SCAPE Document Conversion Service
Sustaining the Value of Digital through Format Transformation Cloud Services

Alex D. Wade
Director – Scholarly Communication
Microsoft Research
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
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
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—SCAlatable 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.
<table>
<thead>
<tr>
<th>AIT Austrian Institute of Technology GmbH</th>
<th>Österreichische Nationalbibliothek</th>
</tr>
</thead>
<tbody>
<tr>
<td>The British Library</td>
<td>Open Planets Foundation</td>
</tr>
<tr>
<td>Internet Memory Foundation</td>
<td>Statsbiblioteket</td>
</tr>
<tr>
<td>Ex Libris Ltd.</td>
<td>Science and Technologies Facilities Council</td>
</tr>
<tr>
<td>Fachinformationszentrum Karlsruhe, Gesellschaft für Wissenschaftlich-Technische Information GmbH</td>
<td>Technische Universität Berlin</td>
</tr>
<tr>
<td>Koninklijke Bibliotheek</td>
<td>Technische Universität Wien</td>
</tr>
<tr>
<td>KEEP SOLUTIONS LDA</td>
<td>The University of Manchester</td>
</tr>
<tr>
<td>Microsoft Research</td>
<td>Université Pierre et Marie Curie – Paris 6</td>
</tr>
</tbody>
</table>
Workflow supported by the SCAPE Azure Services

- **Authentication**
  - Landing page
  - Portal user/visitor
  - External links
  - Login

- **Ingest**
  - Ingest documents
    - Individual documents
    - Collections
  - Manage collection

- **Conversion**
  - Select documents for conversion
  - Format identification
  - Select converters
    - Manual converter selection
    - Automatic converter selection
  - Start Conversion

- **Quality Assurance (Comparison)**
  - Select document(s) for comparison
  - Select comparison operator
  - View visual representation of comparison

- **Reporting and analysis**
  - Analyse ingest data
  - Analyse conversion data
  - Analyse comparison data
  - Generate report/log
  - Select report/log for viewing
SCAPE Azure Service demo
Available conversions

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