

# Sustaining and Enhancing the Value of Digital Assets

DPC Briefing Day: Virtualization and Preservation

*How cloud computing changes how we think about digital preservation'*

Cambridge, 22 July, 2014

Natasa Milic-Frayling

Principal Researcher

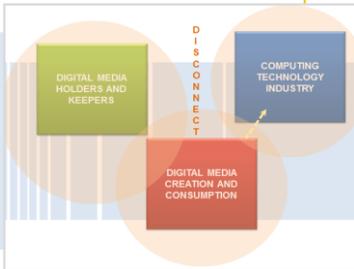
Microsoft Research Cambridge, UK

# Overview



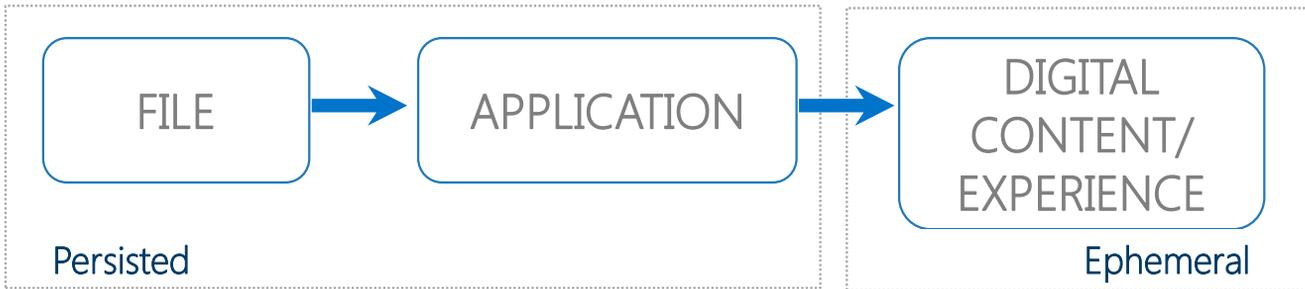
## The Nature of Digital

*Properties and paradoxes of digital*

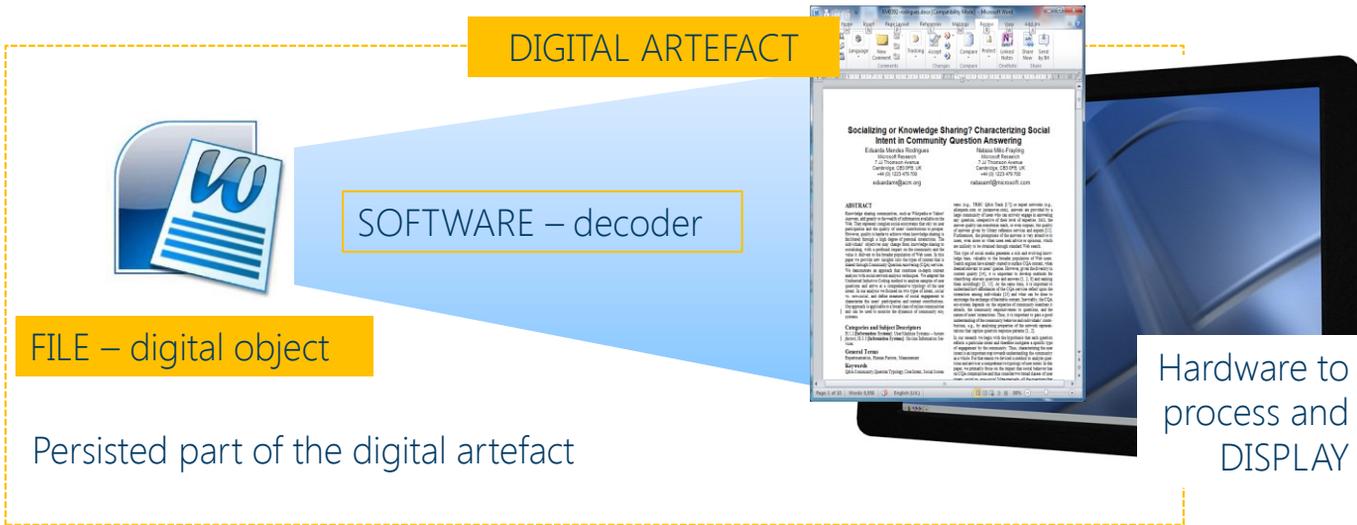


## Sustainability of Digital

*Methods for ensuring content accessibility  
Gaps in the value chain*



PRESERVATION = Persistence + Connection with the contemporary ecosystem.



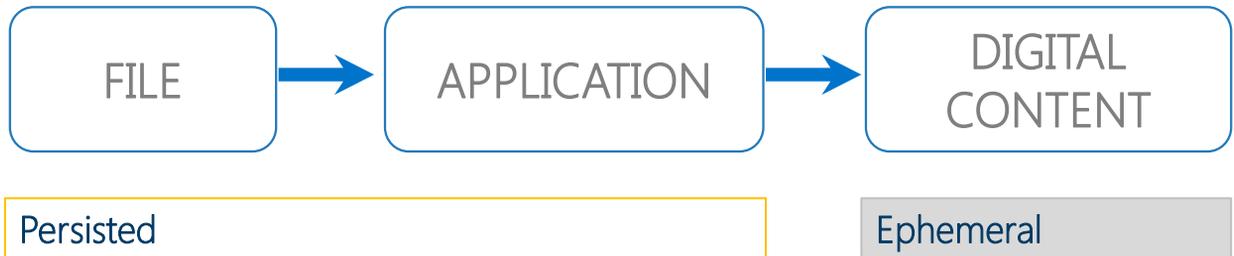
Paradox: we are concerned about storage, yet

*Digital is inherently about processing bits,  
not about storing bits*

# Symbiosis of Files and Applications

Objective of preservation is to ensure that the persisted digital content and applications remain connected with the contemporary computing ecosystem.

PRESERVATION = Persistence + Connection with the contemporary ecosystem.

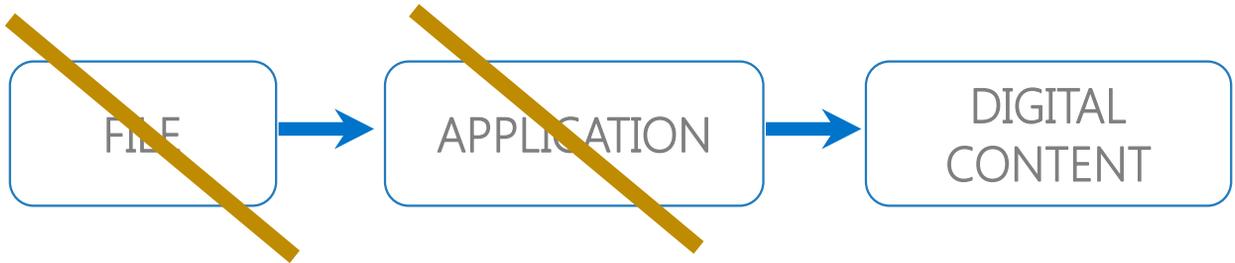


# What do you want to keep 'unchanged'?



- If application is not running in the contemporary environment

# What do you want to keep 'unchanged'?



- If application is not running in the contemporary environment
  - Migrate files and run with a contemporary software

(give up on both the original files and the application)

# What do you want to keep 'unchanged'?



- If application is not running in the contemporary environment
  - Retain the files and port the application to the new environment  
(retain content files by give up on the application, at least partially)

# What do you want to keep 'unchanged'?



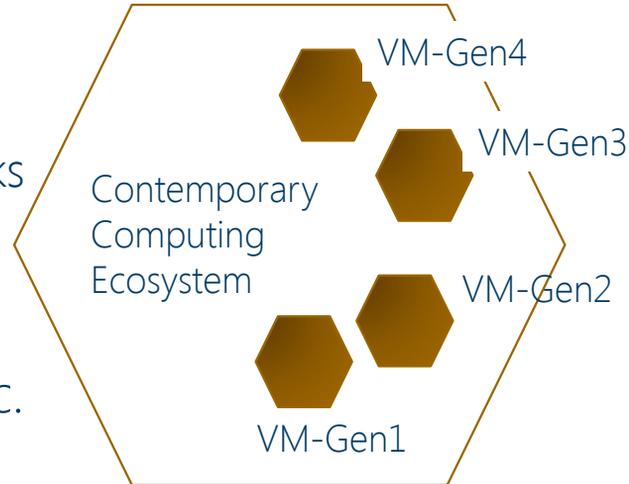
- If application is not running in the contemporary environment
  - Create a virtual machine with the old computing stack and run the original files and software.  
(retain original files and original application; maintain scaffolding)

# Computational Cradles

Sustain and increase the value of digital through

- Virtualization of legacy software + Bridging Services
- Individual computational 'cells' for different generations of software stacks

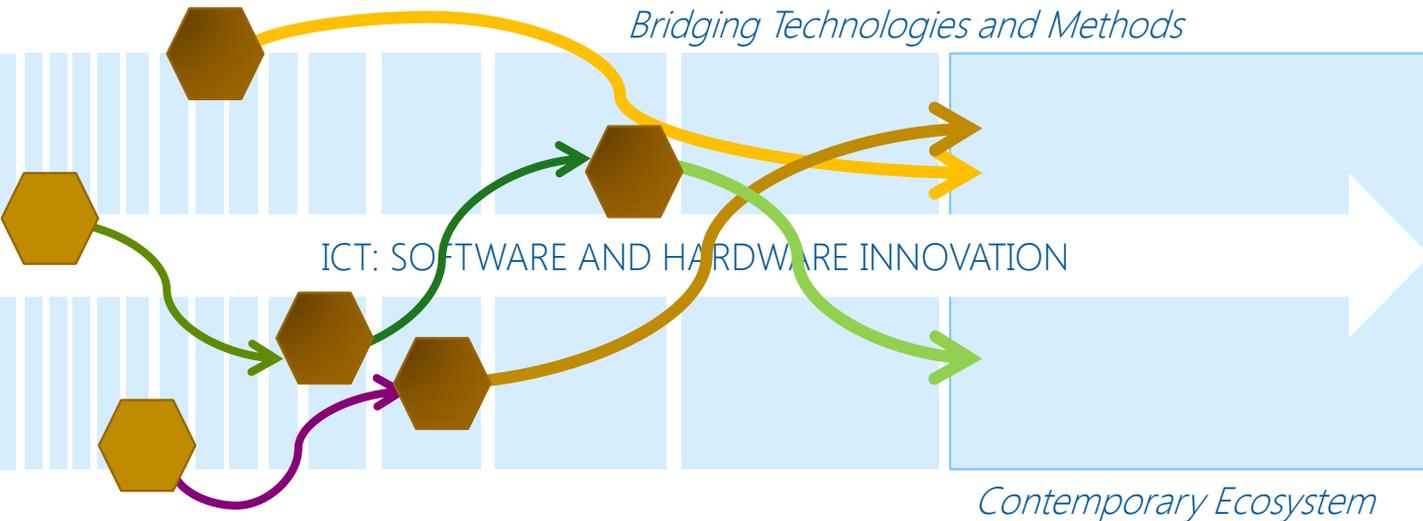
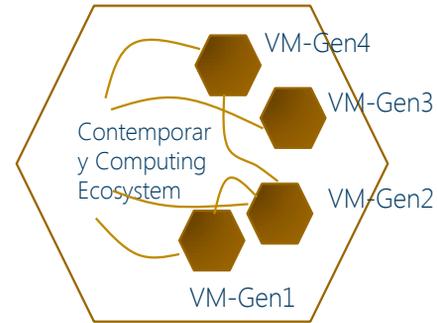
Bridging services: format translators, content extractors, etc.



# Connecting Legacy with the Contemporary Ecosystem

Digital artifact always requires (some software) computation.

No need to give up on the original software!



# Winning strategy

## Virtualization of original software

Ensures access to the digital artefacts

## Format transformation services

On demand transformation within a specific context.

# Winning strategy

## Virtualization of original software

Ensures access to the digital artefacts

## Format transformation services

On demand transformation within a specific context.

extracting value from digital

FORMAT TRANSFORMATION SERVICES

# 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



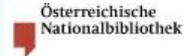
KEEP SOLUTIONS LDA



Microsoft Research Limited



Österreichische Nationalbibliothek



Open Planets Foundation



Statsbiblioteket



Science and Technologies Facilities Council



Technische Universität Berlin



Technische Universität Wien

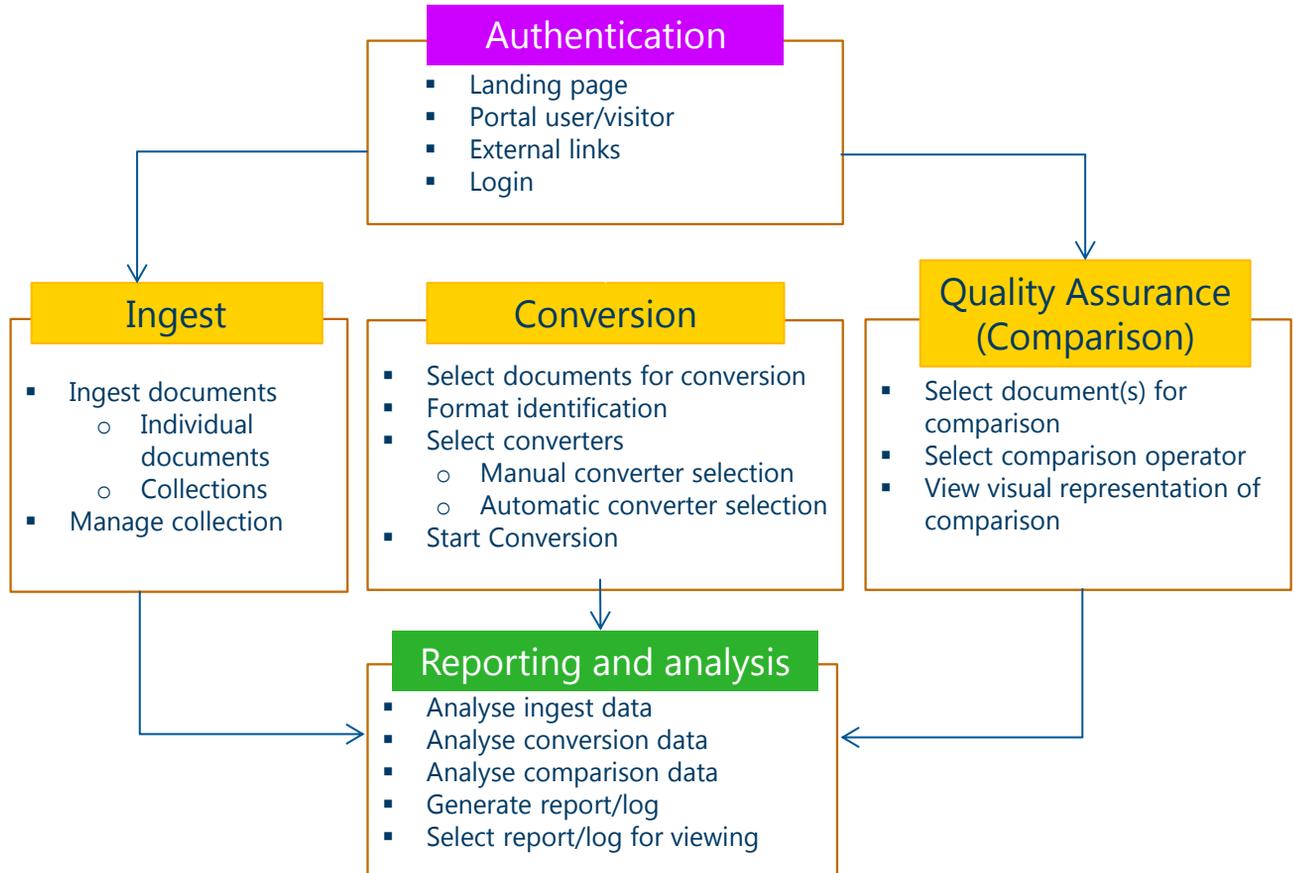


The University of Manchester



Universite Pierre et Marie Curie – Paris 6

# User Interaction with SAZ Services



# Demo: Use of SCAPE Azure Services



# SCAPE: Format Transformation Services on Microsoft Cloud (Azure)

## Format Transformation Options

The screenshot displays the SCAPE web application interface. At the top, there is a navigation bar with the SCAPE logo (SCAble Presentation Environment) and the Microsoft Research logo. Below the navigation bar, there is a sign-in section with fields for "User name:" and "Password:", a "Keep me signed in" checkbox, and a "Sign In" button. The main content area features four large icons representing different services: "manage collections in the cloud", "convert files", "compare conversions", and "view reports". Below these icons is a grid of file format icons and their corresponding supported output formats. The grid is organized as follows:

Input Format	.docx 2007	.docx 97-2003	.doc 97-2003	.docm macro	.dotx template	.dotm macro	.dot template	.odt	.rtf	.mht	.mhtml	.xml	.png Raster	.pdf 1.0	.dz DZOOM	.xps
.docx 2007	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.doc 97-2003	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.docm macro	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.dotx template	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.dotm macro	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.dot template	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.odt	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.rtf	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.mht	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.mhtml	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.xml	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.png Raster	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
.pdf 1.0	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

The bottom of the screenshot shows the Windows taskbar with various application icons and the system tray displaying the time as 15:19 on 28/10/2013.

# Prototype Features

## Results of the Conversion Process

The screenshot displays a web application interface for SCAPE (SCAble Presentation Environment). The browser address bar shows the URL: <http://scapestaging.cloudapp.net:8080/Collections/ConvertedFile>. The page header includes the SCAPE logo and the text "SCAPE: Collections", "cloudapp.net", and "Microsoft Research". A navigation bar at the top right contains links for "Welcome, scape3", "Manage Account", and "Sign Out".

The main content area is titled "Converted Files for the 'Collection DIV (50)' Collection" and includes a link to "[Back to Collections](#)". Below this, a table displays the conversion results for several files. The table has columns for the original file type and various converted formats. Each cell in the table contains a small icon representing the file format, with a green checkmark indicating a successful conversion.

Uploaded Files	DOCX (B2X)	DOCX	DOC	XML	ODT	PDF	XPS	PNG	Deep Zoom
AAI-paper-4.doc <a href="#">View</a>									
SCAPE_Digital_Object_Model_... <a href="#">View</a>									
SCAPE-Connector-API-08-02-2012_v1.docx <a href="#">View</a>									
Raport-i-Auditorit-QHPD-2012-final.docx <a href="#">View</a>									
005661.doc <a href="#">View</a>									

The Windows taskbar at the bottom shows the system clock at 15:15 on 28/10/2013, along with various application icons.

# Prototype Features

## Word'97 Document Transformed into Deep Zoom Format



### SCAPE Digital Object Model

#### Authors

Persons	Role	Partner	Contributions
Matthew Rubin	PI		
Frank Arang	PI		
SP Director		INRA	
Raf Castro		HEPS	

#### Distributors

Partner Org	Role	Partner
SCAPE All		All Partners

#### Revision History

Version	Release	Author	Changes
0.0		Matthew Rubin	2002-09-14
0.1		Matthew Rubin	2002-09-26
0.2		Neil Snowmer	2002-09-27
0.4		Frank Arang	2002-09-29
0.5		Raf Castro	2002-09-29



#### Table of Contents

1	Introduction	1
2	OAS	1
3	METS	2
3.1	A METS Document	2
4	PREMIS	4
5	METS & PREMIS	5
6	SWOSD	6
7	Digital Object Model of SCAPE Repositories	7
7.2	Rosetta	7
7.2.1	Structure	8
7.3	RODA	11
7.3.1	SP	11
7.3.2	AP	11
7.4	ojsDoc	14
8	The SCAPE Digital Object Model	16
8.1	Requirements METS	17
8.2	Example METS Profiles	19
8.3	METS and PREMIS Identifiers	21
8.4	Revision Schemes	21
8.5	Requirements for the OAS Information Packages	22
8.5.1	Definition of a SP	22
8.5.2	Definition of a AP	22
8.5.3	Definition of a DP	22
8.6	Preservation Plans	23
8.6	Summary	23
9	Conclusion	24
10	Glossary	25



#### 11 List of Figures

26

#### 1 Introduction

To be able to implement repository services like the Connector API and the Loader Application we need to agree on a Digital Object Model within the SCAPE project. To ensure that such repository already provides a Digital Object Model but the diversity hinders the SCAPE platform to integrate all the partner repositories. The task of a Digital Object Model has been put on the SCAPE use register on "thompson". This Document tries to resolve this issue.

In order to use the same terminology throughout the document we give a short introduction into the well-known standards like OAS, METS and PREMIS used in the long term preservation world. Some questions related to our domain specific requirements are discussed.

On the next abstract level the reference model for archives, the OAS, model will be discussed in brief. The METS standard describes an XML container for metadata structure of digital objects. METS is widely used for interconnecting different repositories and service components. The PREMIS standard describes a semantic model for preservation metadata, and is widely used in long term preservation. OAS, METS and PREMIS together will be discussed briefly since there are some loose ends to be aware of.

Some of the repositories of the SCAPE members already use METS and PREMIS, but the mere employment of these standards does not guarantee interoperability in between digital repositories. There might for example be significant differences in between two METS documents as they are used for two different repositories. We will describe the current existing data models of the repositories in this document and develop a possible model every repository holder may subscribe to.

#### 2 OAS

OAS is the acronym for an Open Archival System and describes on an abstract level the requirements an archival system for long term preservation has to fulfil. The following functional areas are described by the reference model:

1. Ingest
2. Archival System
3. Data Management
4. Administration
5. Preservation Planning
6. Access

The key terms for this document are SP (Submission Information Package), AP (Archival Information Package) and DP (Dissemination Information Package).

Table 1: Content of an Archival Information Package (AIP) in the SCAPE Digital Object Model

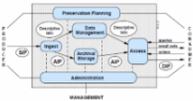


Figure 1: Metadata in the SCAPE Digital Object Model

The current SP definitions of existing SCAPE repositories have significant differences in RODA for example a SP is a compressed ZIP file that contains a METS envelope, in Rosetta a SP may contain several IP (Intellectual Entities). An AP contains technical metadata and metadata important for long term archiving.



	metadata	
Technical metadata	file	Any form, e.g. XML, HTML, PREMIS Object Metadata ... Any form, e.g. PREMIS
Source metadata	file	Info (description, rights, technical) about the original source used to generate the digital object
Rights metadata	file	Any form, e.g. Intellect, copyright(METS), PREMIS Rights Metadata
Digital preservation metadata	file	Any form, e.g. PREMIS Event Metadata
File location	All files that comprise the content of the digital entity. Files are ordered in groups (SP, lang etc.)	



#### 4 PREMIS

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



represent data within a file (e.g. a jpeg in a PDF document), audio data within a MOX file or graphics within a word document.

#### 5 METS & PREMIS

METS is an XML container for structuring metadata in different formats to often used in conjunction with the PREMIS standard for preservation metadata. But one has to consider the existing usage of the METS and PREMIS definitions. There are a few documents available describing best practices and guidelines of how to use PREMIS within METS, see for example [1].



# SCAPE: Format Transformation Services on Microsoft Cloud (Azure)

## Quality Assessment of the Conversion Process

The screenshot displays the SCAPE web application interface. At the top, the browser address bar shows the URL <http://scapestaging.cloudapp.net:8080/Default.aspx?ReturnUri=%2F>. The page header includes the SCAPE logo (Scalable Preservation Environment) and the Microsoft Research logo. Navigation options include "Create Account" and "Sign In".

The main interface features a central navigation bar with four primary actions: "manage collections in the cloud", "convert files", "compare conversions", and "view reports". Below this, the application is divided into two main panels, each displaying a document page (Page 10 of 30 and Page 10 of 29).

The left panel shows a document page with a diagram illustrating the conversion process. The diagram consists of three boxes: "SIP Submission Information Package" (grey), "AIP Archival Information Package" (red), and "DIP Dissemination Information Package" (yellow). Arrows indicate the flow from SIP to AIP, and then from AIP to DIP. The text on the page discusses the Digital Object Model of SCAPE Preservations and the role of the SCAPE system in providing a scalable preservation environment.

The right panel shows a document page with a diagram illustrating the comparison process. The diagram consists of two boxes: "SIP Submission Information Package" (grey) and "AIP Archival Information Package" (red). Arrows indicate the flow from SIP to AIP. The text on the page discusses the Digital Object Model of SCAPE Preservations and the role of the SCAPE system in providing a scalable preservation environment.

At the bottom of the interface, there are controls for "Synchronize pages", "Show full screen", "Page 10 of 30", "Fit page", and "Show Statistics". The Windows taskbar at the bottom shows the system clock at 15:20 on 28/10/2013.

# Prototype Features

## Differences are Assessed and Classified based on 'Importance'



### 7 Digital Object Model of SCAPE Repositories

The SCAPE partners are using different repository implementation such as Rosetta (ExLibris), RODA (Keele), eDocDoc (FZ Karlsruhe) and DOWS (Sb). All of these repositories do have their own Digital Object Model. We will briefly describe these models in the following section.

#### 7.1 Rosetta

Ex Libris Rosetta is a digital-object preservation solution that conforms to the ISO-recognized Open Archival Information System (OAIS). Rosetta allows the SIP and the DIP to have a variety of formats and structures and provides an SDK to support this.

The following chapter describes the AIP, which is stored in a METS XML file in Rosetta's Permanent Repository module. Each AIP describes one IE (Intellectual Entity).

The METS XML is generated in the Staging module during the SIP processing. During processing, the IE information is kept and managed in the database. By the time the SIP is moved to the permanent repository, the METS XML contains all the information regarding the IE, collected from the different database tables.

The information on the METS XML can be reloaded back into the database when the IE is brought from the permanent repository for maintenance (preservation actions, adding representations, and so forth).

The following diagram shows the flow between the three types of information packages:



```
graph LR; SIP[Submission Information Package] --> AIP[Archival Information Package]; AIP --> DIP[Dissemination Information Package];
```



### 7 Digital Object Model of SCAPE Repositories

The SCAPE partners are using different repository implementation such as Rosetta (ExLibris), RODA (Keele), eDocDoc (FZ Karlsruhe) and DOWS (Sb). All of these repositories do have their own Digital Object Model. We will briefly describe these models in the following section.

#### 7.1 Rosetta

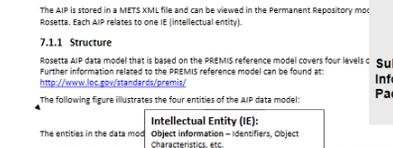
Ex Libris Rosetta is a digital-object preservation solution that conforms to the ISO-recognized Open Archival Information System (OAIS). Rosetta allows the SIP and the DIP to have a variety of formats and structures and provides an SDK to support this.

The following chapter describes the AIP, which is stored in a METS XML file in Rosetta's Permanent Repository module. Each AIP describes one IE (Intellectual Entity).

The METS XML is generated in the Staging module during the SIP processing. During processing, the IE information is kept and managed in the database. By the time the SIP is moved to the permanent repository, the METS XML contains all the information regarding the IE, collected from the different database tables.

The information on the METS XML can be reloaded back into the database when the IE is brought from the permanent repository for maintenance (preservation actions, adding representations, and so forth).

The following diagram shows the flow between the three types of information packages:



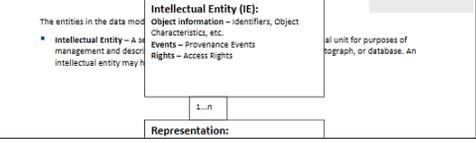
```
graph LR; SIP[Submission Information Package] --> AIP[Archival Information Package]; AIP --> DIP[Dissemination Information Package];
```

The AIP is stored in a METS XML file and can be viewed in the Permanent Repository module. Each AIP relates to one IE (Intellectual Entity).

#### 7.1.1 Structure

Rosetta AIP data model that is based on the PREMIS reference model covers four levels of detail. Further information related to the PREMIS reference model can be found at: <http://www.loc.gov/standards/premis/>

The following figure illustrates the four entities of the AIP data model:



The entities in the data model are:

- Intellectual Entity – A unit of information that has characteristics, etc.
- Submission Information Package – A unit of information that is used for the management and description of an intellectual entity.
- Representation – A unit of information that is used for the management and description of an intellectual entity.

- Synchronize pages
- Show full screen

Page 10 of 30 ▲ ▼ Fit page ▼

Show highlights:  Match  Missing  Error

Page 10 of 29 ▲ ▼ Fit page ▼

Show highlights:  Match  Missing  Error

Show Statistics

# Prototype Features

## Pivot View of the Collection Metadata

Collection | Date created: 2000s > Sort: Author ▾    +

Clear All 

Search... 

Size

Author

Date created 

20th century 28

21st century 775

1900s 1

1990s 27

2000s 774

2010s 1

Custom Range

From Sep 7, 1905 

To Jan 28, 2015 

Page count

Paragraph count

Line count

Character count

Company

Subject

Page Difference (B2X DOCX)

Paragraph Difference (B2X D)

Line Difference (B2X DOCX -)

Word Difference (B2X DOCX)

Good Match Quality (%)

Bad Match Quality (%)

Partial Match Quality (%)

No Match (%)

Average Mismatched Rectan

						
- to	121549 to Cynthia Singh	Dan Ogle, Plant Materials Speciali to Irene Fields	J Clark Salyer to NWTRB	occ to tsong	U.S. Department of Education to Yvette Torres	(no info)

Show full screen

# Winning strategy

## Virtualization of original software

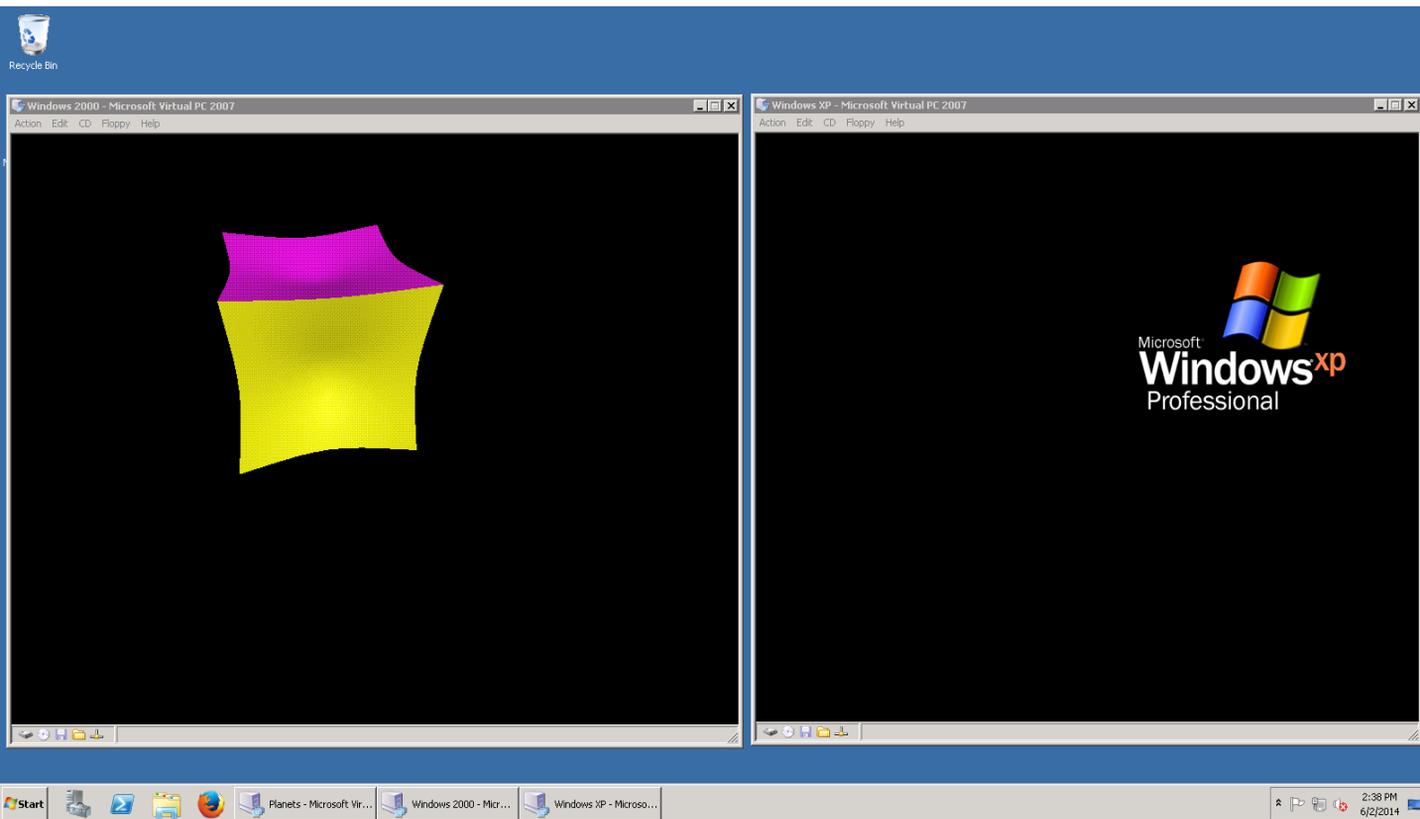
Ensures access to the digital artefacts

## Format transformation services

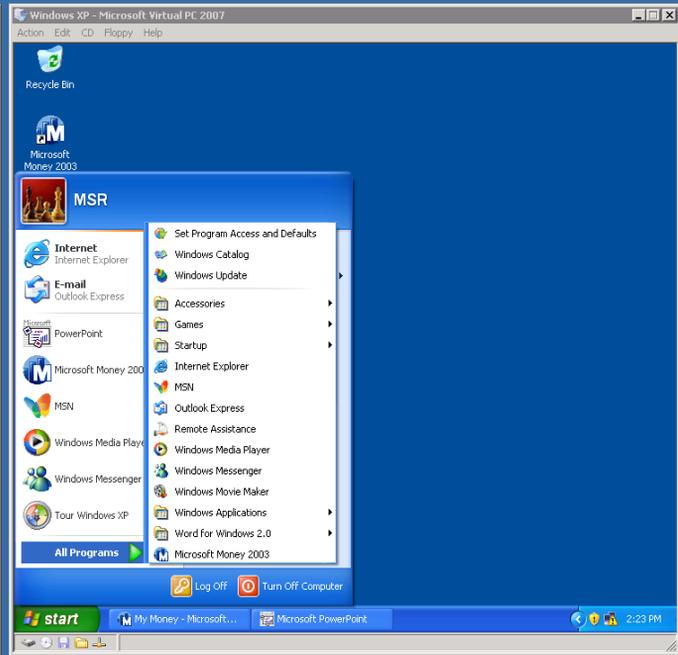
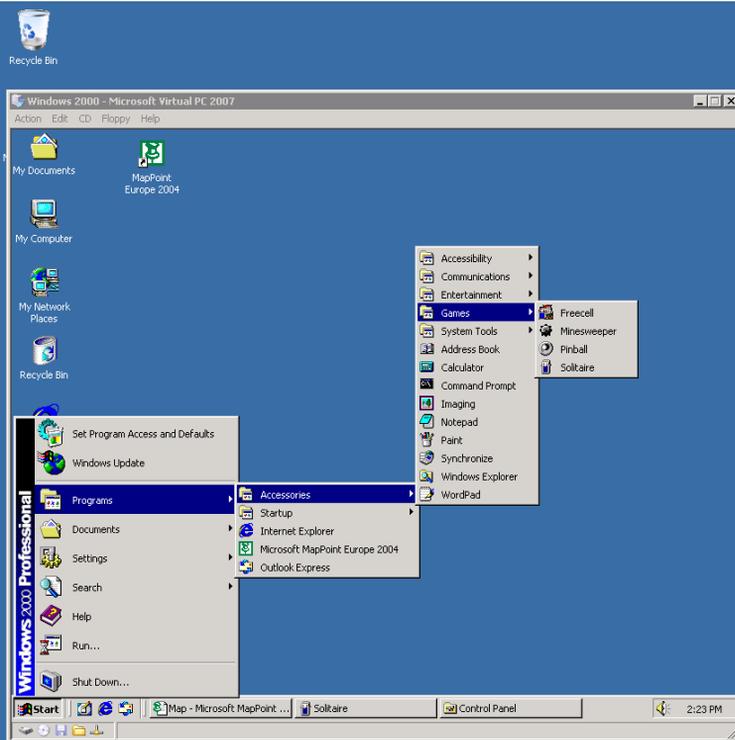
On demand transformation within a specific context.

preserving computation

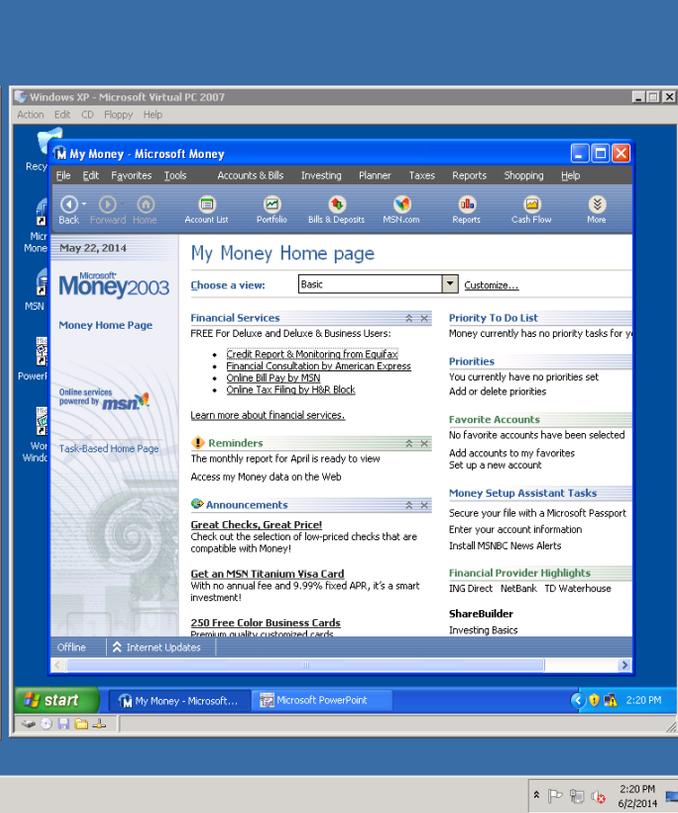
# VIRTUALIZATION OF LEGACY SOFTWARE



Virtual Machine with Windows 2000 (left) and Windows XP (right), running on Microsoft Cloud (Azure)



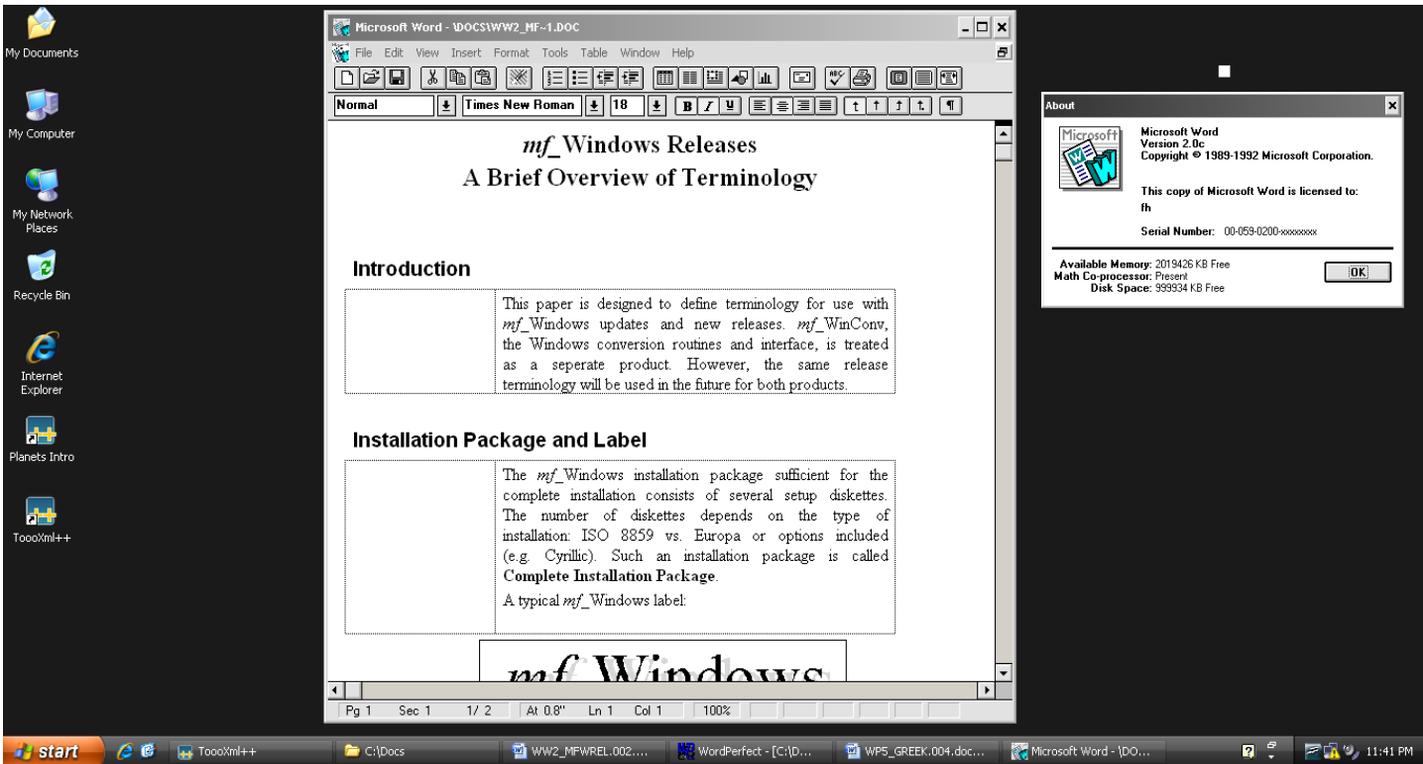
Start menus for Windows 2000 (left) and Windows XP (right),



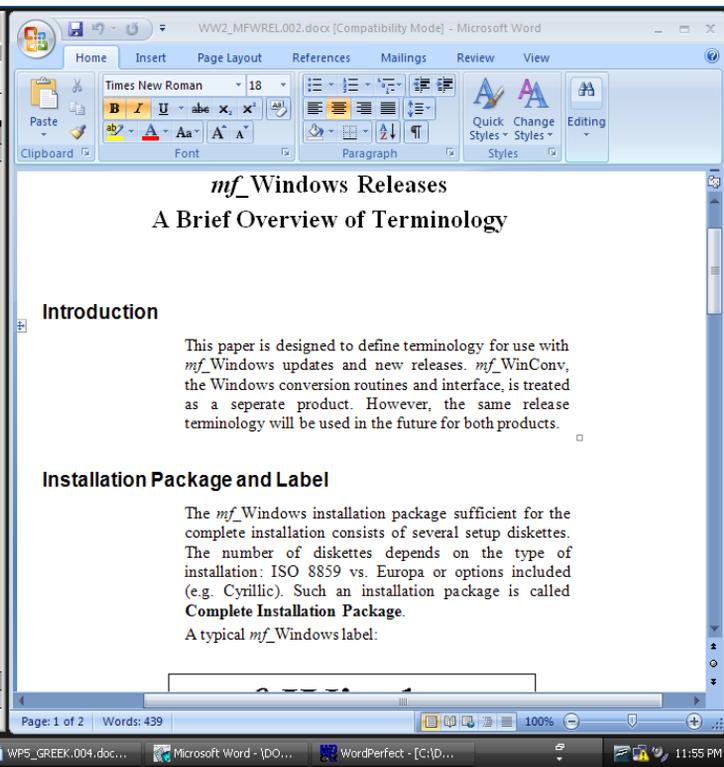
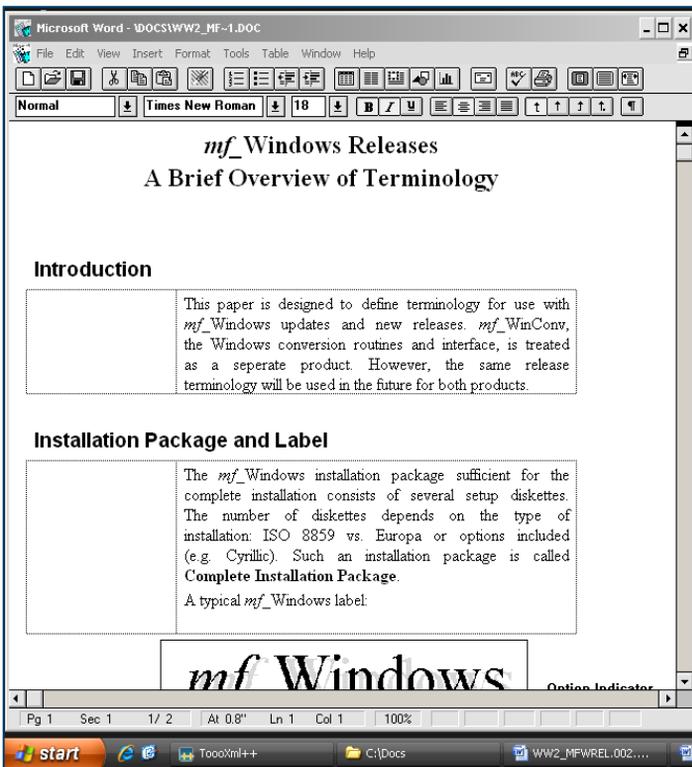
MS Map Point application running on Windows 2000 (left) and MS Money 2003 running on Windows XP (right),

Increasing value of legacy content

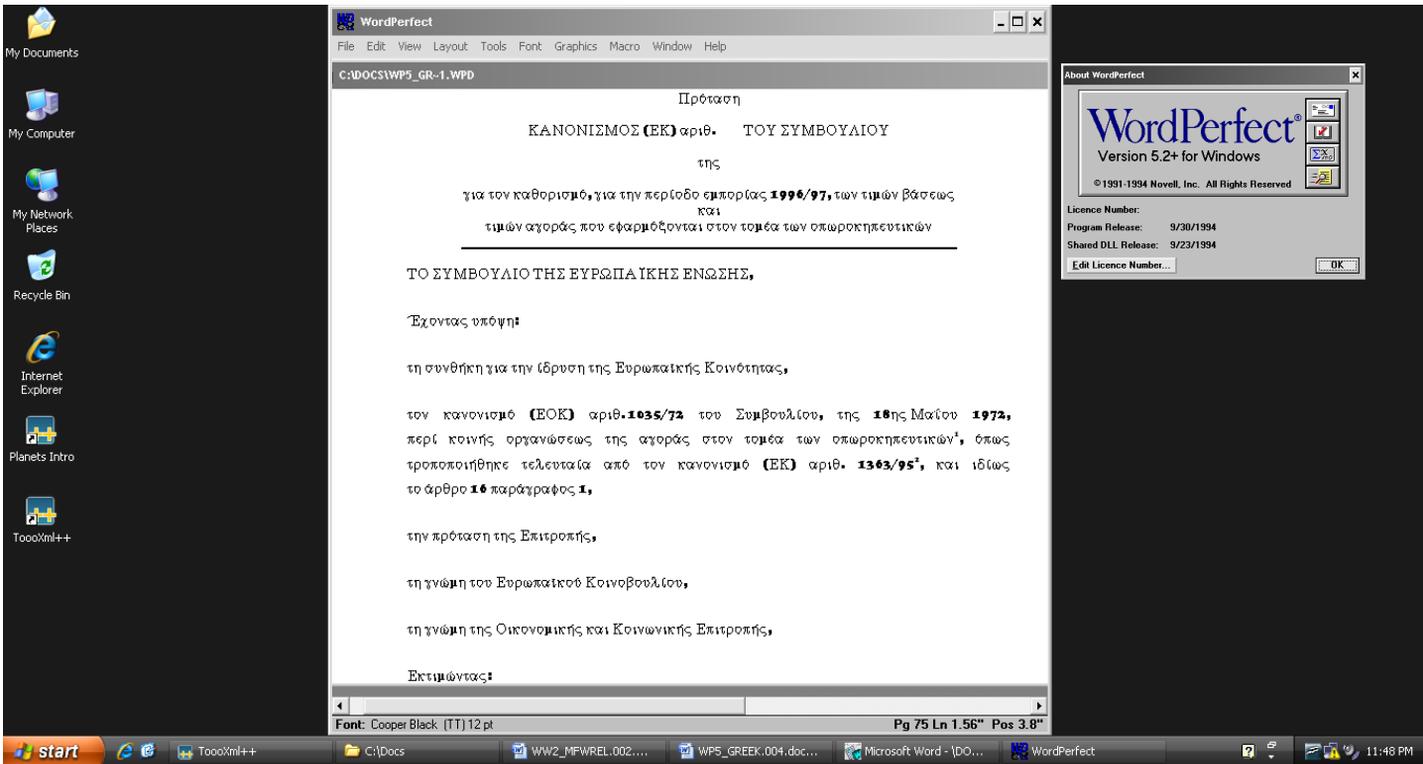
FORMAT TRANSFORMATION



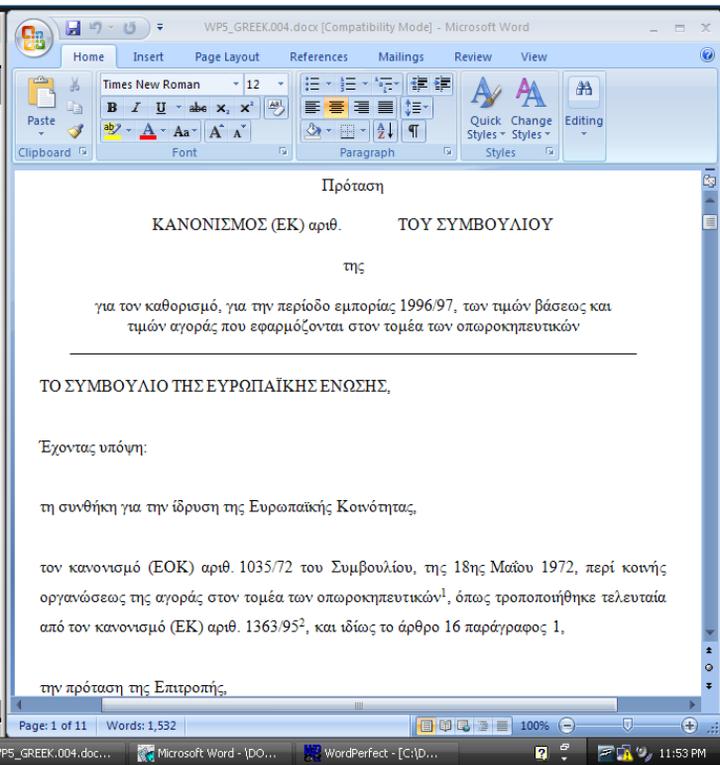
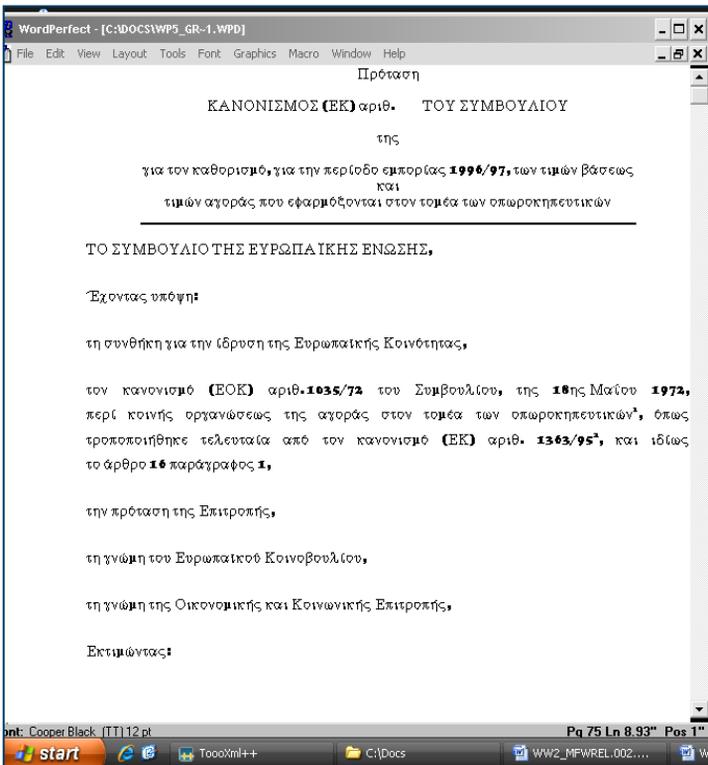
Word document shown in Microsoft Word 2.0 (from 1992)  
Running in the Virtual Machine with Windows XP



Word document in MS Word 2.0 (from 1992) and converted to Open XML format, shown in Office 2007 (right)



Word Perfect document, shown in WordPerfect 5.2 (from 1994)  
Running in the Virtual Machine with Windows XP



Word Perfect document in WordPerfect 5.2 (from 1994) and converted to Open XML format, shown in Office 2007 (right)

# How to cover the cost of long term access?

Manage the services cost to achieve a positive balance between

- value created through immediate explorations of content
- investment needed to sustain perceived value of content in the future.

Assessing the potential value of digital content in the future is difficult.

We can make projections into the near future based on the current needs and opportunities.

Sustainability is possible if the current use scenarios extract sufficient value from the legacy digital assets

# What does the Cloud paradigm offer?

Distributed IT cost and opportunities for extracting value from legacy content:

- Extendible functionality
- Extendible data store
- Scalable computation
- Virtualization
- Common platform for creating services
- Support for client applications on diverse computing platforms.

# Cloud may enable 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.

# Thank you

Natasa Milic-Frayling

[natasamf@microsoft.com](mailto:natasamf@microsoft.com)

Integrated Systems

Microsoft Research Cambridge UK