

SCAPE Document Conversion Service

Sustaining the Value of Digital through
Format Transformation Cloud Services

Alex D. Wade
Director – Scholarly Communication
Microsoft Research

Microsoft[®]

Digital is Empowering

But...

but its properties must be understood in order to use it effectively.
It is dependent on a sophisticated infrastructure and ability to compute.

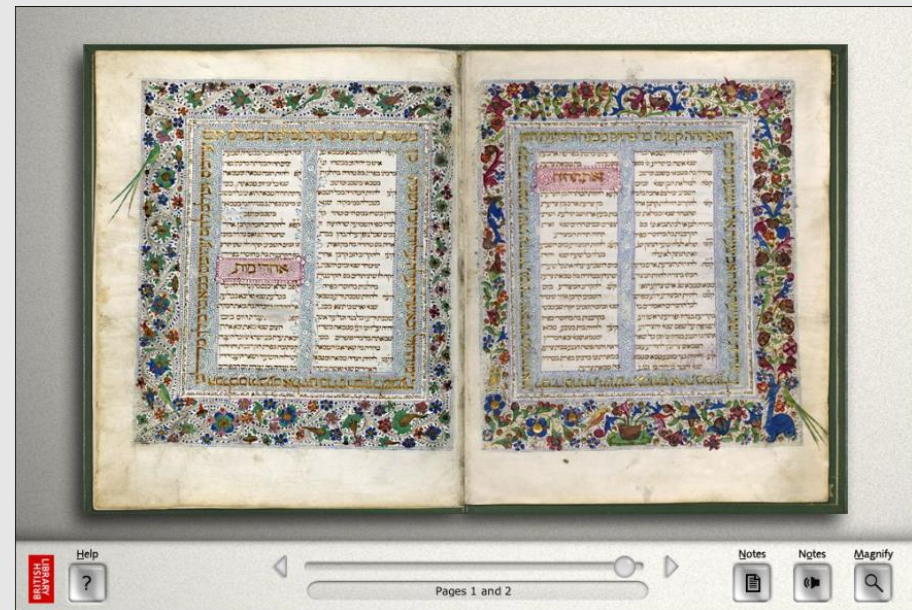
How to ensure that the today's
digital content can be used in the future?

Document formats, software and hardware are becoming obsolete faster than we can ensure the forward compatibility of the content.

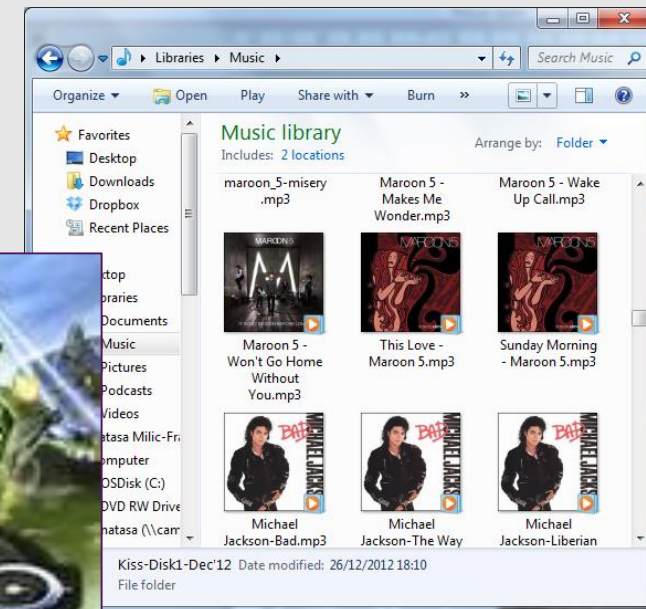
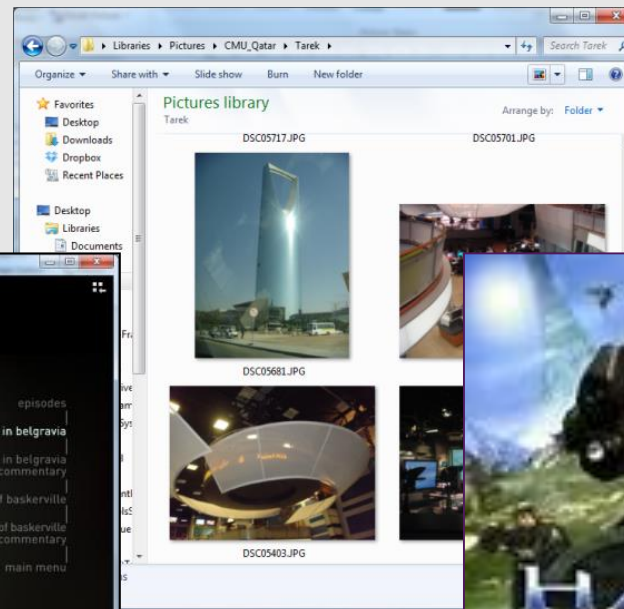
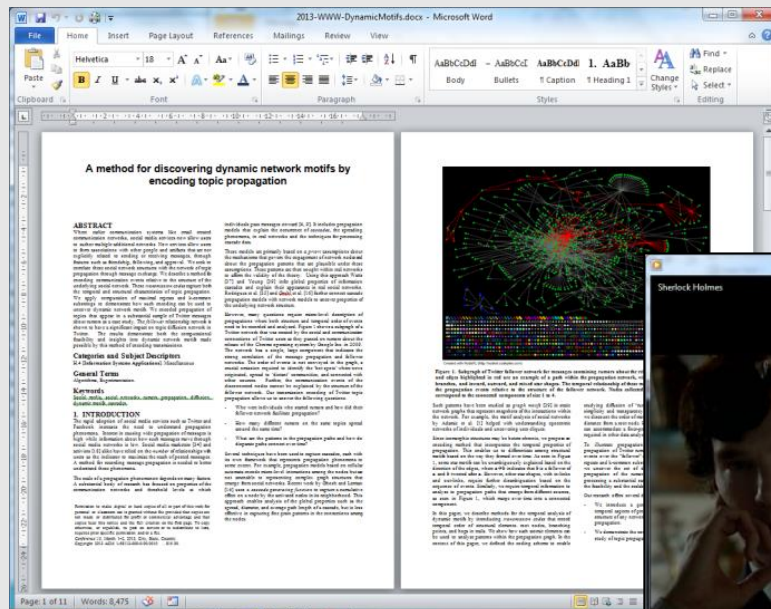


Documents

Digitized collections



Born digital



- Ensure long-term access to Europe's cultural and scientific heritage
 - Improve decision-making about long term preservation
 - Ensure long-term access to valued digital content
 - Control the costs through automation, scalable infrastructure
 - Ensure wide adoption across the user community
 - Establish market place for preservation services and tools
- Build practical solutions
 - Integrate existing expertise, designs and tools
 - Share and build



Preservation and
Long-term
Access through
NETworked
Services

PLANETS Partners 2006-2010



The British Library
National Library, Netherlands
Austrian National Library
State and University Library, Denmark
Royal Library, Denmark



National Archives, UK
Swiss Federal Archives
National Archives, Netherlands

nationaal archief



Hatii at University of Glasgow
University of Freiburg
Technical University of Vienna
University at Cologne



Tessella Plc
IBM Netherlands
Microsoft Research, Cambridge
ARC Seibersdorf research



Format migration of Office documents

Source formats

- WordPerfect 5
- WordPerfect 6
- DOS Word
- Word 2, 6, 95
- Word 97-2003
- RTF
- ODF
- OpenXML

Target formats

- OpenXML
- ODF
- UOF
- HTML
- XCDL (format defined in PLANETS)

SCAPE—SCAlable Preservation Environments



- Develop scalable services for planning and execution of preservation strategies
- Open source platform for semi-automated workflows for large-scale, heterogeneous collections of complex digital objects.

FP7 Project. Started February 2010. Sponsored for 3.5 years.

SCAPE Partners 2010-2014



AIT Austrian Institute of Technology GmbH



The British Library

Internet Memory Foundation



Ex Libris Ltd.



**Fachinformationszentrum
Karlsruhe, Gesellschaft für
Wissenschaftlich-
Technische Information
GmbH**



Koninklijke Bibliotheek

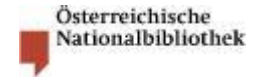


KEEP SOLUTIONS LDA



Microsoft Research

**Österreichische
Nationalbibliothek**



Open Planets Foundation



Statsbiblioteket



**Science and Technologies
Facilities Council**



**Technische Universität
Berlin**



**Technische Universität
Wien**

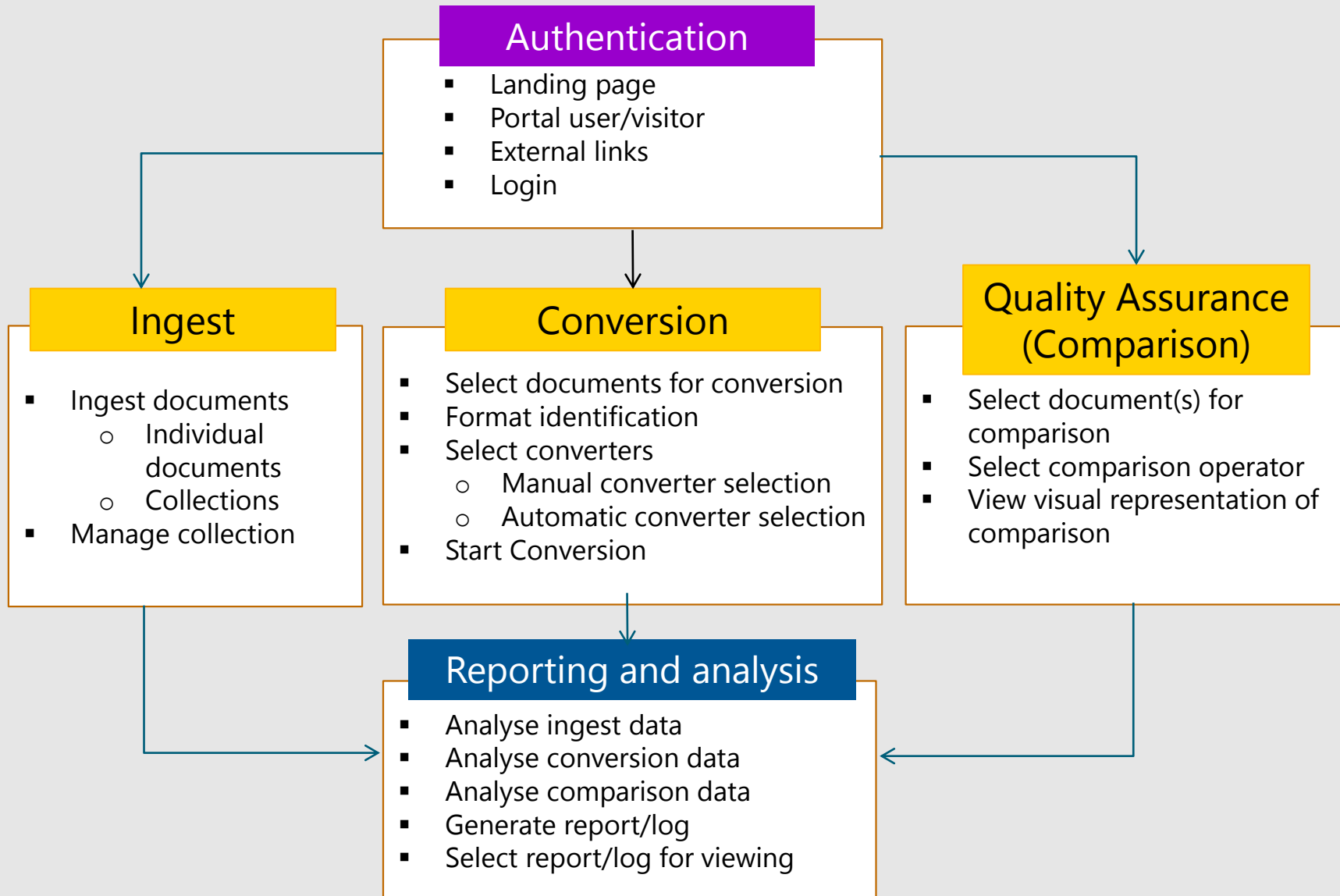
**The University of
Manchester**



**Universite Pierre et Marie
Curie – Paris 6**



Workflow supported by the SCAPE Azure Services



SCAPE Azure Service demo



Available conversions

- Document conversions supported by MSR Azure Services
- General principle: repurpose existing tools.

	.docx 2007	.docx 82X	.doc 97-2003	.docm macro	.dotx template	.dotm macro template	.dot template	.odt	.rtf	.mht	.mhtml	.xml	.png RGBA	.pdf v 9.0	.dz DZOOM	.xps
.docx 2007	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.doc 97-2003	•			•	•	•	•	•	•	•	•	•	•	•	•	•
.docm macro	•		•		•	•	•	•	•	•	•	•	•	•	•	•
.dotx template	•		•	•		•	•	•	•	•	•	•	•	•	•	•
.dotm macro template	•		•	•	•		•	•	•	•	•	•	•	•	•	•
.dot template	•		•	•	•	•		•	•	•	•	•	•	•	•	•
.odt	•															
.rtf	•		•	•	•	•	•			•	•	•	•	•	•	•
.mht	•		•	•	•	•	•	•	•		•	•	•	•	•	•
.mhtml	•		•	•	•	•	•	•	•	•		•	•	•	•	•
.xml	•		•	•	•	•	•	•	•	•	•		•	•	•	•
.png RGBA															•	
.pdf v 9.0													•		•	

Comparison



Format transformation



6

7

8

9

10

11

4 PREMIS

PREMIS is the acronym for Preservation Metadata: Implementation Strategies. The Data Dictionary of PREMIS defines preservation metadata and provides an XML schema³. The PREMIS Data Model⁴ defines Intellectual Entities, Objects, Rights, Agents and Events.

An Object in PREMIS has three subtypes: file, representation and bitstream.

The Intellectual Entity is e.g. a book, a map, photograph or database etc. The Representation of an IE is a set of files including structural Metadata (e.g. described by METS). A File is a named and ordered sequence of bytes known by the operating system that can be written, read and copied. Bitstreams

³ <http://www.loc.gov/standards/premis/premis.xsd>
⁴ <http://www.loc.gov/standards/premis/>

Page 7 of 30

Synchronize pages
Show full screen

Fit page

Show highlights:
All None Error

6

7

8

9

10

11

4 PREMIS

PREMIS is the acronym for Preservation Metadata: Implementation Strategies. The Data Dictionary of PREMIS defines preservation metadata and provides an XML schema³. The PREMIS Data Model⁴ defines Intellectual Entities, Objects, Rights, Agents and Events.

An Object in PREMIS has three subtypes: file, representation and bitstream.

The Intellectual Entity is e.g. a book, a map, photograph or database etc. The Representation of an IE is a set of files including structural Metadata (e.g. described by METS). A File is a named and ordered sequence of bytes known by the operating system that can be written, read and copied. Bitstreams

³ <http://www.loc.gov/standards/premis/premis.xsd>
⁴ <http://www.loc.gov/standards/premis/>

Page 7 of 29

Show Statistics

Fit page

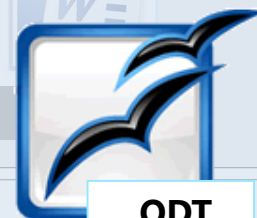
Show highlights:
All None Error Missing

Comparison



Format transformation

Open Office → MS Word



.ODT



.DOCX

6

7

8

9

10

11

SCAPE

4 PREMIS

PREMIS is the acronym for Preservation Metadata: Implementation Strategies. The Data Dictionary of PREMIS defines preservation metadata and provides an XML schema². The PREMIS Data Model³ defines Intellectual Entities, Objects, Rights, Agents and Events.

Intellectual Entities: Content that can be described as a unit (e.g. books, articles, databases).

Rights: Assertion of rights and permissions.

Objects: Discrete units of information in digital form. Can be files, bitstreams or representations.

Agents: People, organizations, or software.

Events: Actions that involve an Object and an Agent known to the system.

An Object in PREMIS has three subtypes: file, representation and bitstream.

Intellectual Entity 1 -- 1 Representation 1 -- 1 File 1 -- 1 Bitstream

Object

The Intellectual Entity is e.g. a book, a map, photograph or database etc. The Representation of an IE is a set of Files including structural Metadata (e.g. described by METS). A File is a named and ordered sequence of bytes known by the operating system that can be written, read and copied. Bitstreams

² <http://www.loc.gov/standards/premis/premis.xsd>

³ <http://www.loc.gov/standards/premis/>

4

Page 7 of 30

Synchronize pages

Show full screen

Show highlights: Match Missing

All None Error

6

7

8

9

10

11

SCAPE

4 PREMIS

PREMIS is the acronym for Preservation Metadata: Implementation Strategies. The Data Dictionary of PREMIS defines preservation metadata and provides an XML schema². The PREMIS Data Model³ defines Intellectual Entities, Objects, Rights, Agents and Events.

Intellectual Entities: Content that can be described as a unit (e.g. books, articles, databases).

Rights: Assertion of rights and permissions.

Objects: Discrete units of information in digital form. Can be files, bitstreams or representations.

Agents: People, organizations, or software.

Events: Actions that involve an Object and an Agent known to the system.

An Object in PREMIS has three subtypes: file, representation and bitstream.

Intellectual Entity 1 -- 1 Representation 1 -- 1 File 1 -- 1 Bitstream

Object

The Intellectual Entity is e.g. a book, a map, photograph or database etc. The Representation of an IE is a set of Files including structural Metadata (e.g. described by METS). A File is a named and ordered sequence of bytes known by the operating system that can be written, read and copied. Bitstreams

² <http://www.loc.gov/standards/premis/premis.xsd>

³ <http://www.loc.gov/standards/premis/>

4

Page 7 of 30

Synchronize pages

Show full screen

Show highlights: Match Missing

All None Error

Screen Print – XPS

OCR Processing

Feature extraction / comparison

What does Cloud offer us?

Extendible functionality

Extendible data store

Scalable computation

Virtualization

Common platform for creating services

Support for client applications on diverse computing platforms

Cloud paradigm may secure the digital future

under the assumptions that:

- Access to digital media becomes one of the primary drivers for innovation and evolution of the ICT ecosystem
 - Customers/digital media producers should demand and pay for long term access provisions at the time of technology acquisition.
- Digital media curation and education become an essential component of digital media services
 - Content creators and content holders need to demonstrate that there is value in combining contemporary and past information to provide compelling and competitive services.

Discussion / Questions?



Microsoft[®]