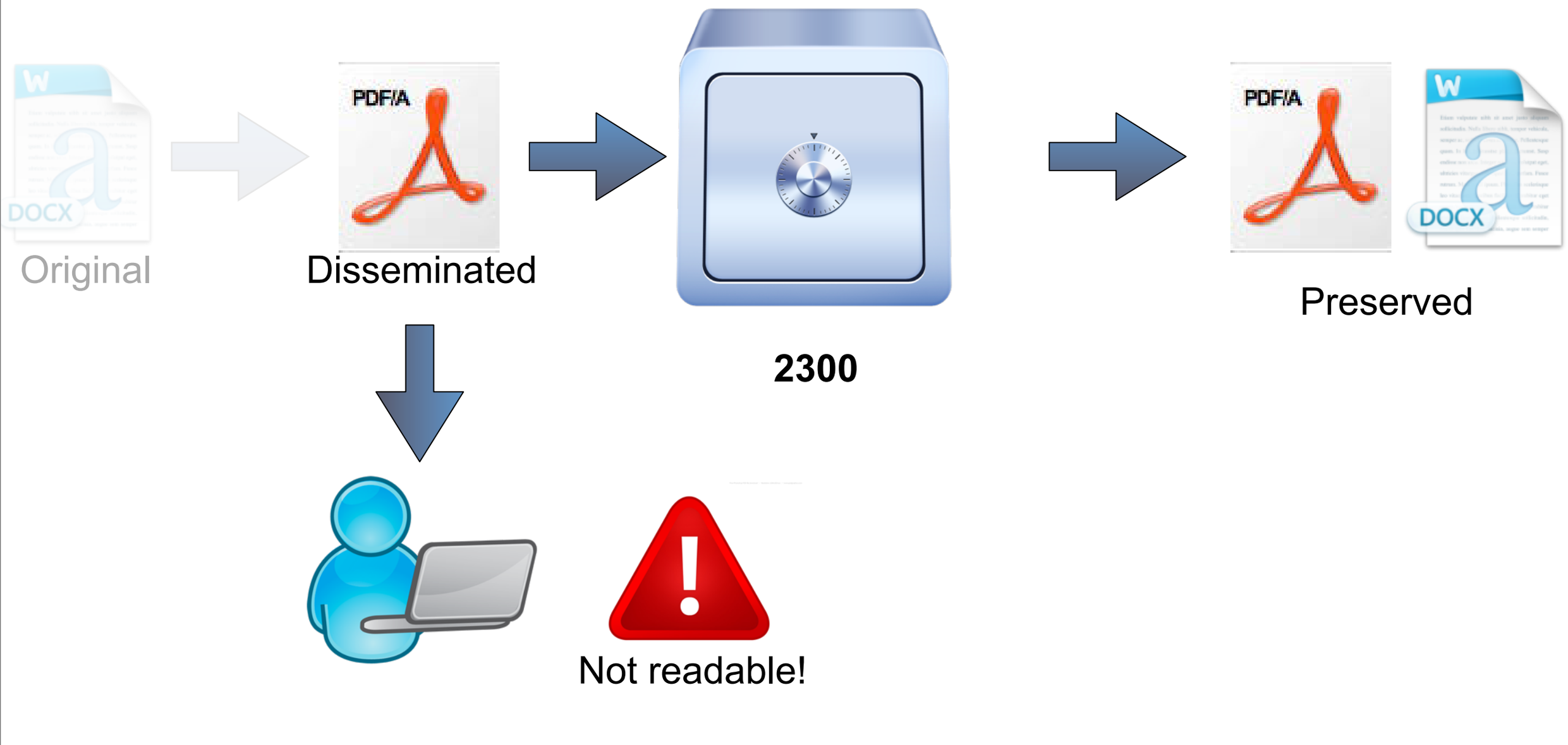


As only disseminated objects are visible to users, they have to match the preserved objects to ensure readability



If disseminated object is unreadable or does not match the original object, it needs to be rebuilt from the original file

Therefore, the original file would also need to be preserved



Except when we are absolutely sure that the preservation version is a perfect copy of the original

Original

JPEG 2000
(1:10)

Except when we are absolutely sure that the preservation version is a perfect copy of the original

Original

JPEG 2000
(1:10)

Quality metrics (PSNR...)

Except when we are absolutely sure that the preservation version is a perfect copy of the original

Original

JPEG 2000
(1:10)

Quality metrics (PSNR...)

Jpylizer

Except when we are absolutely sure that the preservation version is a perfect copy of the original

Original

JPEG 2000
(1:10)

Quality metrics (PSNR...)

Jpylizer

Quality Control

The SIP is extracted from the IR. The P&D service creates the AIP and possibly multiple DIPs for the object

Institutional Repository

Preservation
and
Dissemination
Service

The SIP is extracted from the IR. The P&D service creates the AIP and possibly multiple DIPs for the object

Institutional Repository

Preservation
and
Dissemination
Service

1!

The SIP is extracted from the IR. The P&D service creates the AIP and possibly multiple DIPs for the object

Institutional Repository

Preservation
and
Dissemination
Service

1!

The SIP is extracted from the IR. The P&D service creates the AIP and possibly multiple DIPs for the object

Institutional Repository

Preservation
and
Dissemination
Service

1!

The SIP is extracted from the IR. The P&D service creates the AIP and possibly multiple DIPs for the object

Institutional Repository

Preservation
and
Dissemination
Service

1!

The SIP is extracted from the IR. The P&D service creates the AIP and possibly multiple DIPs for the object

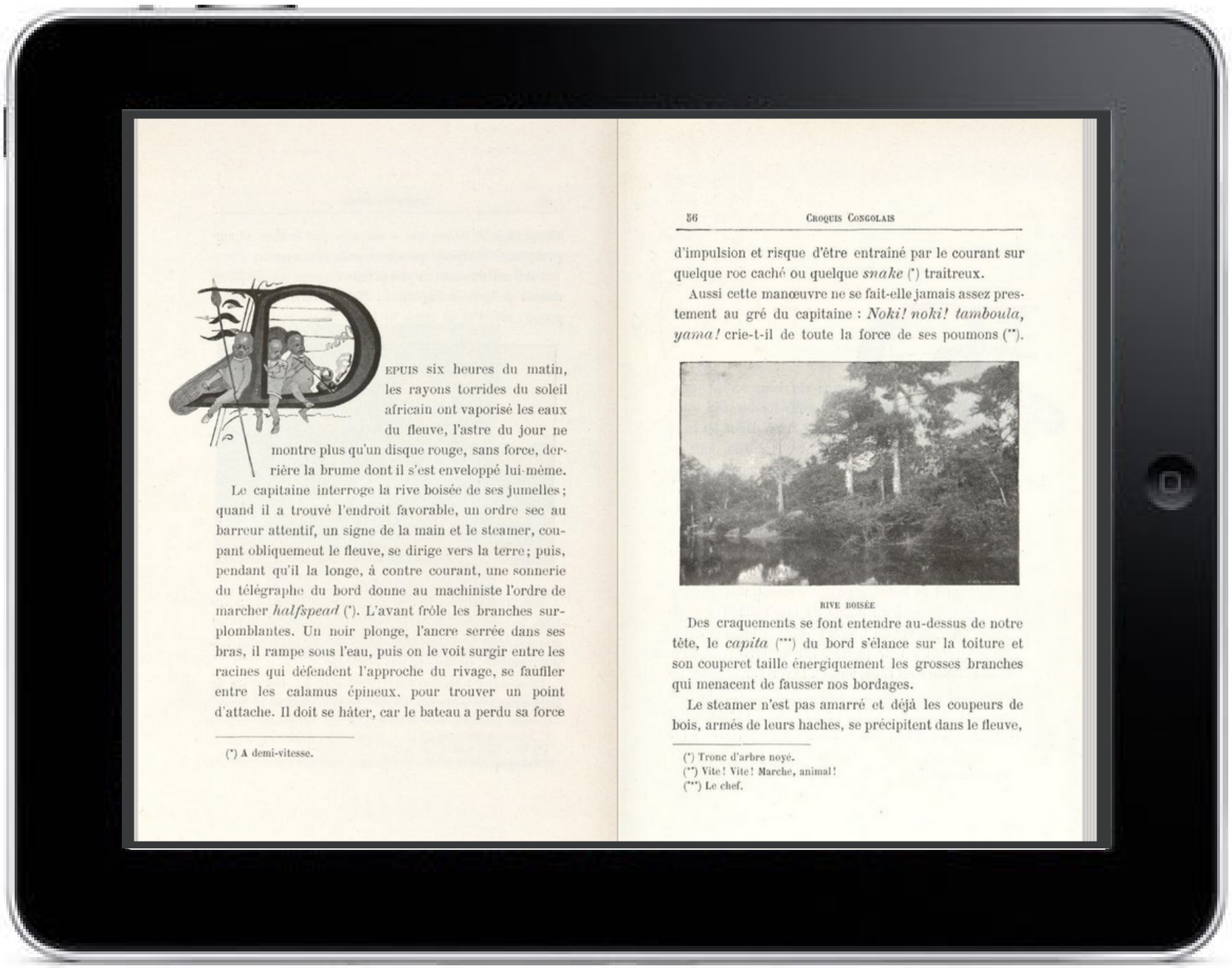
Institutional Repository

1..N

Preservation
and
Dissemination
Service

1!

For books, we are using the Internet Archive Bookreader format



DEPUIS six heures du matin, les rayons torrides du soleil africain ont vaporisé les eaux du fleuve, l'astre du jour ne montre plus qu'un disque rouge, sans force, derrière la brume dont il s'est enveloppé lui-même.

Le capitaine interroge la rive boisée de ses jumelles ; quand il a trouvé l'endroit favorable, un ordre sec au barreur attentif, un signe de la main et le steamer, coupant obliquement le fleuve, se dirige vers la terre ; puis, pendant qu'il la longe, à contre courant, une sonnerie du télégraphe du bord donne au machiniste l'ordre de marcher *halfspead* (*). L'avant frôle les branches surplombantes. Un noir plonge, l'ancre serrée dans ses bras, il rampe sous l'eau, puis on le voit surgir entre les racines qui défendent l'approche du rivage, se faufiler entre les calamus épineux, pour trouver un point d'attache. Il doit se hâter, car le bateau a perdu sa force d'impulsion et risque d'être entraîné par le courant sur quelque roc caché ou quelque *snake* (*) traîtreux.

(*) A demi-vitesse.

Aussi cette manœuvre ne se fait-elle jamais assez prestement au gré du capitaine : *Noki! noki! tamboula, yama!* crie-t-il de toute la force de ses poumons (**).

RIVE BOISÉE

Des craquements se font entendre au-dessus de notre tête, le *capita* (***) du bord s'élance sur la toiture et son couperet taille énergiquement les grosses branches qui menacent de fausser nos bordages.

Le steamer n'est pas amarré et déjà les coupeurs de bois, armés de leurs haches, se précipitent dans le fleuve,

(*) Tronc d'arbre noyé.
(**) Vite ! Vite ! Marche, animal !
(***) Le chef.

Internet Archive BookReader format offers several advantages

GNU GPLv3 license

Completely customizable

Based on standards

Safe and reliable

Over 1M books available online in this format

Internet Archive BookReader format offers several advantages

GNU GPLv3 license

Completely customizable

Based on standards

Safe and reliable

Over 1M books available online in this format

The implementation offers many interesting features for users:

- full-text search supporting metacharacters
- book structure visible in the progress bar
- link to the repository item
- text-to-speech in english
- text-to-speech in french

AIPs & DIPs are concepts...
What about the hard reality of institutional repositories ?

Institutional
Repository

DSPACE

AIPs & DIPs are concepts...
What about the hard reality of institutional repositories ?

Institutional Repository

DSPACE

item {handle}

- bitstream 1: Tech md
- bitstream 2: JP2.zip
- bitstream 3: bookid_meta.xml
- bitstream 4: bookid_files.xml
- bitstream 5: scandata.xml
- bitstream 6: alto.xml
- bitstream 7: txt
- bitstream 8: PDF

AIPs & DIPs are concepts...
What about the hard reality of institutional repositories ?

Institutional Repository

DSPACE

item {handle}

- bitstream 1: Tech md
- bitstream 2: JP2.zip
- bitstream 3: bookid_meta.xml
- bitstream 4: bookid_files.xml
- bitstream 5: scandata.xml
- bitstream 6: alto.xml
- bitstream 7: txt
- bitstream 8: PDF

How can we describe the relations between all those files?

What has to be changed?

New high-volume digitization workflow

New description model for all digital objects

A preservation strategy for all digital objects

Intellectual entity and representations

item {handle}

- bitstream 1: Tech md
- bitstream 2: JP2.zip
- bitstream 3: bookid_meta.xml
- bitstream 4: bookid_files.xml
- bitstream 5: scandata.xml
- bitstream 6: alto.xml
- bitstream 7: txt
- bitstream 8: PDF

Intellectual entity and representations

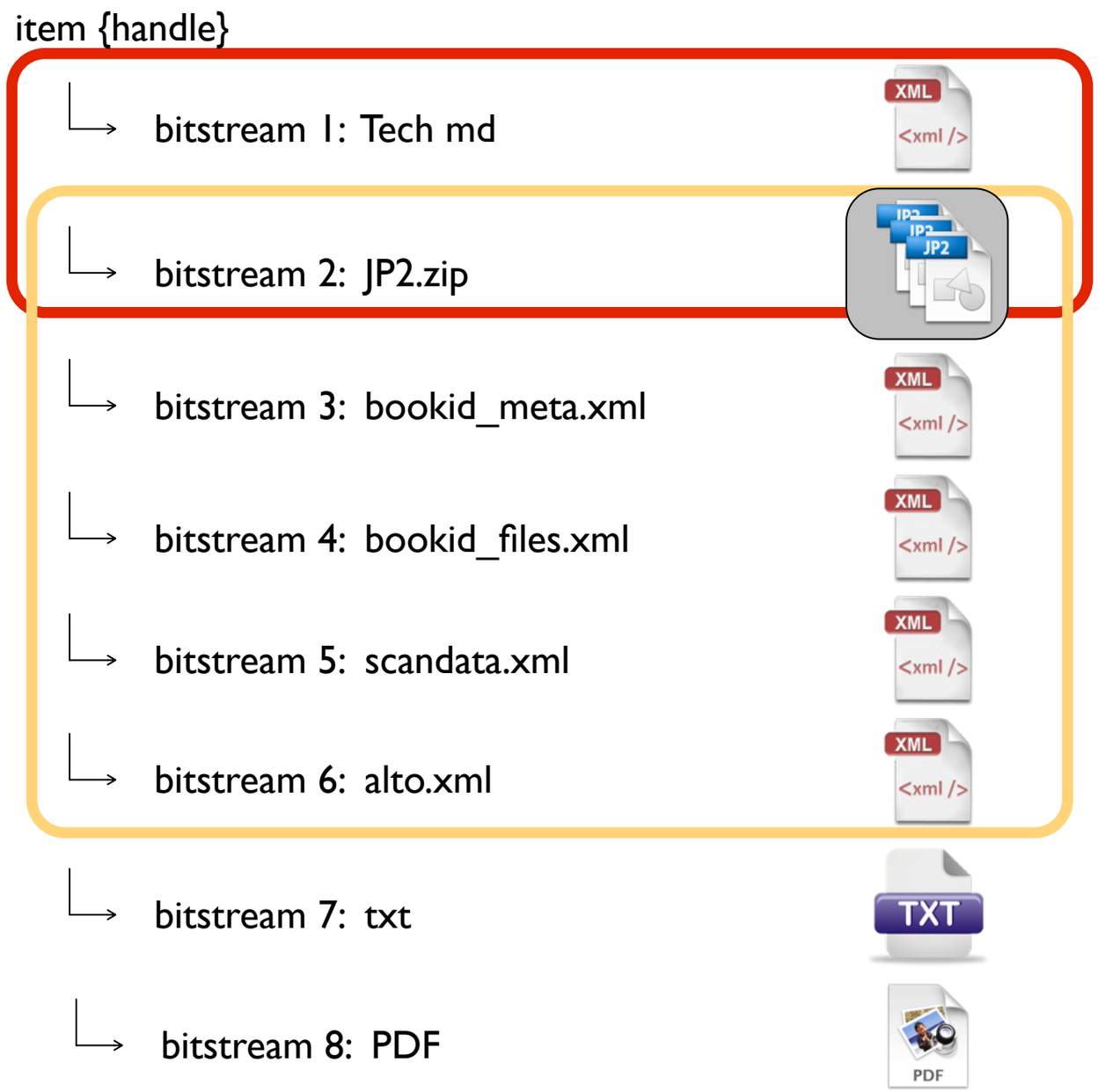
Representation of an IE

item {handle}

- bitstream 1: Tech md
- bitstream 2: JP2.zip
- bitstream 3: bookid_meta.xml
- bitstream 4: bookid_files.xml
- bitstream 5: scandata.xml
- bitstream 6: alto.xml
- bitstream 7: txt
- bitstream 8: PDF

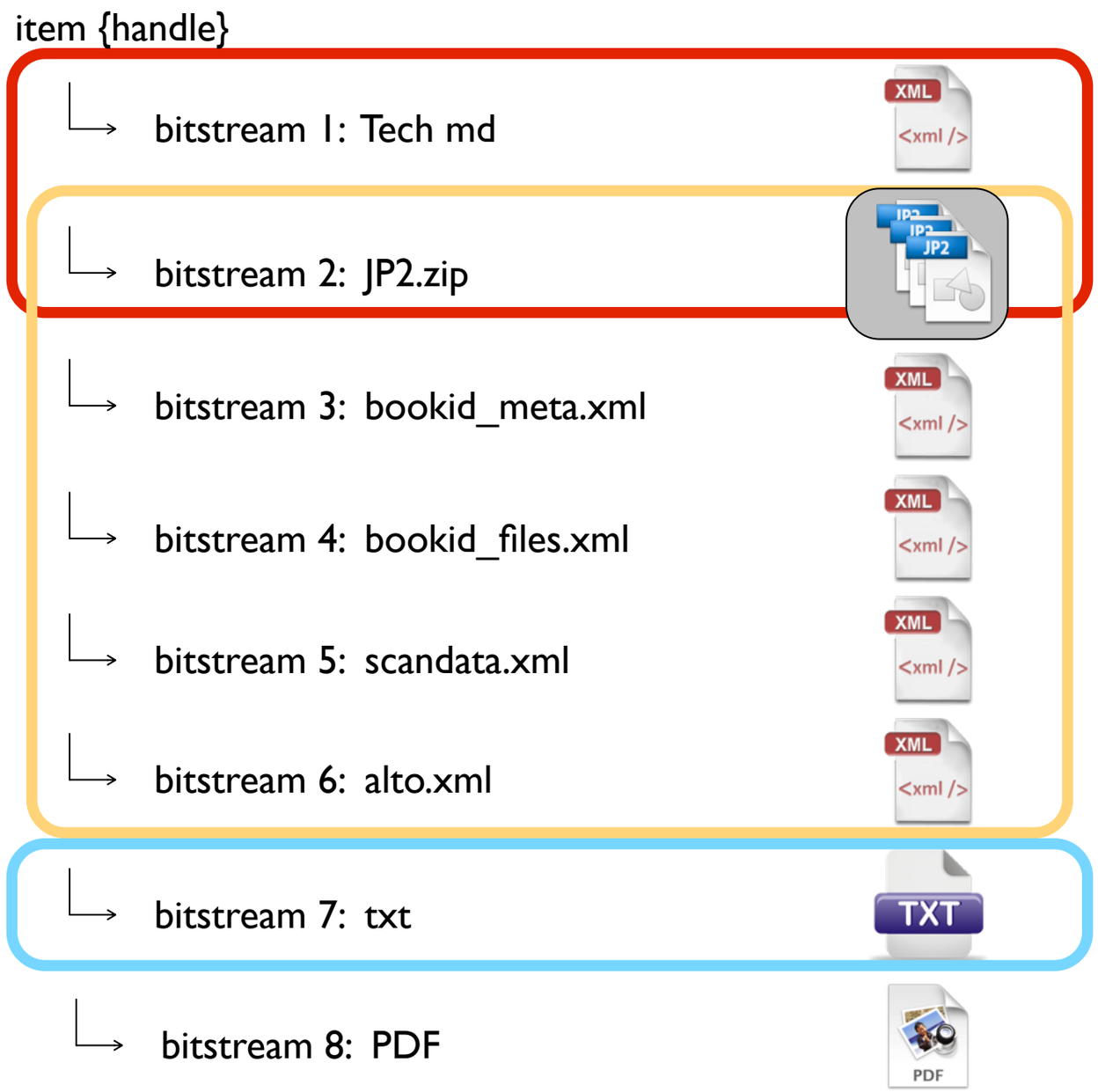
Intellectual entity and representations

Representation of an IE



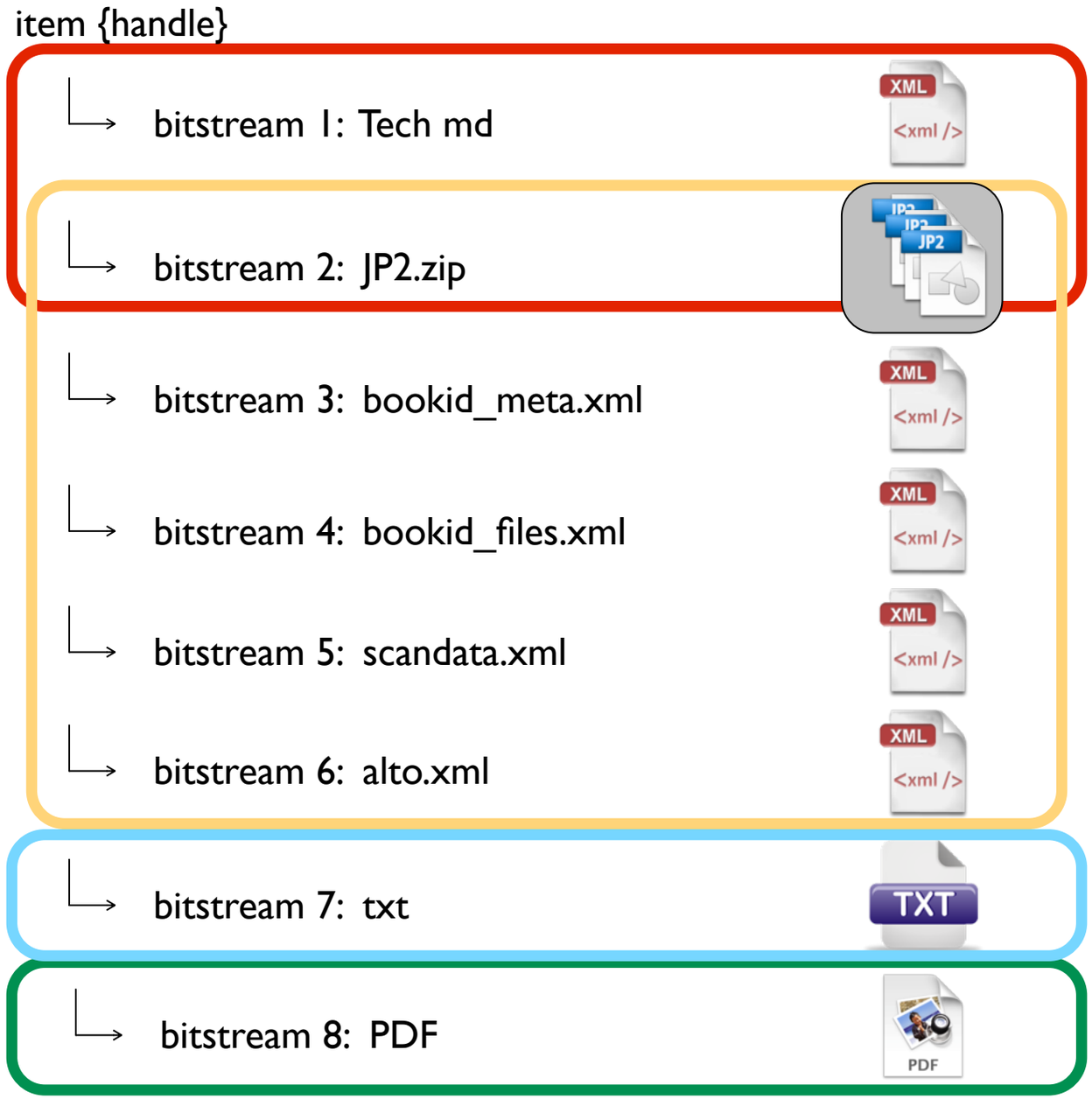
Intellectual entity and representations

Representation of an IE



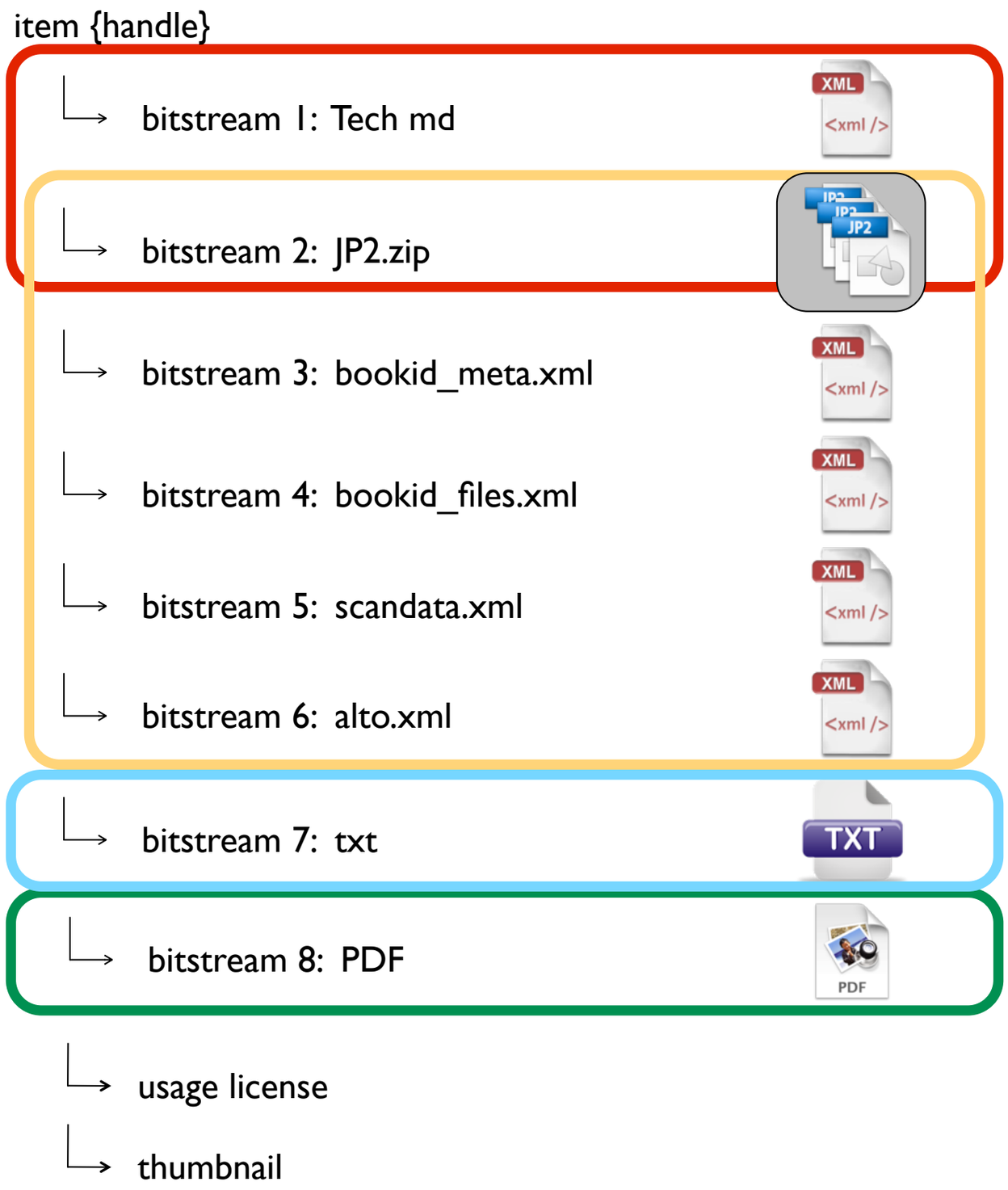
Intellectual entity and representations

Representation of an IE



Intellectual entity and representations

Representation of an IE



IE and representations of a thesis

Intellectual entity

- the thesis, with some descriptive metadata

3 representations of type “**submission**”

- digitization of paper thesis
- electronic version of thesis submitted by PhD student
- official university version

IE and representations of a thesis

5 representations for “**dissemination**” purposes

- Internet Archive BookReader
- raw txt
- ALTO xml
- generated PDF/A derived from paper digitization
- PDF/A version of whatever files submitted by PhD student + signed license (PDF/A file)

IE and representations of a thesis

1 representation for “**archival**“ purposes:

- tech MD + jp2.zip

- minimal information needed permitting us to generate all dissemination representations f an intellectual entity

IE and representations of a thesis

In total for 1 thesis

- 1 intellectual entity
- 1 descriptive metadata for the IE
- 3 submission representations
- 5 dissemination representations
- 1 archival representation

General problem statement

Any **intellectual entity** can be represented in many ways.

Each of these **representations** can be thought of as

- A set of files
- Relationships between these files
- Metadata about these files
  - descriptive, technical, digital provenance, rights
- Metadata about the intellectual entity itself
- Relationships with other representations

3 **types of representations** can be distinguished

- submission, dissemination, archival

How do we store all this information in our repository?

Supplementary datafiles

Files that come with the thesis, report, article,... any publication

- citable
- reuse in other publications
- are themselves (very) complex objects

It could therefore be more appropriate in some cases to handle these supplementary files as distinct intellectual entities

Relationships with other IEs to be expressed through e.g. **`<mods:relatedItem>`**

METS: the solution to describe an IE

METS is an XML container with different sections

- structMap: structure of representation
- fileSec: list of the content files
- amdSec:
  - techMD: PREMIS, MIX, TextMD, VideoMD, AudioMD
  - digiprovMD: PREMIS
  - rightsMD
- dmdSec: descriptive MD of the IE (MODS)

METS: the solution to describe an IE

```
<mets>
   <amdSec>
        <techMD>
                    - PREMIS - UUID
                    - Fixity (SHA256)
                    - Size
                    - PRONOM file type
                    - FITS output (Jhove, exiftool, Droid, NLNZ, OIS, ffident)
        </techMD>

        <digiprovMD>
                    PREMIS.EVENT for every action performed by Archivematica:
                    (normalization, antivirus, etc)
        </digiprovMD>
   </amdSec>

    <fileSec>
        <file> URL to content file </file>
    </fileSec>

    <structMap>
        <div> fptr </div>
    </structMap>
</mets>
```

METS profiles

METS very powerful descriptive tool and therefore generic.

Especially for interoperability, we need to agree on vocabularies, agree on how to describe specific document types (e.g. phd theses)...

Several profiles have been developed and are registered at LC. One of these suits our needs for a generic description model for repository content: agnostic of specific document types.

ECHO Dep Generic METS Profile for Preservation and Digital Repository Interoperability (*University of Illinois at Urbana-Champaign, Grainger Engineering Library Information Center*)

http://www.loc.gov/standards/mets/profiles/00000015.html

Other profiles could be extensions of this one, in order to target specific document types.

Our METS profile

Being inspired by this profile, we propose:

All representations of one IE are described in one METS document

1 representation = 1 structMap

structMap “type” xml attribute specifies the purpose of the representation: submission, dissemination, archival

Our METS profile

structMap is made up of elements, representing hierarchical structure of the representation

The “order” attribute describes the sequence of the structural elements of the representation

The “type” attribute specifies the purpose of the structural element (e.g. ToC, cover page, chapter, body of text, ...)

administrative (amdSec) and descriptive (dmdSec) metadata are attached to any and any content file through {ID, IDREF} correlations

Our METS profile

- taxonomy for “use”, “type” attributes, and “premis.event”, “premis.agent”, … - various proposals in METS profiles

- examples METS profiles:

Australian METS profile 1.0 (National Library of Australia) http://www.loc.gov/standards/mets/profiles/00000018.html

San Diego Complex Object METS profile http://www.loc.gov/standards/mets/profiles/00000028.html

! Need for standardization

Representations in METS

METS document

amdSec

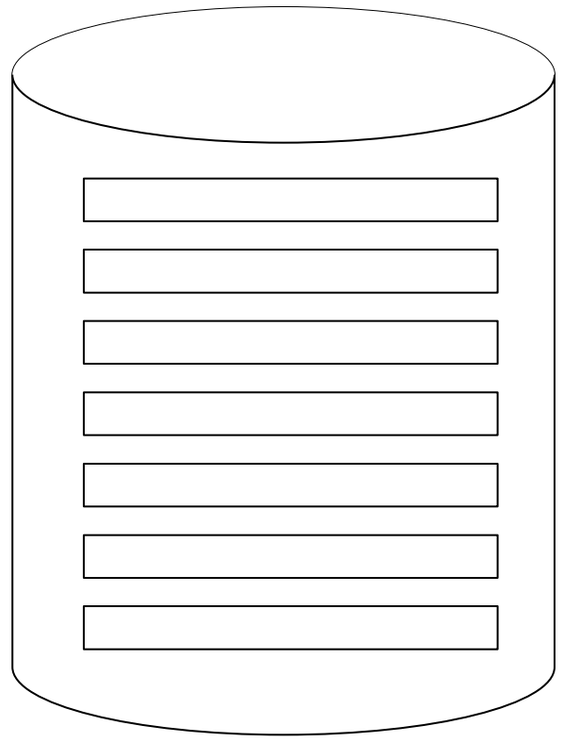
fileSec

structMap

dmdSec- IE descriptive md

- Title
- Authors
- Publication status
- Dates
- Keywords

Content files



Representations in METS

METS document

amdSec

fileSec

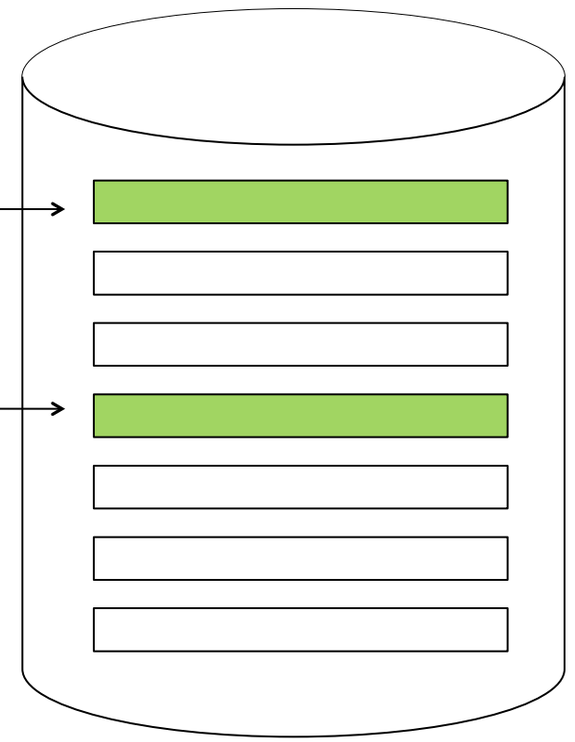
structMap

S

IE descriptive md

- Title
- Authors
- Publication status
- Dates
- Keywords

Content files



Representations in METS

METS document

amdSec

fileSec

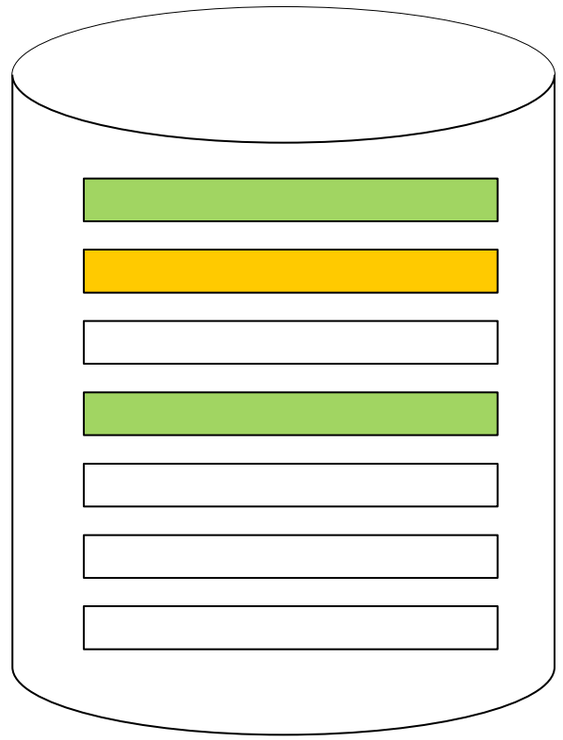
structMap

D
S

IE descriptive md

- Title
- Authors
- Publication status
- Dates
- Keywords

Content files



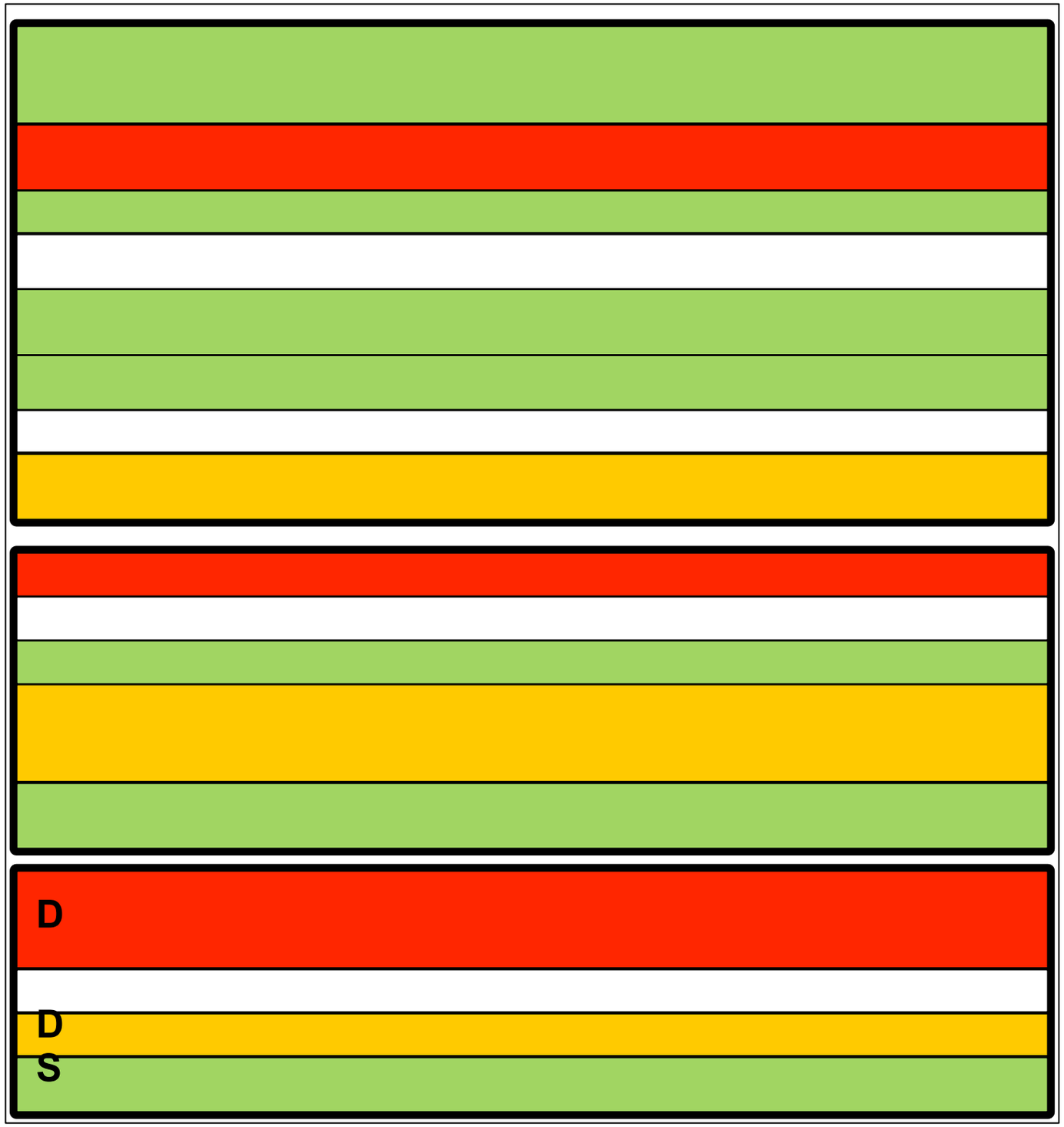
Representations in METS

METS document

amdSec

fileSec

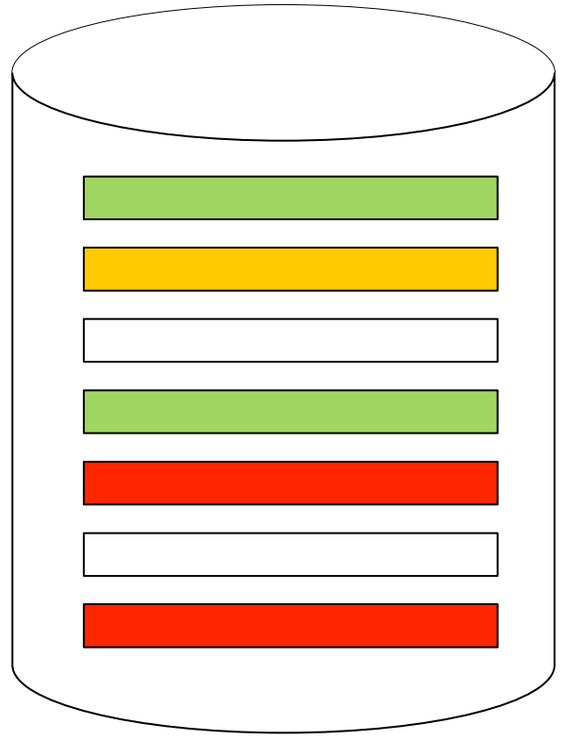
structMap



IE descriptive md

- Title
- Authors
- Publication status
- Dates
- Keywords

Content files



Representations in METS

METS document

amdSec

fileSec

structMap

D

D

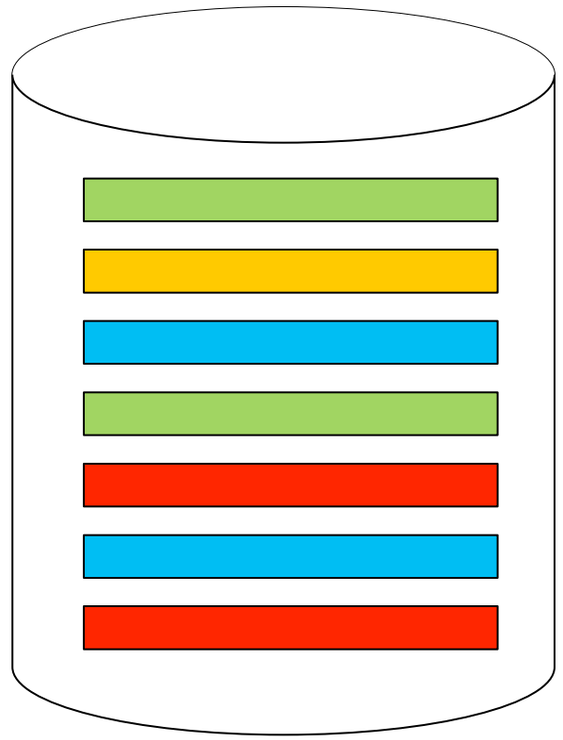
D

S

IE descriptive md

- Title
- Authors
- Publication status
- Dates
- Keywords

Content files



All parts of a representation are linked together through ID,IDREF pairs

Simple example: technical metadata attached to some file

```
<structMap TYPE="dissemination" >
        <div TYPE="coverpage"
ADMID="adm-1">
                <fptr ADMID="adm-2">
                        <Flocat
xlink:href="…">
<amdSec>
        <techMD ID="adm-2">
                PREMIS:object
```

Mapping to DSpace

- 1 intellectual entity = 1 DSpace item
- descriptive metadata of the intellectual entity is maintained in the SQL database
- exactly 1 bundle is attached to this item:
  - 1 bitstream for the METS document
  - as many bitstreams as there are content files

We do not use the DSpace ‘bundle’ hierarchical structure to express structure of the digital object: all of this sits in the METS document

Mapping to DSpace

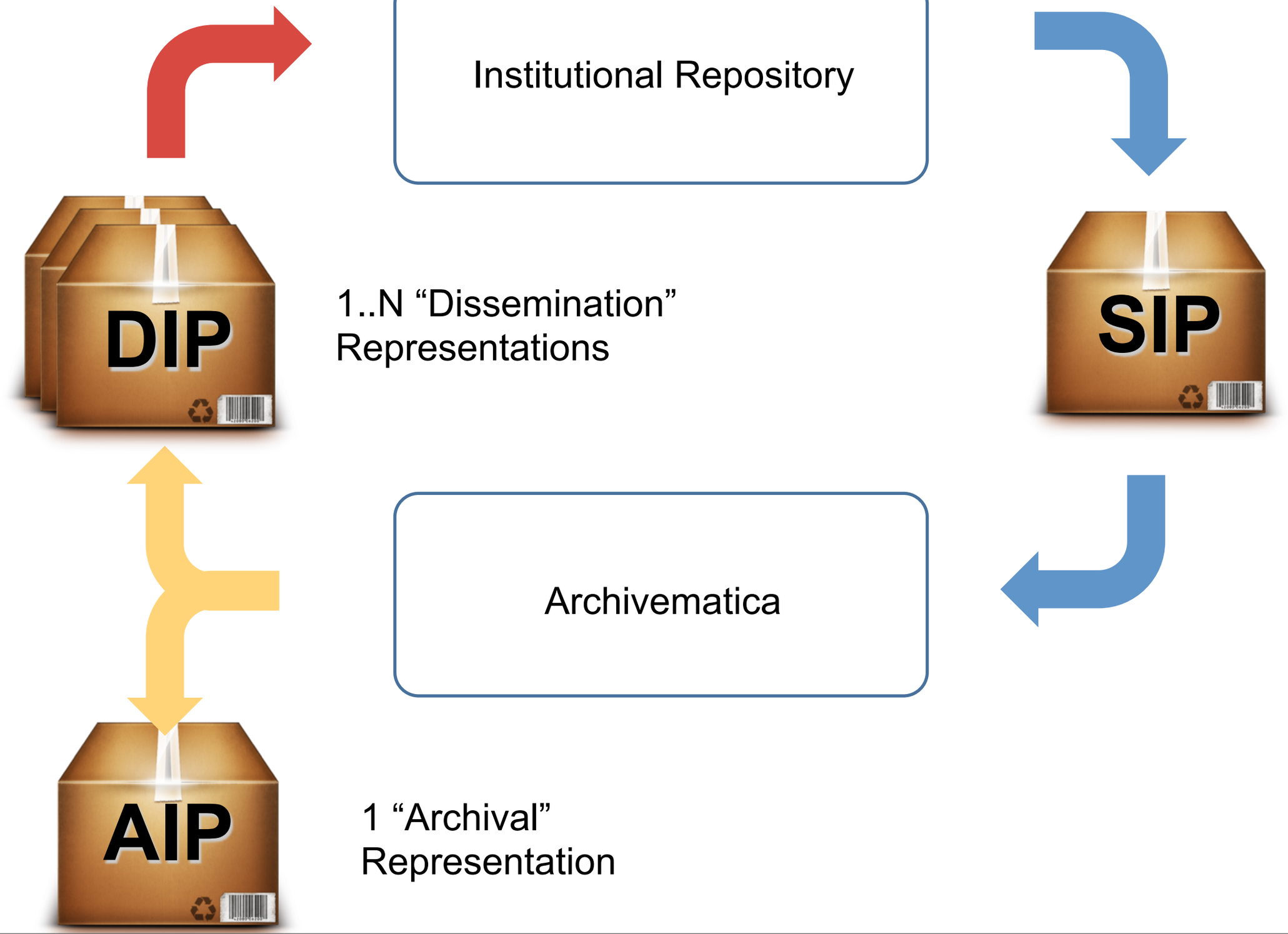
- All identifiers in the METS document should be based on the handle of the item

  !! All identifiers in METS document must be unique

- The fptr in of the content files is an URL to the bitstream in DSpace

```
<fptr ADMID="adm-2">
   <Flocat
         xlink:href="https://dipot.ulb.ac.be/dspace/bitstream/2013/103504/1/
                        duvosquel.pdf">
```

Import and export of representations

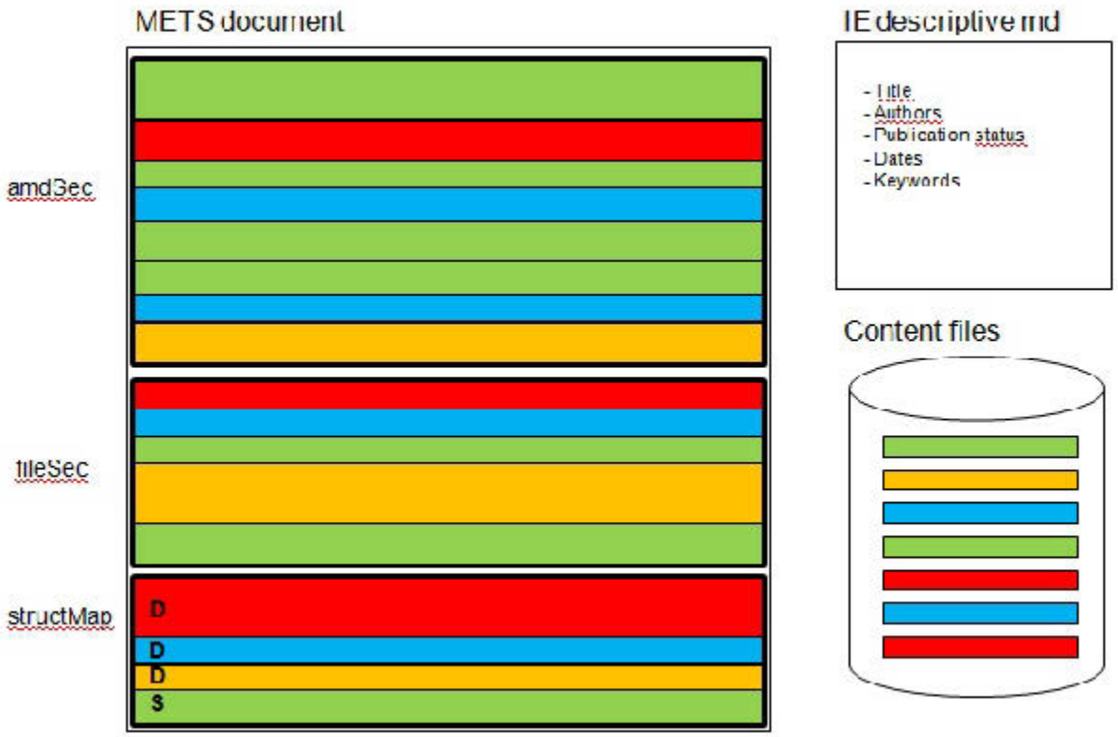


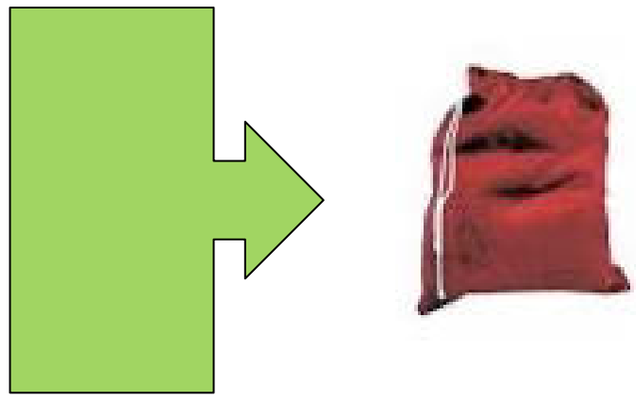
Export as SIP for Archivematica

Make a new METS document containing all relevant information from the original METS documents:

1. Select all “submission” structMaps and follow identifiers
2. get descriptive metadata from DSpace SQL-database, map to MODS and add as dmdSec to the new METS document
3. get appropriate content files together and create a new BagIt of the content files and the new METS document

Export as SIP for Archivematica





Archivematica

Import a DIP generated by Archivematica

- open the BagIt that Archivematica delivers

- extract content files from the AM BagIt and generate DSpace bitstreams (generating bitstream IDs)

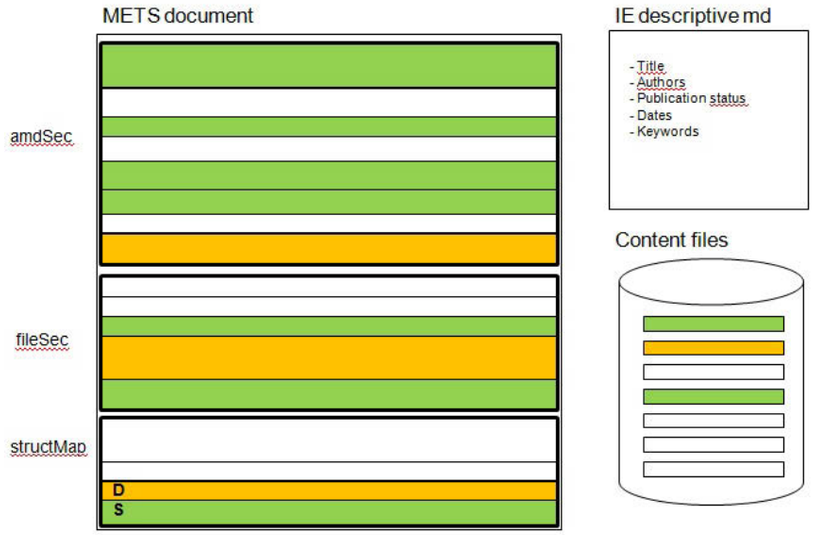
- filepaths in the AM structMap must be replaced by URLs to the created bitstreams

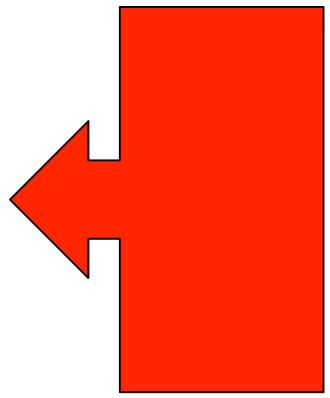
- extract the relevant parts from the AM METS document (structMap, fileSec, amdSec)

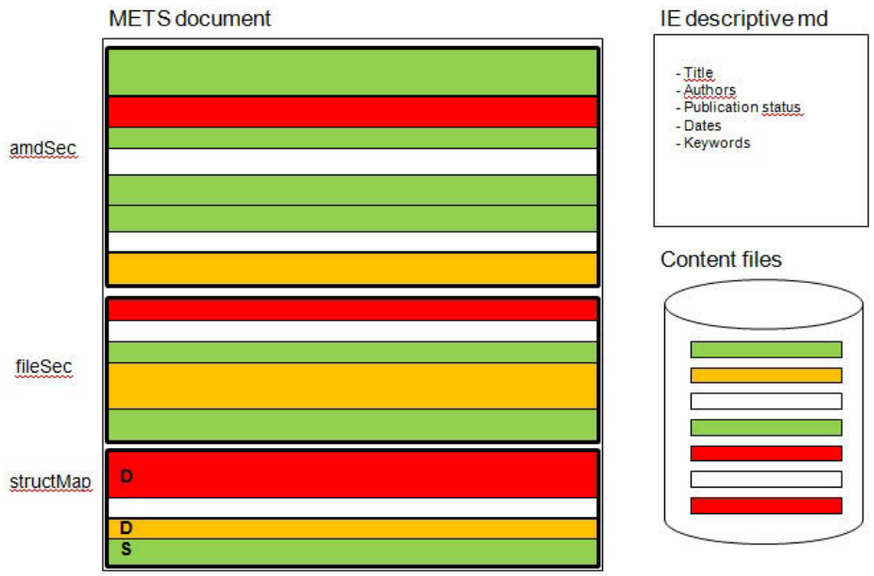
- reassign all identifiers based on the Dspace item handle

- insert all these xml sections to the METS document for the intellectual entity

Import a DIP generated by Archivematica







Archivematica

Export a DIP to DI-fusion

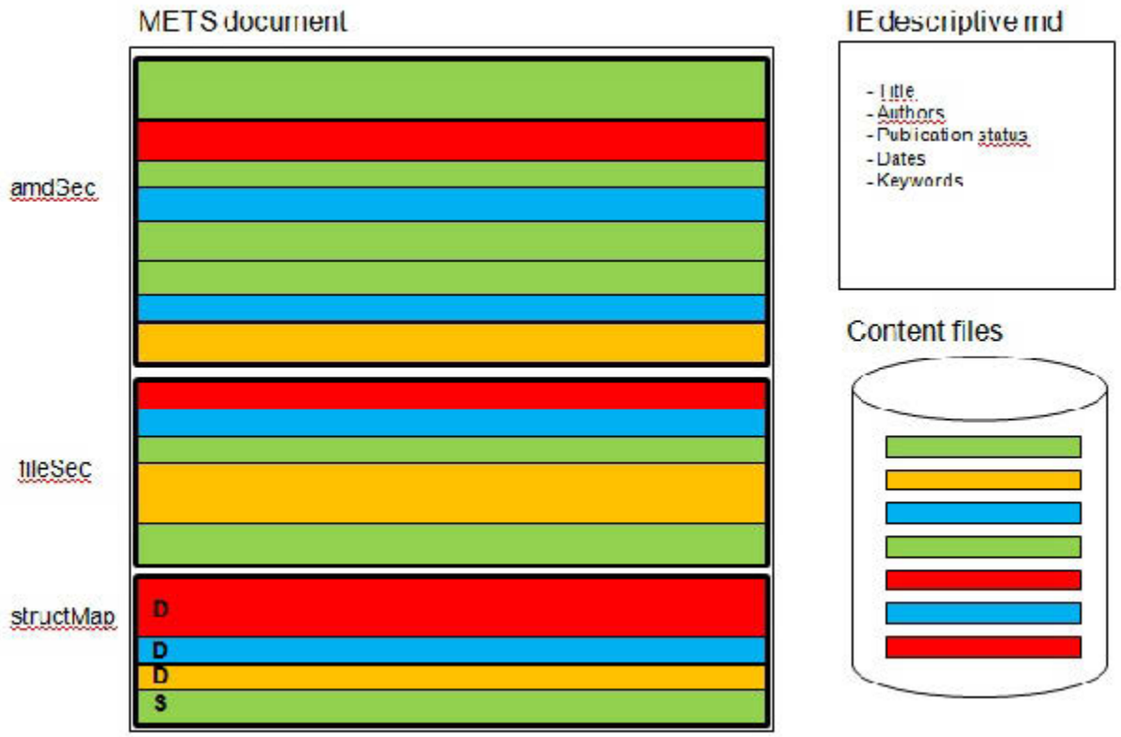
Make a new METS document containing all relevant information from the original METS document:

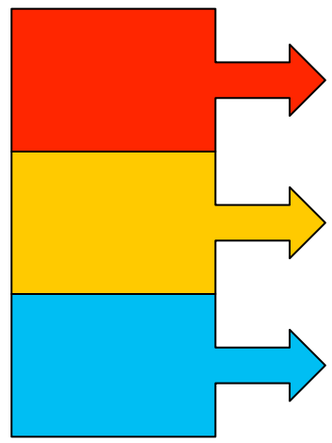
1. select all “dissemination” structMaps, follow identifiers, and get all relevant amdSec and fileSec sections together

2. get descriptive metadata from DSpace SQL database, map to MODS and add as dmdSec to the new METS document

Deliver this new METS XML document to DI-fusion

Export a DIP to DI-fusion





METS

DI-fusion

OAI-PMH

Import AIP from Archivematica

- Archivematica generates AIP for the intellectual entity

- This AIP gets pushed to our **preservation repository** (listen to Anthony in 2 minutes)

- we keep track of this event in the repository as an archival representation:

  - file pointer in the structMap is an URL to the Preservation Repository

What has to be changed?

New high-volume digitization workflow

New description model for all digital objects

A preservation strategy for all digital objects

LEARNING RESOURCES

University Libraries have two main **missions**

To guarantee access to objects
selected by curators

To preserve those objects
especially our own production

University Libraries have two main **missions**

To guarantee access to objects
selected by curators

To preserve those objects
especially our own production

In the digital era, those missions are compromised:

- we have lost access control on some digital objects (journal access via subscription)
- the vulnerability of digital objects

Multiple dangers threaten our archives

Natural disasters

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Internal or external attacks

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Internal or external attacks

- **Authentification**

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Internal or external attacks

- **Authentification**

Human failure

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Internal or external attacks

- **Authentification**

Human failure

- **Independent site technical admin**

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Internal or external attacks

- **Authentification**

Human failure

- **Independent site technical admin**

Economic breakdown

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Internal or external attacks

- **Authentification**

Human failure

- **Independent site technical admin**

Economic breakdown

- **Cost control**

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Storage media failure

Internal or external attacks

- **Authentification**

Human failure

- **Independent site technical admin**

Economic breakdown

- **Cost control**

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Storage media failure

- **Data monitoring**

Internal or external attacks

- **Authentification**

Human failure

- **Independent site technical admin**

Economic breakdown

- **Cost control**

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Storage media failure

- **Data monitoring**

Internal or external attacks

- **Authentification**

Media obsolescence

Human failure

- **Independent site technical admin**

Economic breakdown

- **Cost control**

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Storage media failure

- **Data monitoring**

Internal or external attacks

- **Authentification**

Media obsolescence

- **Media migration**

Human failure

- **Independent site technical admin**

Economic breakdown

- **Cost control**

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Storage media failure

- **Data monitoring**

Internal or external attacks

- **Authentification**

Media obsolescence

- **Media migration**

Human failure

- **Independent site technical admin**

Format obsolescence

Economic breakdown

- **Cost control**

Multiple dangers threaten our archives

Natural disasters

► **Geo-replication (3f+1)**

Storage media failure

► **Data monitoring**

Internal or external attacks

► **Authentification**

Media obsolescence

► **Media migration**

Human failure

► **Independent site technical admin**

Format obsolescence

► **Format migration**

Economic breakdown

► **Cost control**

Multiple dangers threaten our archives

Natural disasters

► **Geo-replication (3f+1)**

Storage media failure

► **Data monitoring**

Internal or external attacks

► **Authentification**

Media obsolescence

► **Media migration**

Human failure

► **Independent site technical admin**

Format obsolescence

► **Format migration**

Economic breakdown

► **Cost control**

Management issues

Multiple dangers threaten our archives

Natural disasters

- **Geo-replication (3f+1)**

Storage media failure

- **Data monitoring**

Internal or external attacks

- **Authentification**

Media obsolescence

- **Media migration**

Human failure

- **Independent site technical admin**

Format obsolescence

- **Format migration**

Economic breakdown

- **Cost control**

Management issues

- **Independent site administrations**

The preservation solution will have to integrate all those aspects

Authentification

Format Migration

Independent Administration

Independent Tech Management

Media Migration

Geo-replication

Data Monitoring

Cost control

Three types of bit-level preservation solutions

Tessella
Digital Preservation

DURACLOUD™

LOCKSS

third-party integrated solutions

cloud storage

distributed preservation network

Control of preservation

LEARNING RESOURCES

We want to be “active players of preservation”, not “passive clients of third-party preservation services“ [Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

Technical issues:

- No control on the archiving technical policy
- Is migration to another provider possible?

We want to be “active players of preservation”, not “passive clients of third-party preservation services“

[Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

Technical issues:

- No control on the archiving technical policy
- Is migration to another provider possible?

No more technical problems...

at the cost of many legal and economical issues

By outsourcing one of two main missions, we would become **brokers** in preservation and not **curators** anymore.

By outsourcing one of two main missions, we would become **brokers** in preservation and not **curators** anymore.

By outsourcing one of two main missions, we would become **brokers** in preservation and not **curators** anymore.

How can we be **actors** of the preservation of our digital objects and not passive clients of third-party preservation services ?

Ok... But can we handle preservation ourselves?

Yes we can. But not on our own!

Private LOCKSS Networks: community-based distributed preservation network

Private LOCKSS Networks: community-based distributed preservation network

http://difusion.academiewb.be/

Private LOCKSS Networks: community-based distributed preservation network

http://difusion.academiewb.be/

geo-replication (100km, safe place)

Private LOCKSS Networks: community-based distributed preservation network

http://difusion.academiewb.be/

geo-replication (100km, safe place)
+7 nodes (Byzantine fault tolerance)

Private LOCKSS Networks: community-based distributed preservation network

http://difusion.academiewb.be/

geo-replication (100km, safe place)
+7 nodes (Byzantine fault tolerance)
low-cost hardware

Private LOCKSS Networks: community-based distributed preservation network

http://difusion.academiewb.be/

geo-replication (100km, safe place)
+7 nodes (Byzantine fault tolerance)
low-cost hardware
low TCO over the long term

Private LOCKSS Networks: community-based distributed preservation network

http://difusion.academiewb.be/

geo-replication (100km, safe place)
+7 nodes (Byzantine fault tolerance)
low-cost hardware
low TCO over the long term
regular monitoring of archives integrity

Private LOCKSS Networks: community-based distributed preservation network

http://difusion.academiewb.be/

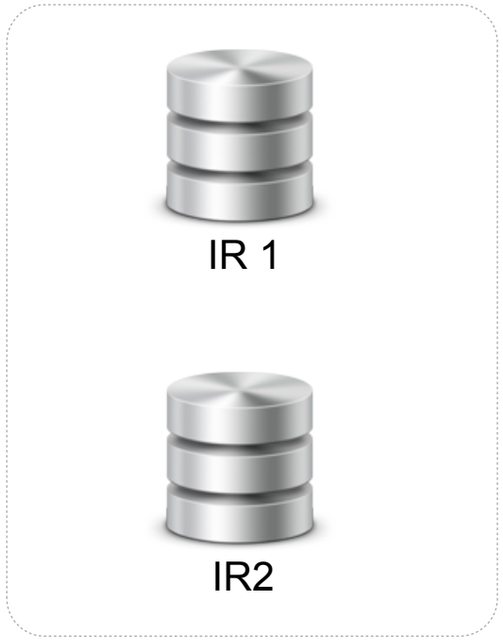
geo-replication (100km, safe place)
+7 nodes (Byzantine fault tolerance)
low-cost hardware
low TCO over the long term
regular monitoring of archives integrity
no long-term secrets, rate-limit changes

Private LOCKSS Networks: community-based distributed preservation network

http://difusion.academiewb.be/

geo-replication (100km, safe place)
+7 nodes (Byzantine fault tolerance)
low-cost hardware
low TCO over the long term
regular monitoring of archives integrity
no long-term secrets, rate-limit changes
10 active PLNs accross the world

How are we planning to perform preservation?



Dissemination repositories

How are we planning to perform preservation?

IR 1

IR2

Dissemination repositories

Back up
(EqualLogic)

How are we planning to perform preservation?

IR 1

IR2

Dissemination repositories

Back up
(EqualLogic)

LOCKSS boxes

x 7

Offline boxes

x 2

Preservation repositories

How are we planning to perform preservation?

IR 1

IR2

Dissemination repositories

manifest
IR1

manifest
IR2

Archival Unit Publisher

LOCKSS boxes

x 7

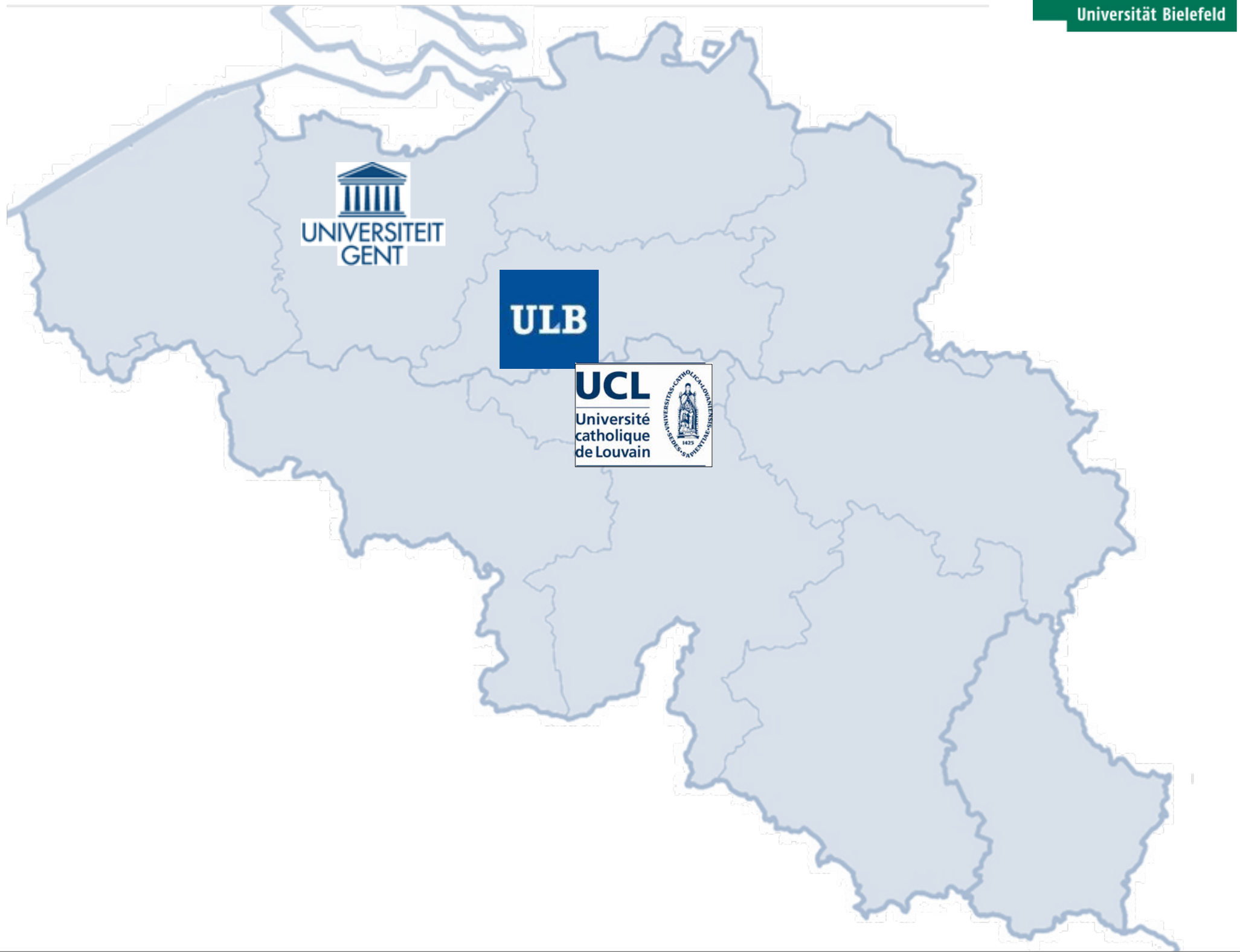
Offline boxes

x 2

Preservation repositories

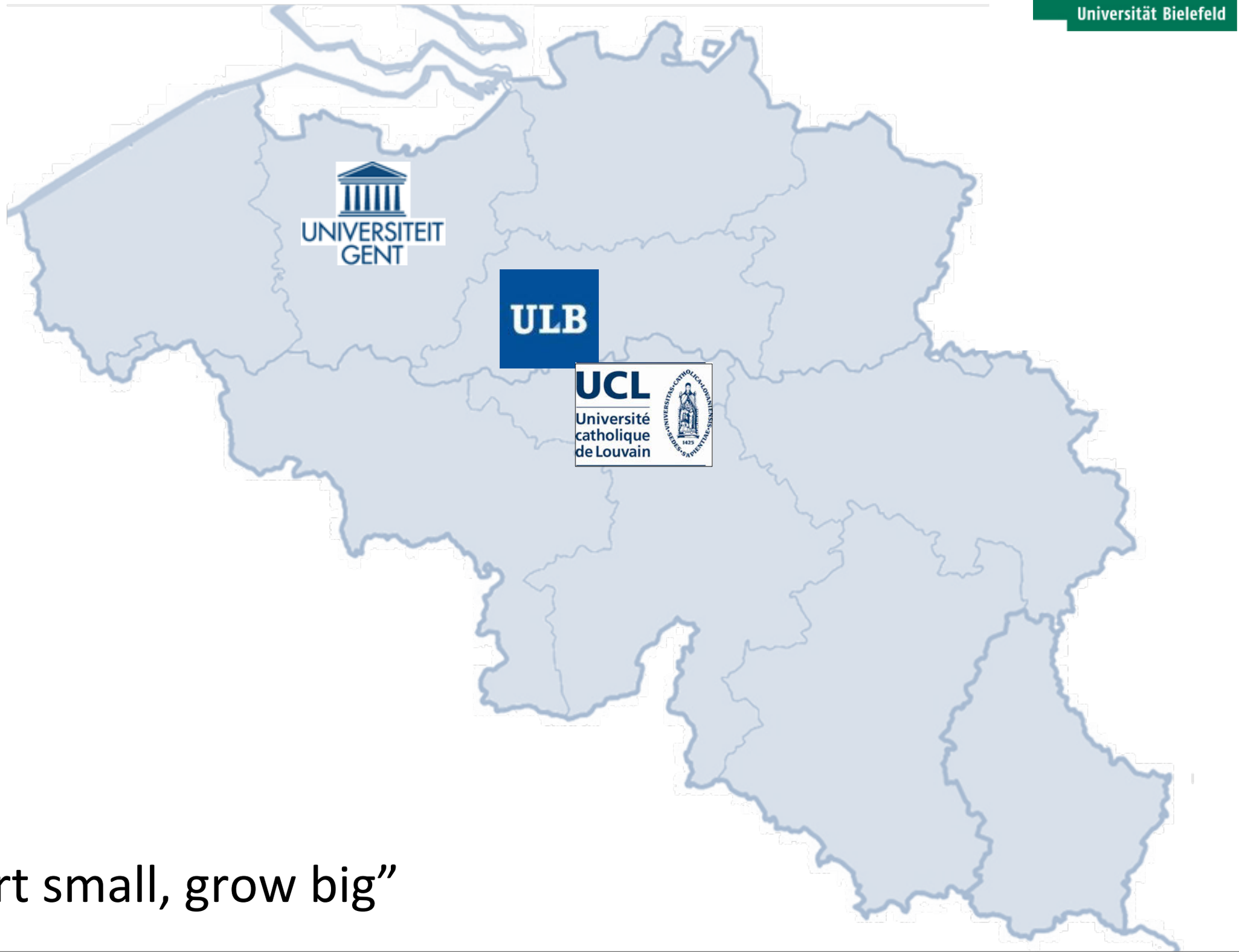
Back up
(EqualLogic)

Who is currently involved?



LEARNING RESOURCES

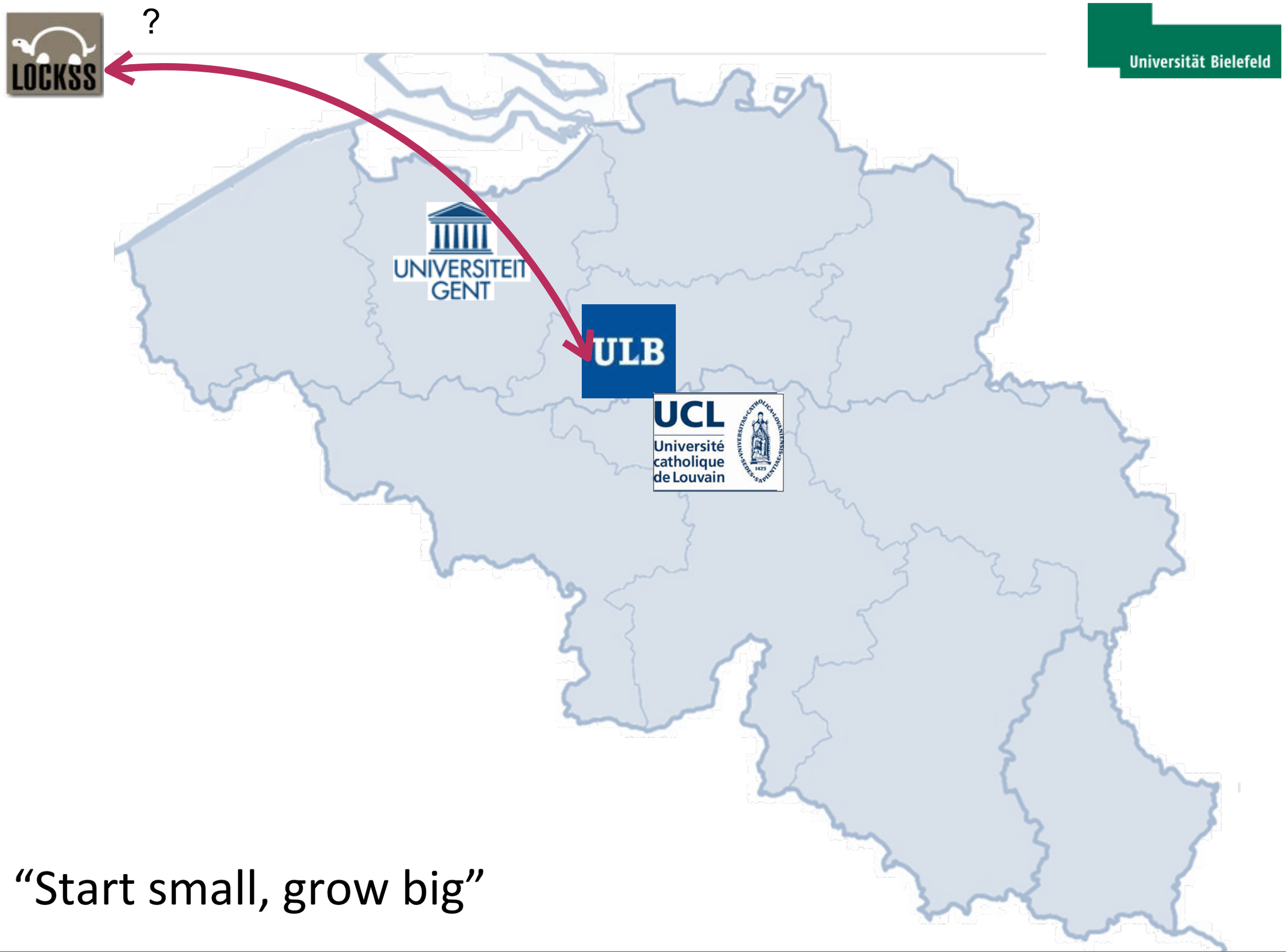
Who is currently involved?



"Start small, grow big"

Who is currently involved?

?



“Start small, grow big”

Belgium offers attractive assets for international partners

Belgium offers attractive assets for international partners

Belgium offers attractive assets for international partners

Belgium offers attractive assets for international partners

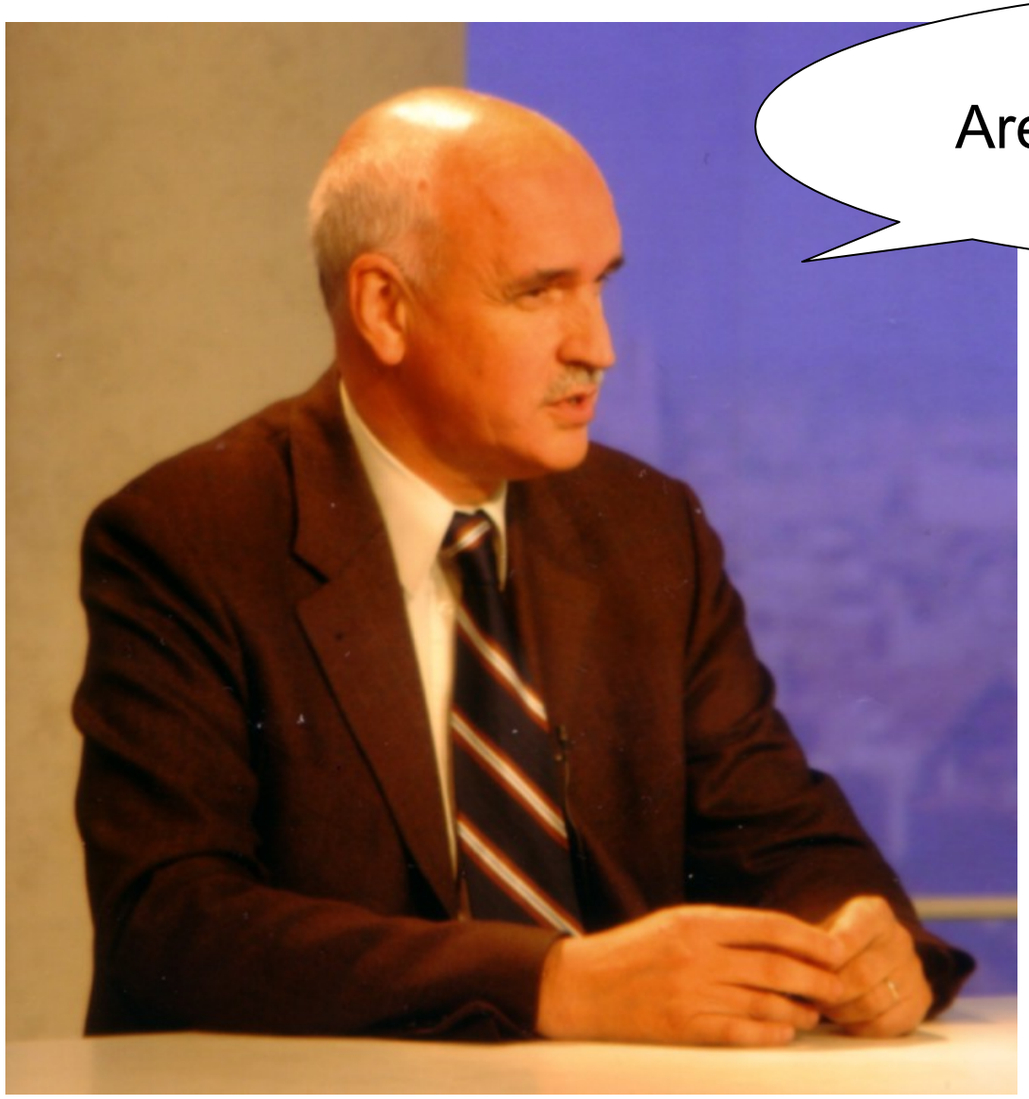
Belgium offers attractive assets for international partners

Belgium offers attractive assets for international partners

Belgium offers attractive assets for international partners

Conclusion

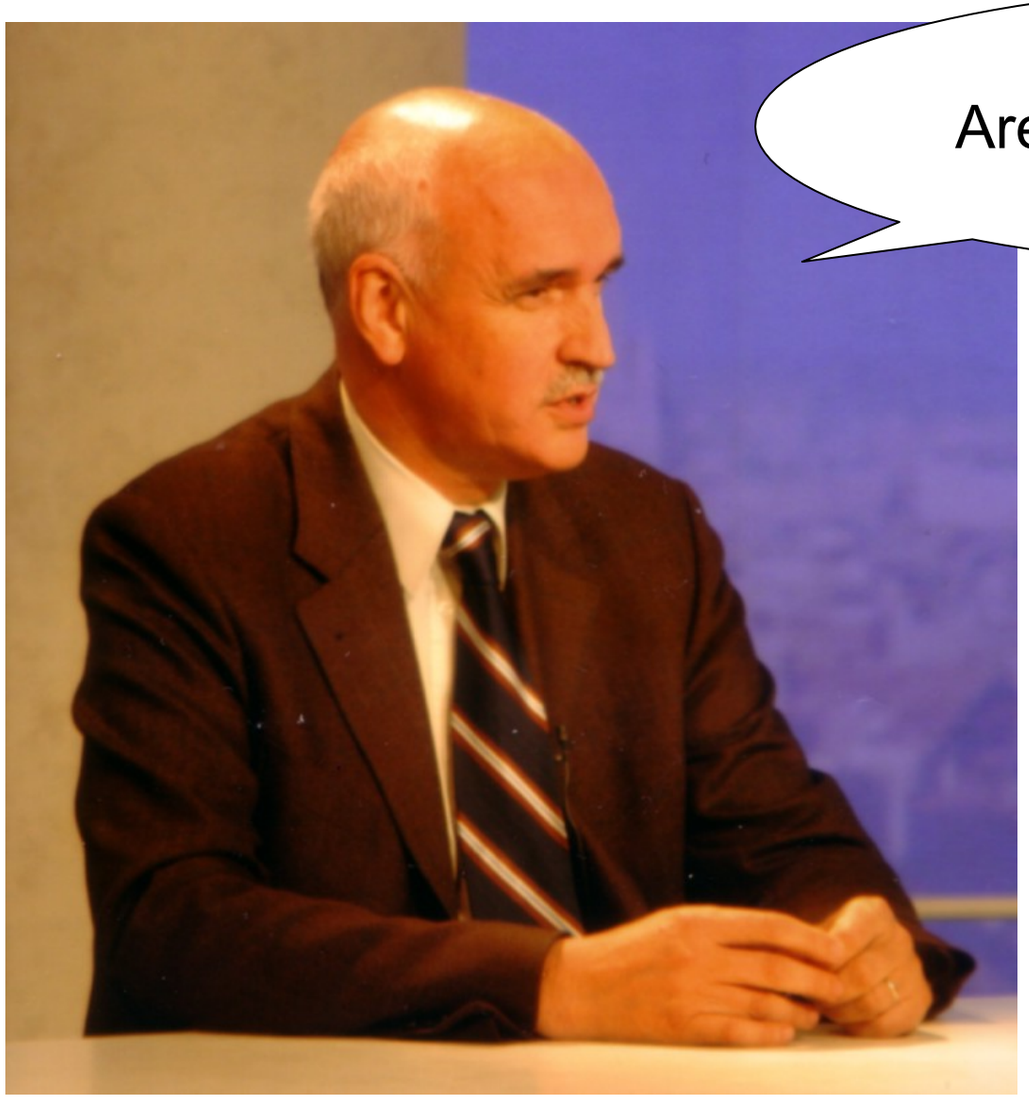
Conclusion



Are we done yet?

Well, not yet... But so far we have:

Conclusion

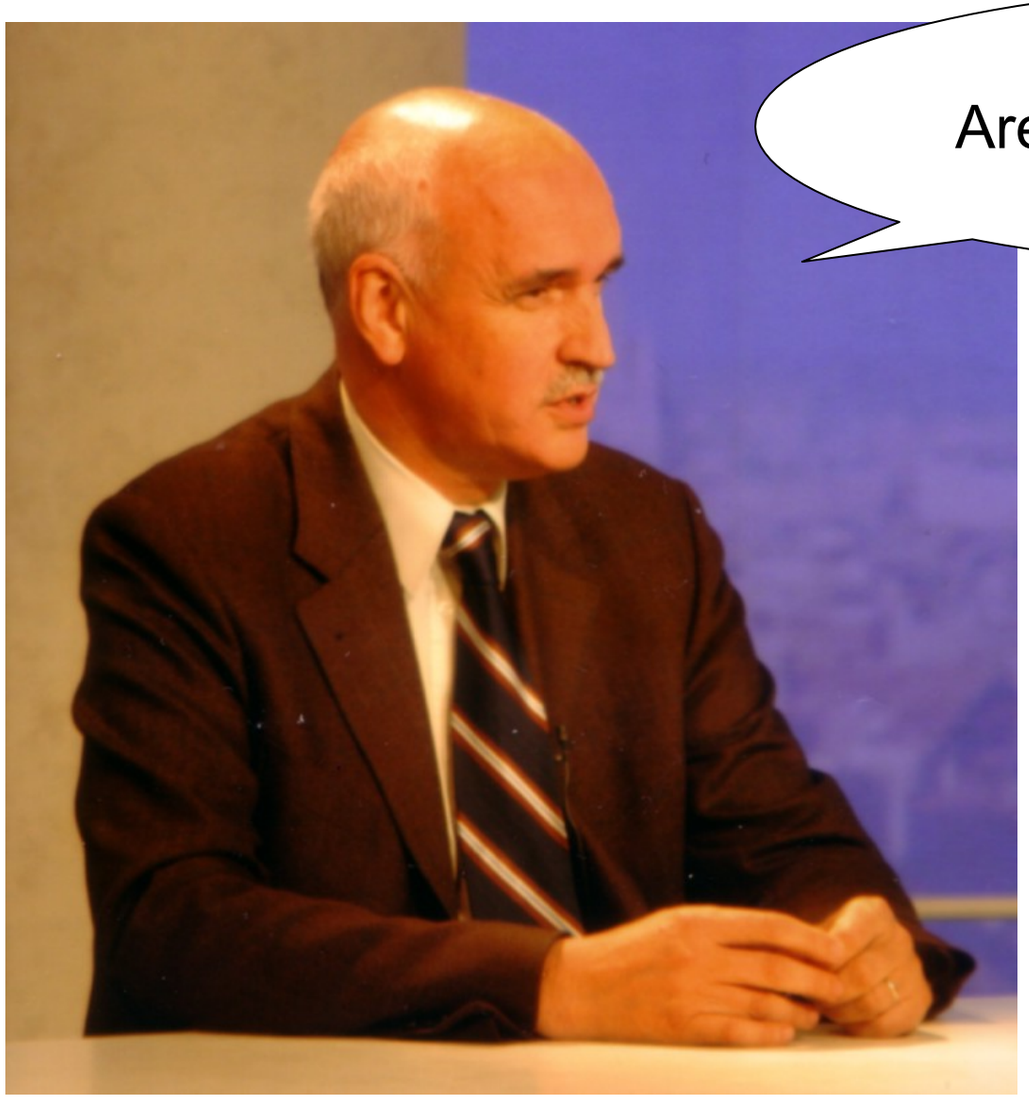


Are we done yet?

Well, not yet... But so far we have:

- an efficient high-volume digitization workflow

Conclusion

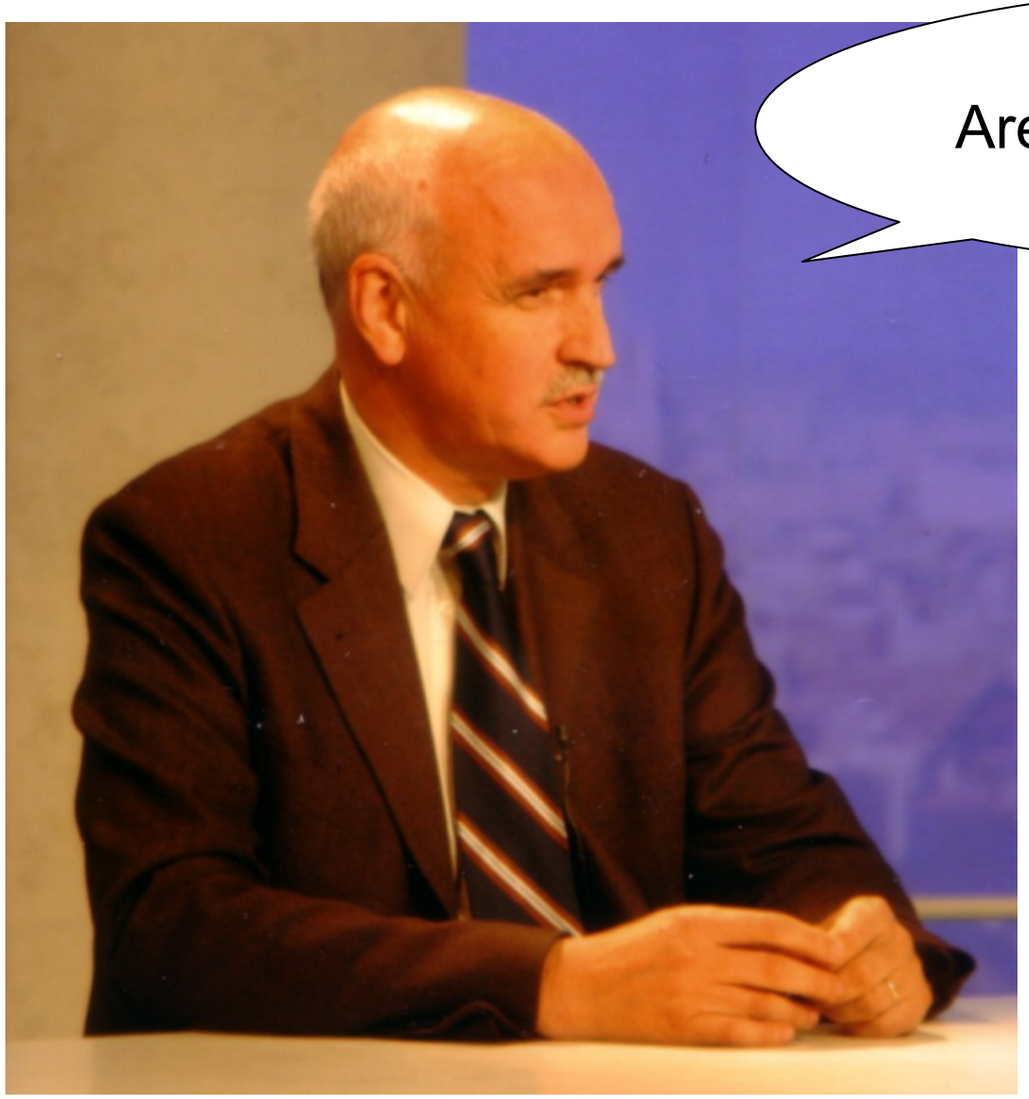


Are we done yet?

Well, not yet... But so far we have:

- an efficient high-volume digitization workflow
- a description model for our complex objects

Conclusion



Are we done yet?

Well, not yet... But so far we have:

- an efficient high-volume digitization workflow

- a description model for our complex objects

- a distributed bit-level preservation solution, possibly with new collaborations after OR2013

LEARNING RESOURCES

JPEG2000 a été adopté comme format de préservation par de nombreuses institutions

KB
Koninklijke Bibliotheek

HARVARD UNIVERSITY LIBRARY

BRITISH LIBRARY

Smithsonian Libraries

Library and Archives Canada

nb.no

INTERNET ARCHIVE

Structure of an AIP

AIP-[UUID] - BagIt

- bagit.txt
- bag-info.txt
- tagmanifest-md5.txt
- manifest-sha512.txt
- data/
  - METS.[UUID].xml
  - logs/
  - metadata/
  - objects/

Structure of an AIP

AIP-[UUID] - BagIt

- bagit.txt
- bag-info.txt
- tagmanifest-md5.txt
- manifest-sha512.txt
- data/
  - METS.[UUID].xml
  - logs/
  - metadata/
  - objects/

bagit.txt

```
BagIt-version: 0.97
Tag-File-Character-Encoding: UTF-8
```

Structure of an AIP

AIP-[UUID] - BagIt

- bagit.txt
- bag-info.txt
- tagmanifest-md5.txt
- manifest-sha512.txt
- data/
  - METS.[UUID].xml
  - logs/
  - metadata/
  - objects/

bag-info.txt

```
Payload-Oxum: 10419323.53
Bagging-Date: 2012-07-13
Bag-Size: 9.9 MB
```

Structure of an AIP

AIP-[UUID] - BagIt

- bagit.txt
- bag-info.txt
- tagmanifest-md5.txt
- manifest-sha512.txt
- data/
  - METS.[UUID].xml
  - logs/
  - metadata/
  - objects/

tagmanifest-md5.txt

```
[MD5hash]   bagit.txt
[MD5hash]   bag-info.txt
[MD5hash]   manifest-sha512.txt
```

manifest-sha512.txt

```
[SHA512hash]   data/METS.[UUID].xml
[SHA512hash]   data/objects/bitstream
[SHA512hash]   ...
```

Structure of an AIP

METS.[UUID].xml

```
<mets xsi:schemaLocation:"http//www.loc.gov/METS/...">
  - <amdSec ID = "amdSec_1">
     - <techMD ID = "techMD_1">
          - PREMIS - UUID
          - Fixity (SHA256)
          - Size
          - PRONOM file type
          - FITS output (Jhove, exiftool, Droid, NLNZ, OIS, ffident)
       </techMD>
     - <digiprovMD ID ="digiprovMD_1">
          PREMIS.EVENT for every action performed by Archivematica:
          (normalization, antivirus, etc)
       </digiprovMD>
    </amdSec>

 - <fileSec>
       - fileGrp "original"
          href to original files
       - fileGrp "Preservation"
          href to every bitstream of the object
    </fileSec>
  - <structMap>
          fptr to every files
    </structMap>
</mets>
```

Structure of an AIP

AIP-[UUID] - BagIt

- bagit.txt
- bag-info.txt
- tagmanifest-md5.txt
- manifest-sha512.txt
- data/
  - METS.[UUID].xml
  - logs/
    - metadata/
    - objects/

normalizationLog.txt

```
Log of the output of the tools
used for normalization
```

Structure of an AIP

AIP-[UUID] - BagIt

- bagit.txt
- bag-info.txt
- tagmanifest-md5.txt
- manifest-sha512.txt
- data/
  - METS.[UUID].xml
  - logs/
  - metadata/
  - objects/

metadata/

This folder contains log files from tools used during SIP ingest:

- clamAVScan.txt
- extraction.log
- filenameCleanup.log
- FileUUIDs.log
- METS.xml

Structure of an AIP

AIP-[UUID] - BagIt

- bagit.txt
- bag-info.txt
- tagmanifest-md5.txt
- manifest-sha512.txt
- data/
  - METS.[UUID].xml
  - logs/
  - metadata/
  - objects/

objects/

This folder contains the bitstreams related to the preserved object:

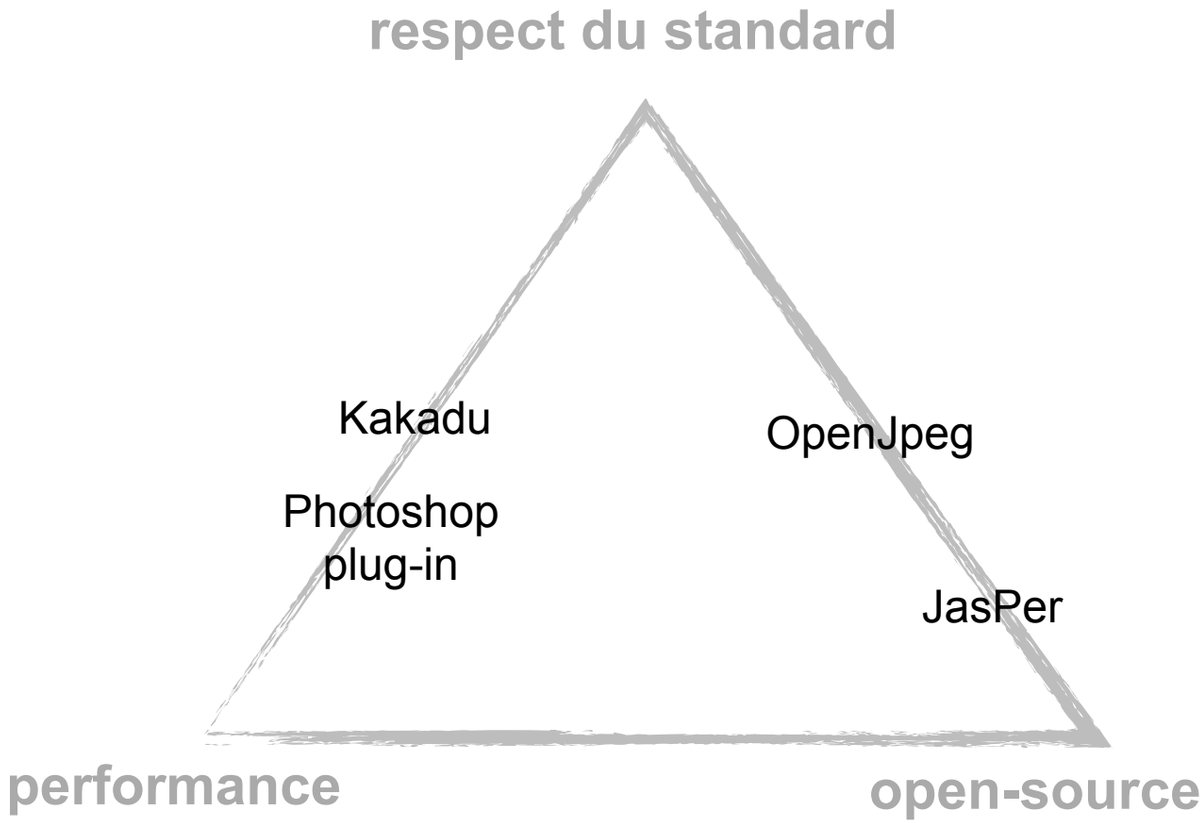
- bitstream_6593.pdf
- bitstream_6595
- bitstream_6903.txt
- mets.xml

JPEG2000 présente toutefois deux inconvénients majeurs

Support limité des browsers

JPEG2000 présente toutefois deux inconvénients majeurs

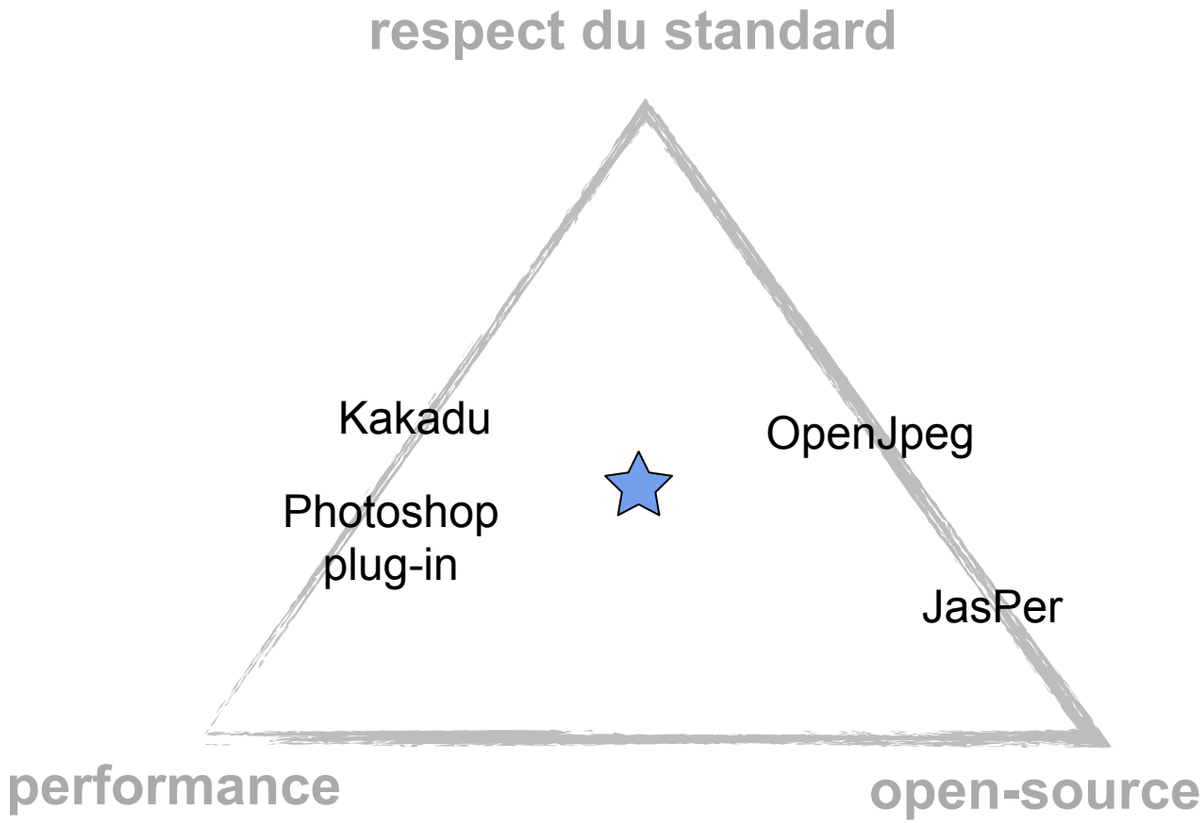
Support limité des browsers



Pas d'implémentation à la fois open-source, standardisée et performante

JPEG2000 présente toutefois deux inconvénients majeurs

Support limité des browsers



Pas d'implémentation à la fois open-source, standardisée et performante

If a *new type* of DIP is needed, it will be built from the AIP extracted from the AU publisher by Archivematica

AU publisher
(wwwarch)

bundle "**Preservation**"

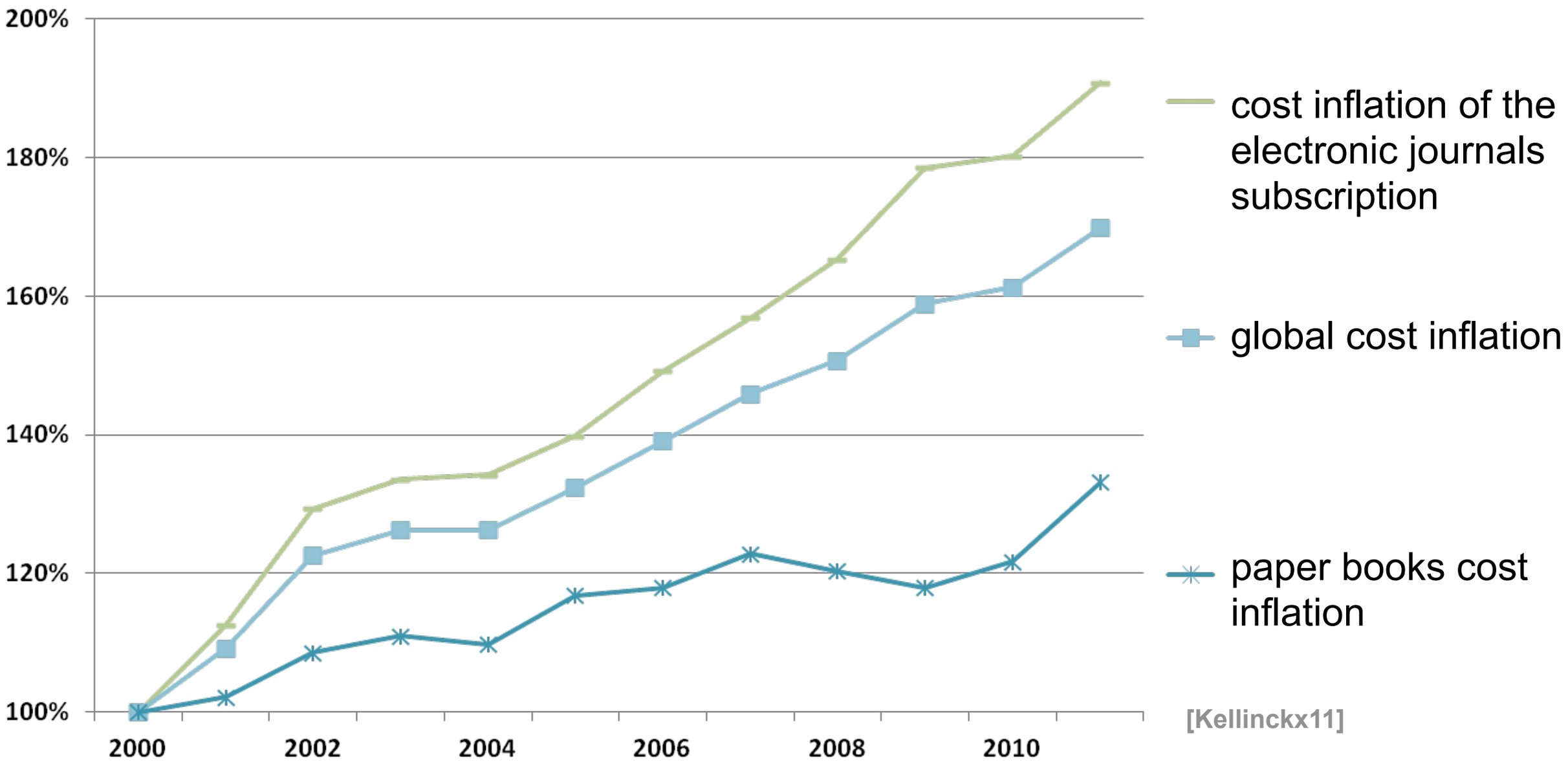
Institutional
Repository

OAI-PMH

SQL
database

New image server
(dishow)

We made a considerable mistake 20 years ago by delegating electronic journal publication to commercial companies: we are still suffering from this decision



90% inflation of the ULB publication subscription budget over the last 11 years

Cloud-based solutions: outsourcing bit-archival to commercial partners

commercial integrated solutions

cloud storage

preservation network

Cloud-based storage is exposed to the same legal and economical issues: average yearly byte cost reduction is 3%

Cloud-based storage is exposed to the same legal and economical issues: average yearly byte cost reduction is 3%

Windows Azure

amazon web services

Google CLOUD STORAGE

rackspace

Cloud cost:
~1000$/TB/year

LEARNING RESOURCES

Cloud-based storage is exposed to the same legal and economical issues: average yearly byte cost reduction is 3%

Windows Azure

amazon web services

Google CLOUD STORAGE

rackspace

DURACLOUD

Cloud cost:
~1000$/TB/year
+ 7000$ (broker)

Cloud-based storage is exposed to the same legal and economical issues: average yearly byte cost reduction is 3%

Windows Azure

amazon web services

Google CLOUD STORAGE

rackspace

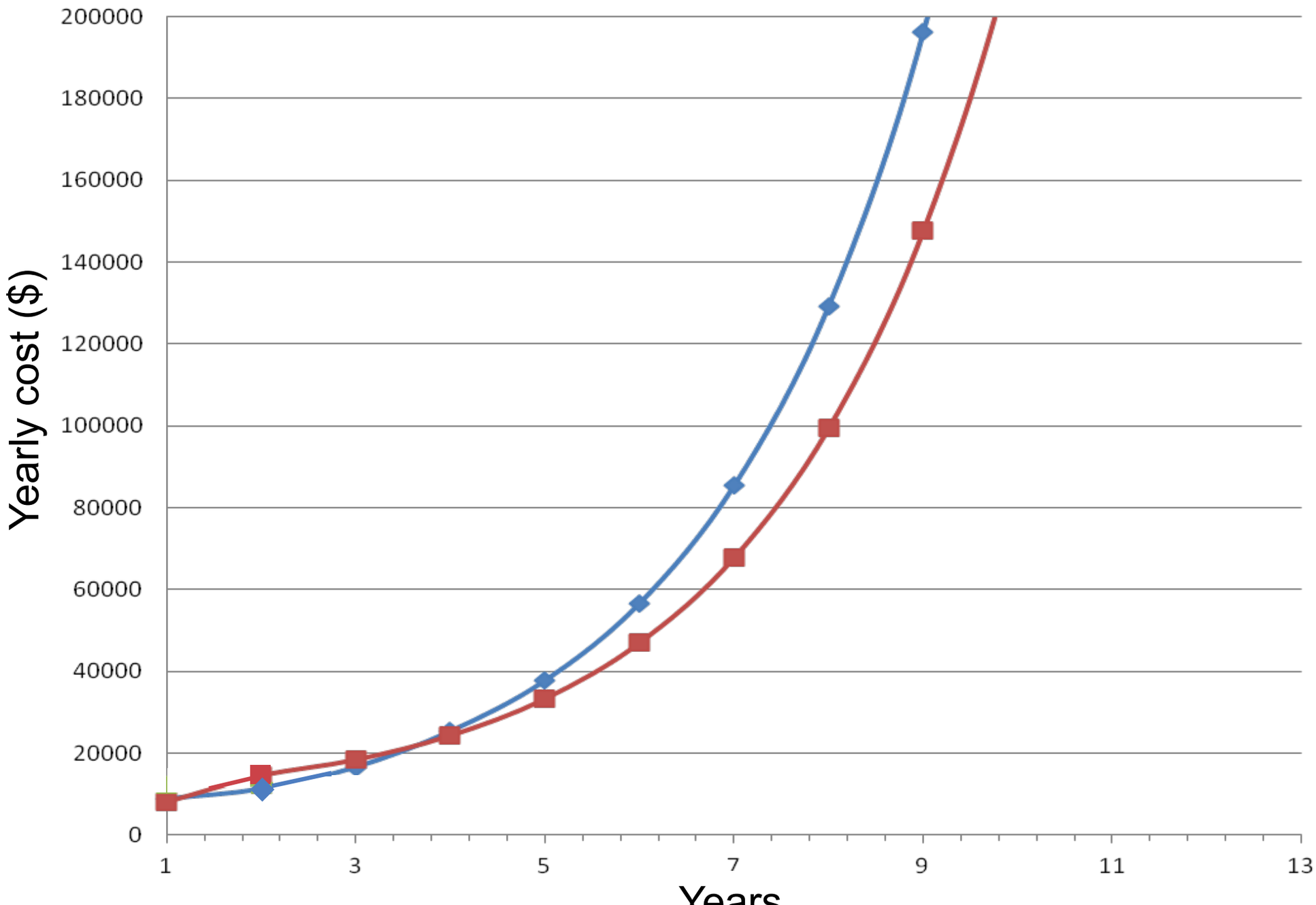
DURACLOUD

Cloud cost:
~1000$/TB/year
+ 7000$ (broker)

Capacity growth:
+60% / year

Cloud-based storage is exposed to the same legal and economical issues: average yearly byte cost reduction is 3%

Windows Azure · amazon web services · Google CLOUD STORAGE · rackspace · DURACLOUD



Cloud cost:
~1000$/TB/year
+ 7000$ (broker)

Capacity growth:
+60% / year

We expect a yearly storage growth rate of **60%** (a rather conservative hypothesis)

What should we preserve?
The ULB case study

What should we preserve? The ULB case study

DI·fusion

Institutional Repository

- Scientific publications
- DSpace
- +78000 references
- +10000 PDF
- 2 GB

What should we preserve? The ULB case study

DI·fusion

Institutional Repository

- Scientific publications
- DSpace
- +78000 references
- +10000 PDF
- 2 GB

Digithèque

- Digitized books and publications
- Symphony
- +220000 digitized pages
- 3TB (raw TIFF) - 1TB (J2K)
- 50 GB PDF files

What should we preserve? The ULB case study

Institutional Repository
- Scientific publications
- DSpace
- +78000 references
- +10000 PDF
- 2 GB

Digithèque
- Digitized books and publications
- Symphony
- +220000 digitized pages
- 3TB (raw TIFF) - 1TB (J2K)
- 50 GB PDF files

BicTel
- Electronic Thesis Dissertations
- ETD Software
- +1250 PDF text
- 1 GB

What should we preserve? The ULB case study

Institutional Repository
- Scientific publications
- DSpace
- +78000 references
- +10000 PDF
- 2 GB

Digithèque
- Digitized books and publications
- Symphony
- +220000 digitized pages
- 3TB (raw TIFF) - 1TB (J2K)
- 50 GB PDF files

BicTel
- Electronic Thesis Dissertations
- ETD Software
- +1250 PDF text
- 1 GB

Iconothèque
- Image collections
- ContentDM
- +12000 JPEG images
- 2 GB

What should we preserve? The ULB case study

Institutional Repository
- Scientific publications
- DSpace
- +78000 references
- +10000 PDF
- 2 GB

Digithèque
- Digitized books and publications
- Symphony
- +220000 digitized pages
- 3TB (raw TIFF) - 1TB (J2K)
- 50 GB PDF files

BicTel
- Electronic Thesis Dissertations
- ETD Software
- +1250 PDF text
- 1 GB

Iconothèque
- Image collections
- ContentDM
- +12000 JPEG images
- 2 GB

Required total storage capacity: **5TB**

Every year, the AU Pubisher harvests the IR for AIPs marked as “not preserved” for each collection (defined by doc type)

Archival Unit Publisher
(wwwarch)

dspace-oai
verb=ListRecords
doctype=thesis
preservation=false

<ListRecords>

</ListRecords>

Institutional Repository
(dipot)

For each collection of AIPs, the AU Publisher creates a new AU and adds the corresponding URL to the Manifest Page

Archival Unit Publisher
(wwwarch)

LOCKSS Manifest Page

- RepositoryA
- RepositoryB
- RepostioryC

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

In the following months, the AU Publisher is slowly harvested by offline and online preservation repositories which collects all the AUs

Archival Unit
Publisher
(wwwarch)

LOCKSS Manifest Page

- RepositoryA
- RepositoryB
- RepostioryC

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

In the following months, the AU Publisher is slowly harvested by offline and online preservation repositories which collects all the AUs

Archival Unit Publisher
(wwwarch)

LOCKSS Manifest Page

- RepositoryA
- RepositoryB
- RepostioryC

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

LOCKSS boxes
(stoarch)

x 7

Offline boxes
(stoarch)

x 7

Preservation repositories

PLN in practice: structure of an AU

2012-ETD-ULB-Archival Unit

- **manifest.html**
- **metadata.xml**
- 

- 

- ...
- 

PLN in practice: structure of an AU

2012-ETD-ULB-Archival Unit

- **manifest.html**
- **metadata.xml**
- 

- 

- ...
- 

 can be exported as a (W)ARC file

PLN in practice: structure of an AU

2012-ETD-ULB-Archival Unit

ARCHIVE

- manifest.html
- metadata.xml
- AIP 1
- AIP 2
- AIP N

AIP-[UUID] - BagIt

- bagit.txt
- bag-info.txt
- tagmanifest-md5.txt
- manifest-sha512.txt
- data/
  - METS.[UUID].xml
  - logs/
  - metadata/
  - objects/

Schedule

- Introduction to Digital Preservation
- Media Type Preservation Planning
- Current State of Preservation Tools
- E-Theses Preservation
- PDF to PDF/A Migration Workflow

  <BREAK>
- Archivematica
- The Trappist Method for the Dissemination and Preservation of Digital Objects
- **Q and A**