

# 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

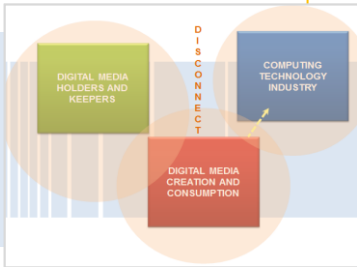
Microsoft Research Cambridge, UK

# Overview



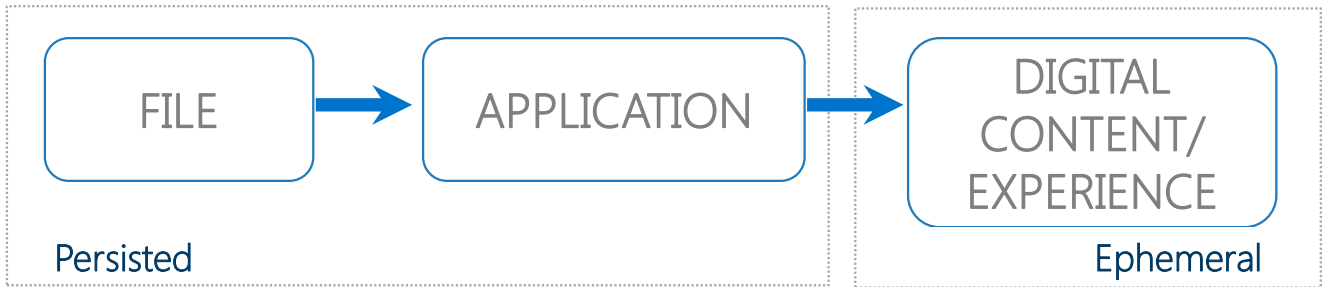
## The Nature of Digital

*Properties and paradoxes of digital*

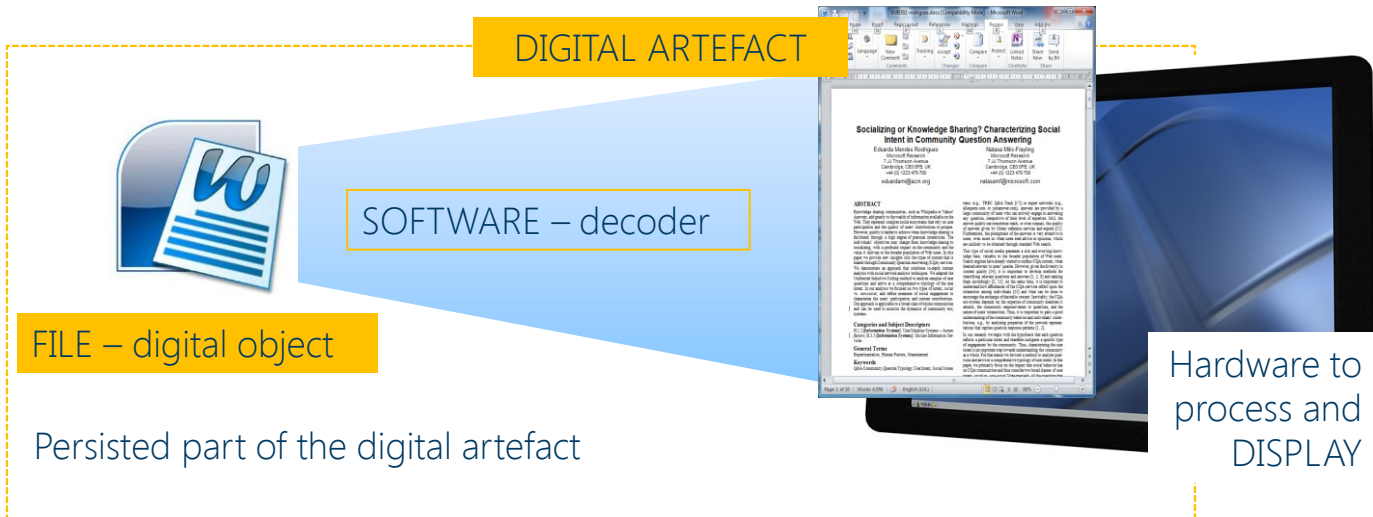


## Sustainability of Digital

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



PRESERVATION = Persistence + Connection with the contemporary ecosystem.



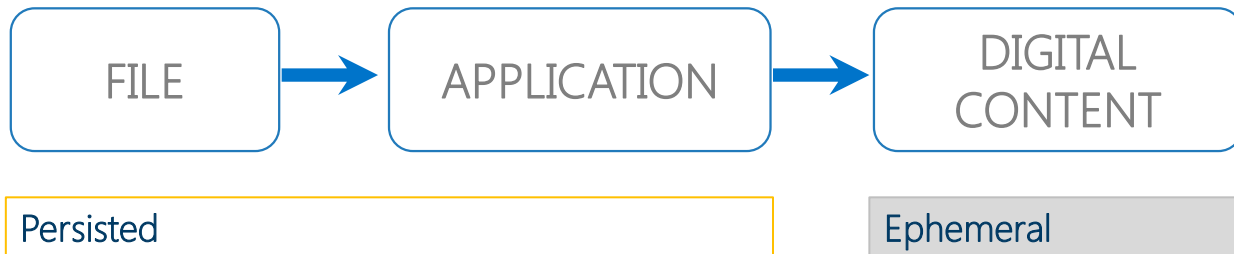
Paradox: we are concerned about storage, yet

*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.

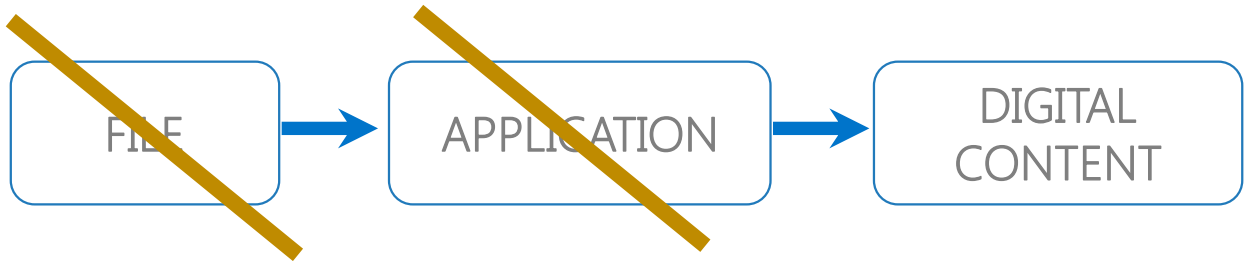


# What do you want to keep 'unchanged'?



- 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'?



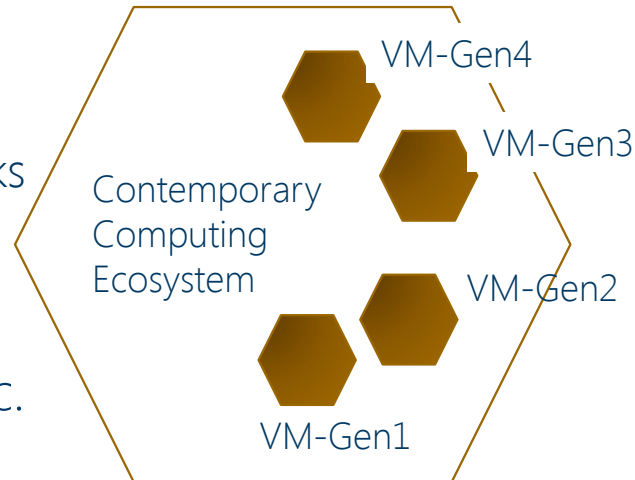
- If application is not running in the contemporary environment
  - Create a virtual machine with the old computing stack and run the original files and software.  
(retain original files and original application; maintain scaffolding)

# 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

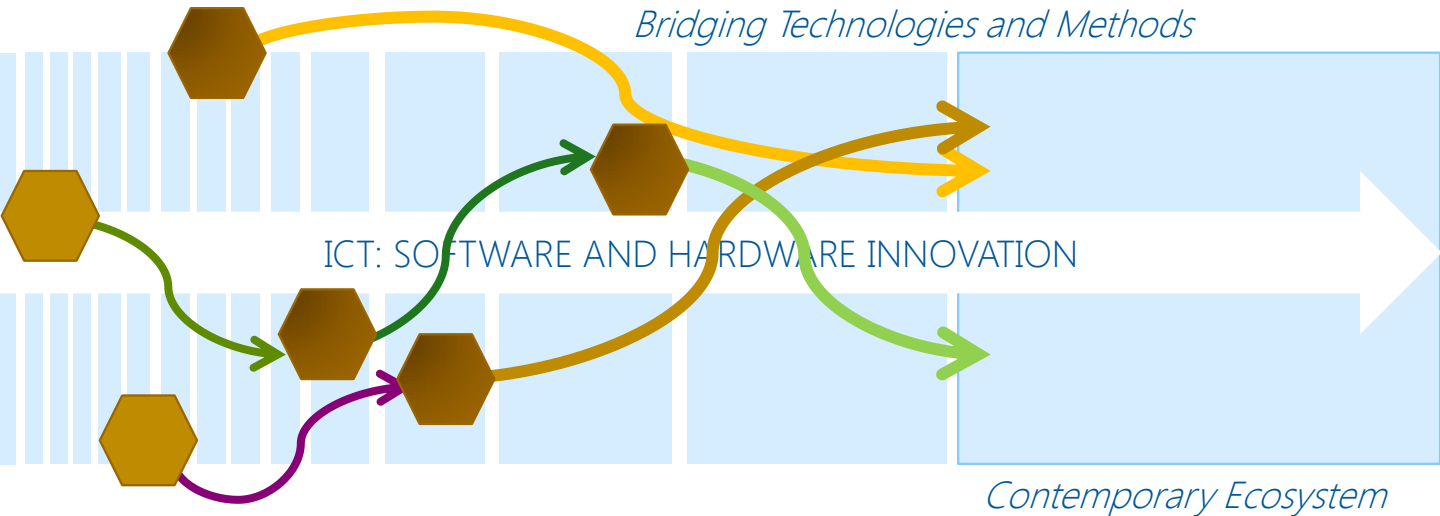
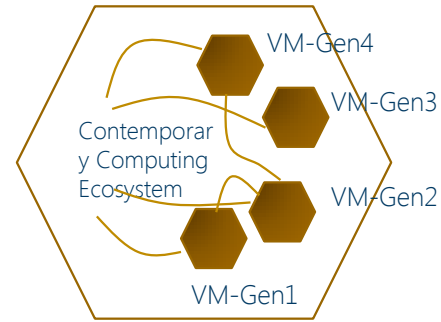
Bridging services: format translators, content extractors, etc.



# 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



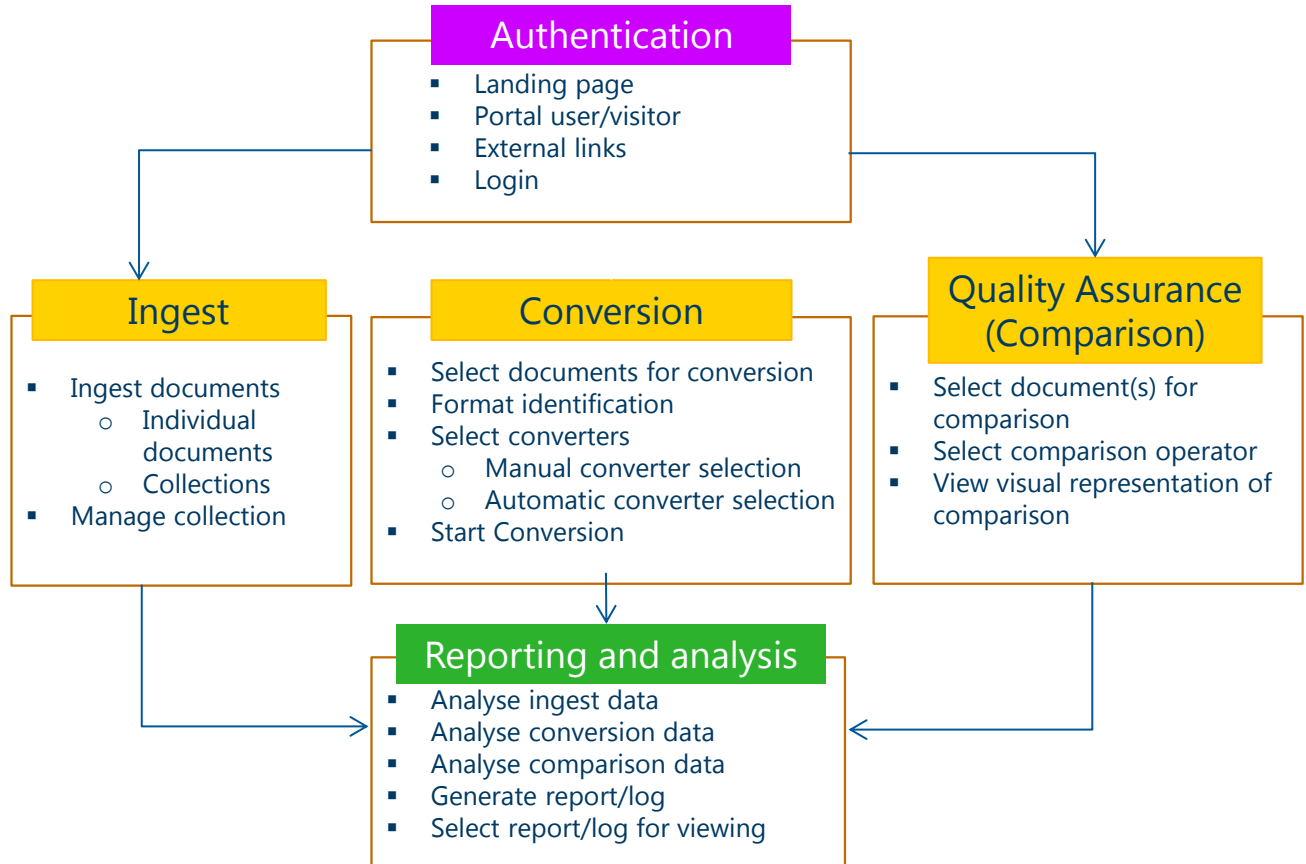
The University of Manchester



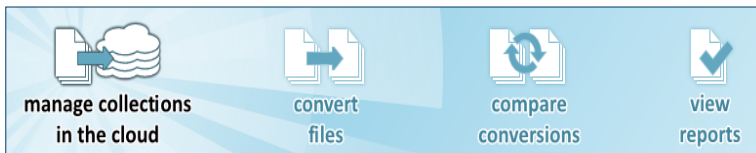
Universite Pierre et Marie Curie – Paris 6



# User Interaction with SAZ Services



# Demo: Use of SCAPE Azure Services



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

## Format Transformation Options

SCAPE Sign In

digital heritage internatio... MUM 2013 About http--cs.gsu (2) SkyDrive Pro PNC 2013 Annual Confere... SIGCHI conferences subm... scape Sign In 2013 IEEE International Co... Natasha Milic-Frayling Cambridge Taxi Request T...

Create Account | Sign In

Microsoft Research

User name: Password: ☐ Keep me signed in

manage collections in the cloud

convert files

compare conversions

view reports

	.docx 2007	.docx 97-2003	.doc 97-2003	.docm macro	.dotx template	.dotm macro	.dot template	.odt	.rtf	.mht	.mhtml	.xml	.png 300x	.pdf 1.0	.dz 1.0	.xps
.docx 2007	•	•														
.doc 97-2003			•													
.docm macro				•												
.dotx template					•											
.dotm macro						•										
.dot template							•									
.odt								•								
.rtf									•							
.mht										•						
.mhtml											•					
.xml												•				
.png 300x													•			
.pdf 1.0														•		

15:19 28/10/2013

# Prototype Features

## Results of the Conversion Process

http://scapestaging.cloudapp.net:8080/Collections/ConvertedFile ... SCAPE: Collections cloudapp.net SCAPE: Convert... SCAPE: Deep Zoom SCAPE: Converted ... Compare Files Compare Files

Welcome, scape3 | [Manage Account](#) | [Sign Out](#)

**SCAPE**  
SCAPE Presentation Environment

Microsoft  
**Research**

Converted Files for the "Collection DIV (50)" Collection

[Back to Collections](#)

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

15:15 28/10/2013

# Prototype Features

## Word'97 Document Transformed into Deep Zoom Format



### SCAPE Digital Object Model

#### Authors

Persons	Role	Partner	Contribution
Matthew Hahn	PM		
Frank Arang	PM		
W. Brumley	PM		
Raf Castro	PM		

#### Distributors

Persons	Role	Partner
SCAPE All		All Partners

#### Revision History

Person	Status	Change	Date	Changes
1.0	Matthew Hahn		2002-09-14	
1.1	Matthew Hahn		2002-09-26	
1.2	W. Brumley		2002-09-27	
1.3	Frank Arang		2002-09-28	
1.4	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	SWOOSIS	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	UP	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	Extension Schemes	21
8.5	Requirements for the OAS Information Packages	22
8.5.1	Definition of a UP	22
8.5.2	Definition of a AP	22
8.5.3	Definition of a DIP	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 discuss that each repository already provides a Digital Object Model but the diversity hinders the SCAPE platform to integrate all the partner repositories. The lack of a Digital Object Model has been put on the SCAPE top register on "Bampernet". 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 a XML container for metadata structure of digital objects. METS is widely used for interoperability between repositories and service components. The PREMIS standard describes a semantic model for preservation metadata, and is widely used in long term preservation. Using METS and PREMIS together will be discussed briefly since there are some loose ends one has 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 be for example for significant differences in between how METS documents as they are used in 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 Storage
3. Data Management
4. Administration
5. Preservation Planning
6. Access

The key terms for this document are DIP (Dissemination Information Package), AP (Archival Information Package) and OP (Dissemination Information Package).

Figure 1: Overview of the OAS Information Packages



Figure 1: Overview of the OAS Information Packages

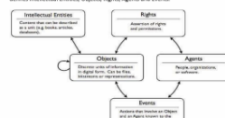
The current DIP definitions of existing SCAPE repositories have significant differences in RODA for example a DIP is a compressed ZIP file that contains a METS envelope, in Rosetta a DIP may contain several DIP (Intellectual Entities). An AP contains technical metadata and metadata important for long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



	metadatas	
Technical metadata	dip	Any form, e.g. METS, METS, PREMIS Object Metadata ... Any form, e.g. PREMIS
Source metadata	dip	info (descriptive, rights, technical) about the existing source used to generate the digital object
Rights metadata	dip	Any form, e.g. rights, copyright, PREMIS Rights Metadata
Digital preservation metadata	dip	Any form, e.g. PREMIS Event Metadata
File location	All files that comprise the content of the digital entity. Files are ordered in groups (DIP, group 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 jpg in a PDF document, audio data within a MP3 file or graphics within a word document.

#### 5 METS & PREMIS

METS is an XML container for structuring metadata in different formats is often used in conjunction with the PREMIS standard for preservation metadata. But one has to consider the existing effort 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:



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

## Quality Assessment of the Conversion Process

The screenshot displays the SCAPE web application interface, which is designed for managing digital collections and performing format transformations. The interface is divided into several sections:

- Header:** Includes the SCAPE logo (Storable Content Preservation Environment) and the Microsoft Research logo. Navigation links for "Create Account" and "Sign In" are present.
- Authentication:** A login section with fields for "User name:" and "Password:", a "Keep me signed in" checkbox, and a "Sign In" button.
- Main Navigation:** Four large icons represent the core functions: "manage collections in the cloud", "convert files", "compare conversions", and "view reports".
- Document Viewer:** The central area shows two side-by-side document previews. The left document is titled "7: digital Object Model of SCAPE Preservation" and features a diagram illustrating the workflow: SIP (Submission Information Package) → AIP (Archival Information Package) → DIP (Dissemination Information Package). The right document is a comparison of conversions, showing differences between two versions of the same document.
- Footer:** Includes a "Synchronize pages" checkbox, a "Show full screen" checkbox, and page navigation controls (Page 10 of 30 and Page 10 of 29).

The interface is running in a web browser, with the address bar showing the URL: <http://scapestaging.cloudapp.net:8080/Default.aspx?ReturnUrl=%2F>. The taskbar at the bottom shows various application icons, including Internet Explorer, Word, and PowerPoint.

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

- ☒ Synchronize pages
- ☒ Show full screen

# 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 DOCX)

Line Difference (B2X DOCX)

Word Difference (B2X DOCX)

Good Match Quality (%)

Bad Match Quality (%)

Partial Match Quality (%)

No Match (%)

Average Mismatched Rectangles

☒ Show full screen

- to

121549 to Cynthia Singh

Dan Ogle, Plant Materials Specialist to Irene Fields

J Clark Salyer to NWTRB

occ to tsong

U.S. Department of Education to Yvette Torres

(no info)



# Winning strategy

## Virtualization of original software

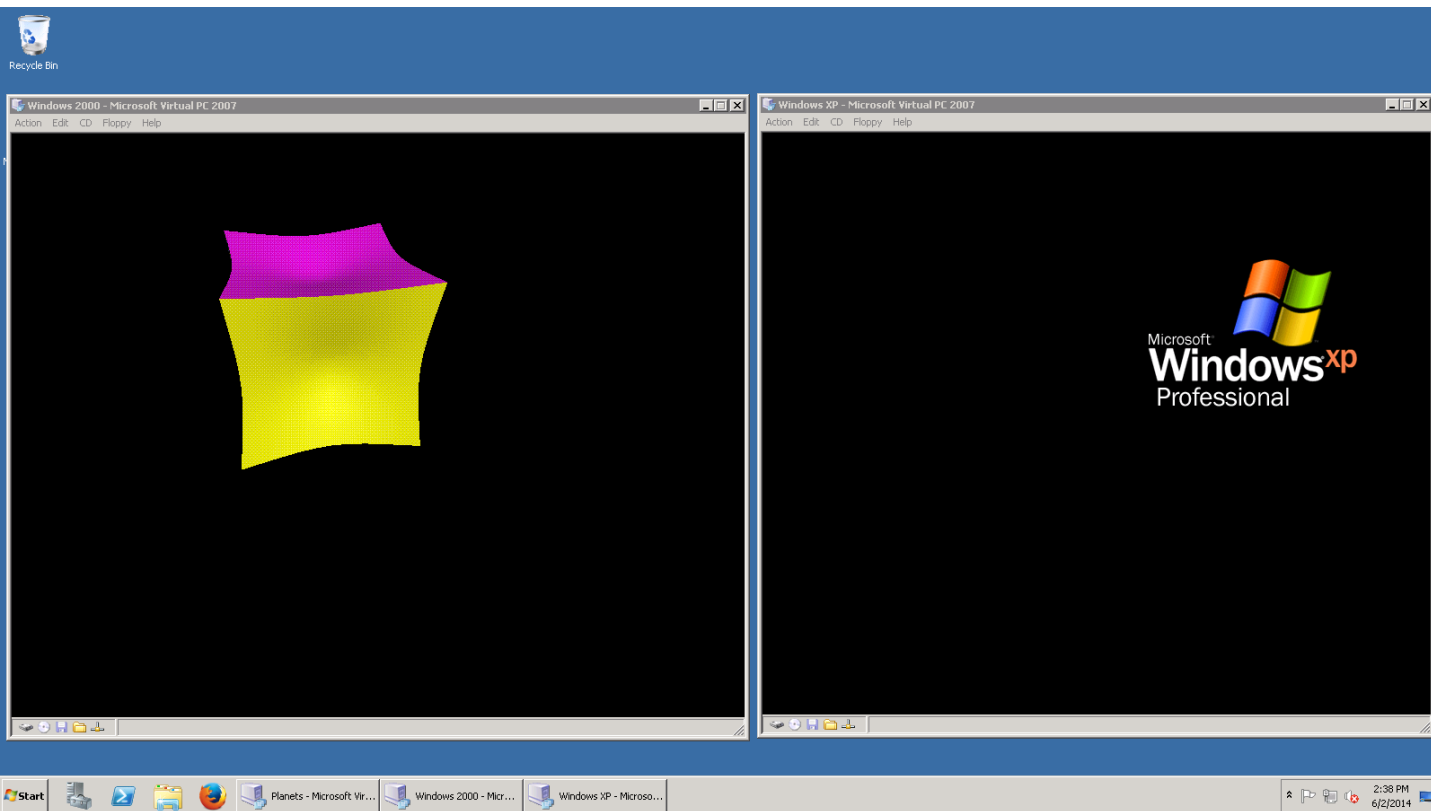
Ensures access to the digital artefacts

## Format transformation services

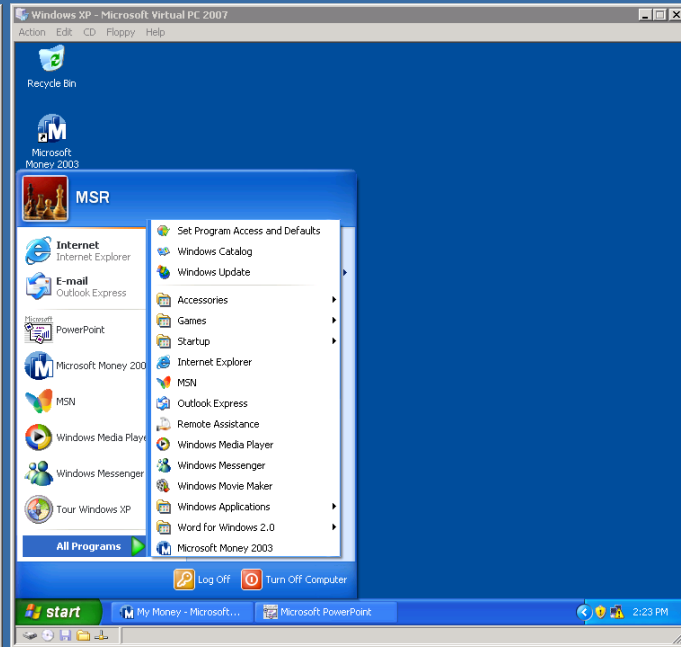
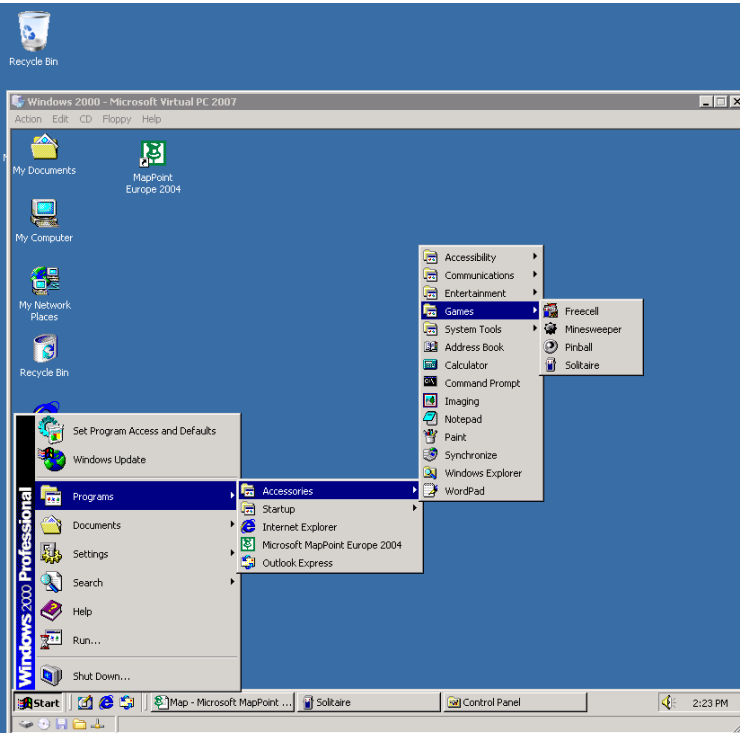
On demand transformation within a specific context.

preserving computation

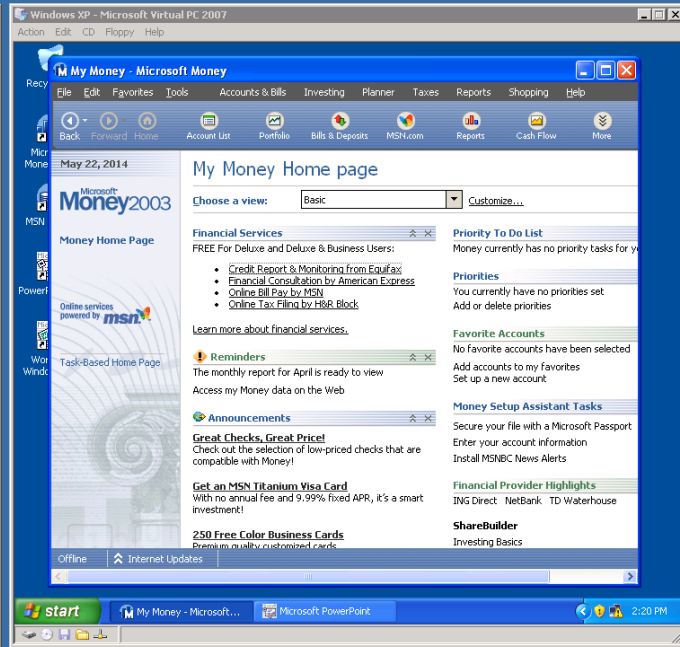
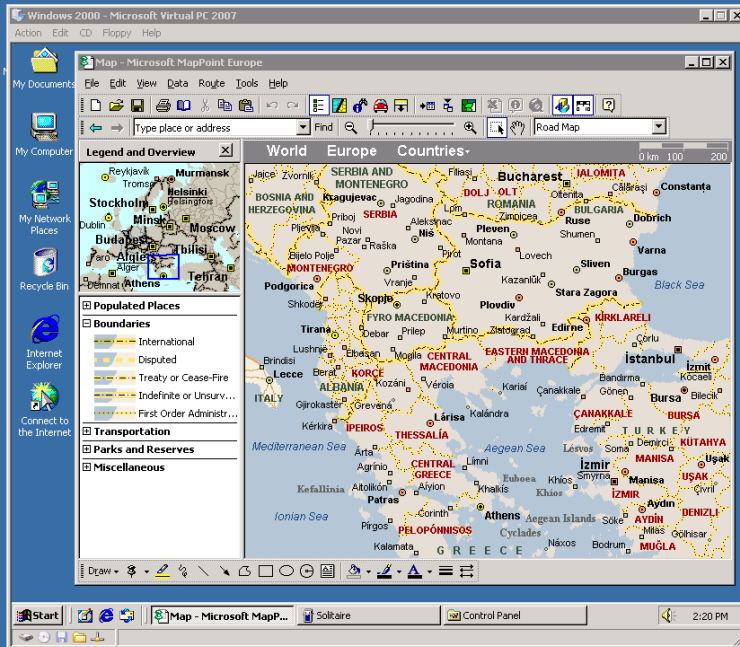
# VIRTUALIZATION OF LEGACY SOFTWARE



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



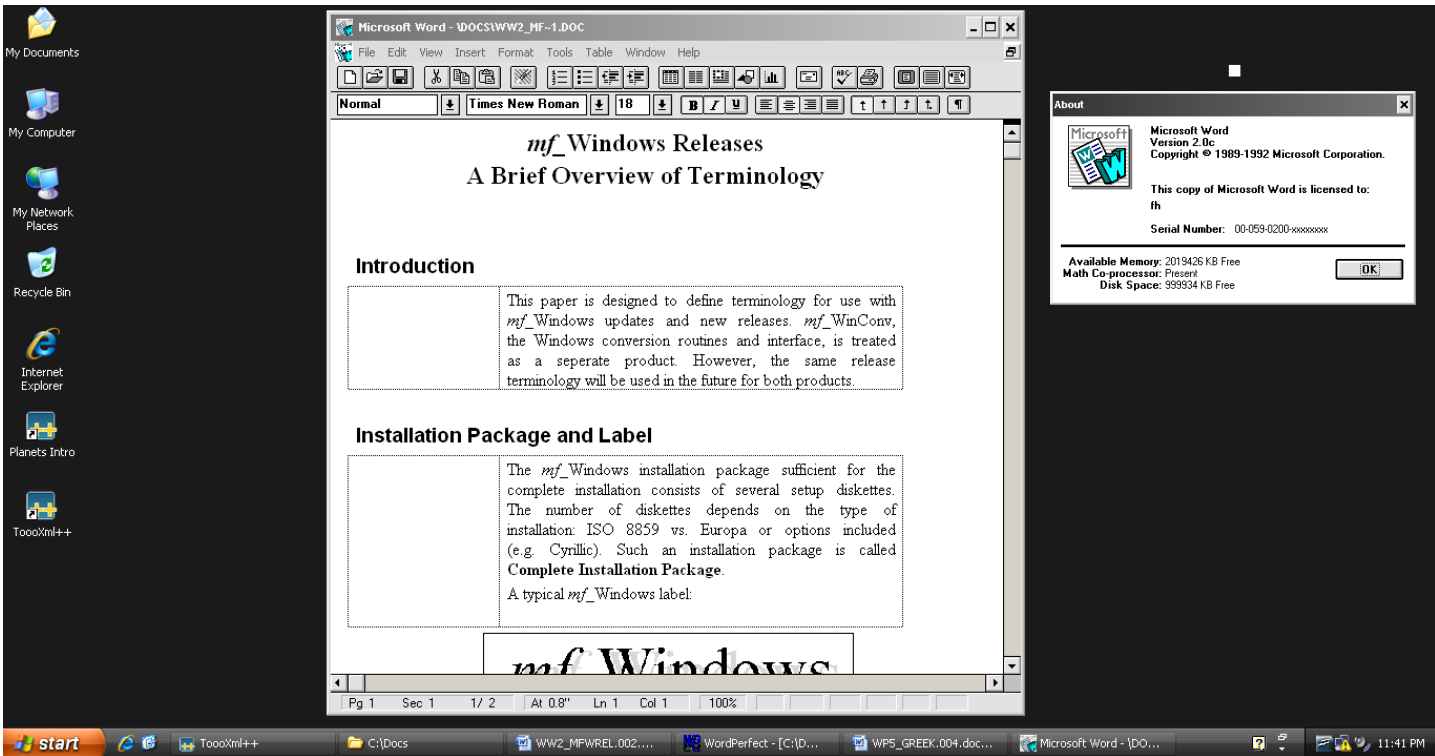
Start menus for Windows 2000 (left) and Windows XP (right),



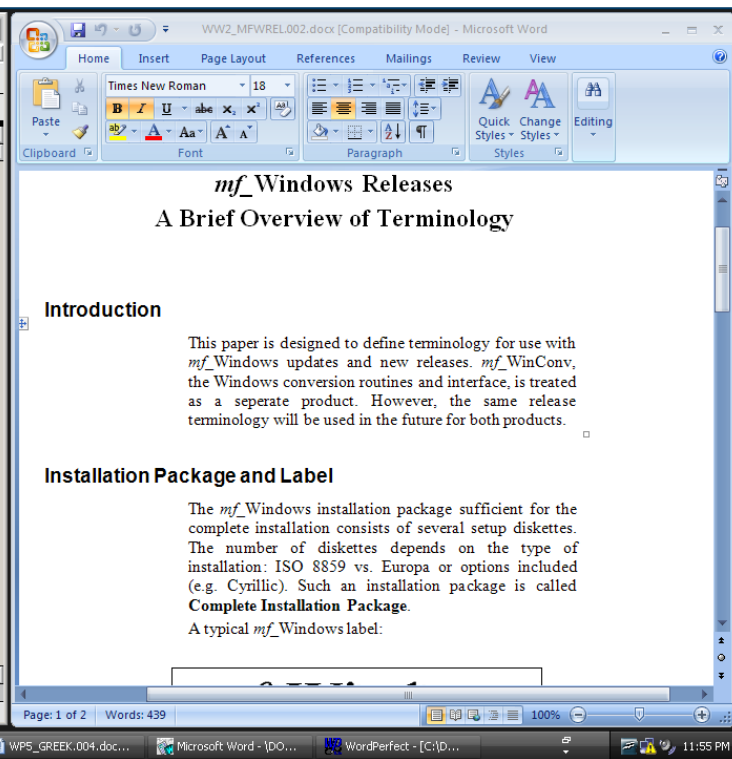
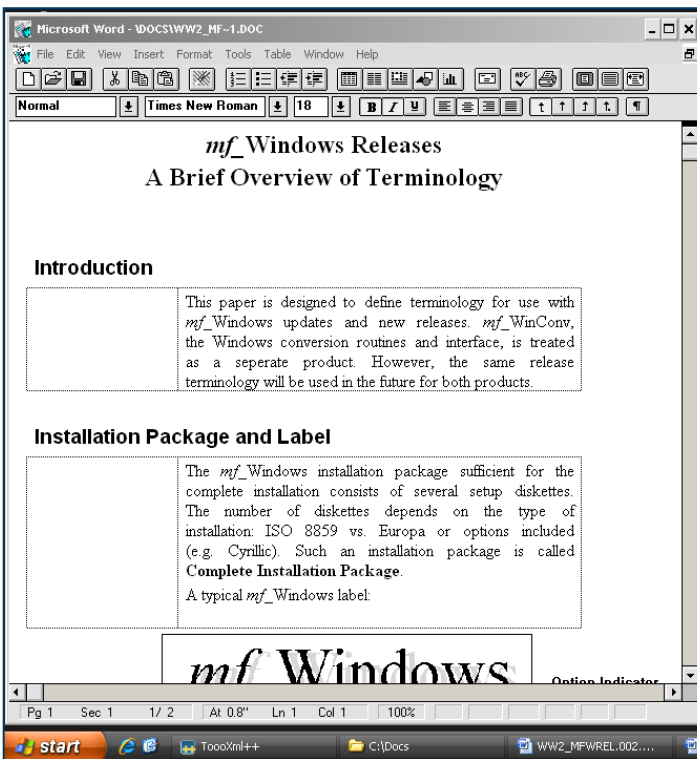
MS Map Point application running on Windows 2000 (left) and MS Money 2003 running on Windows XP (right),

Increasing value of legacy content

FORMAT TRANSFORMATION

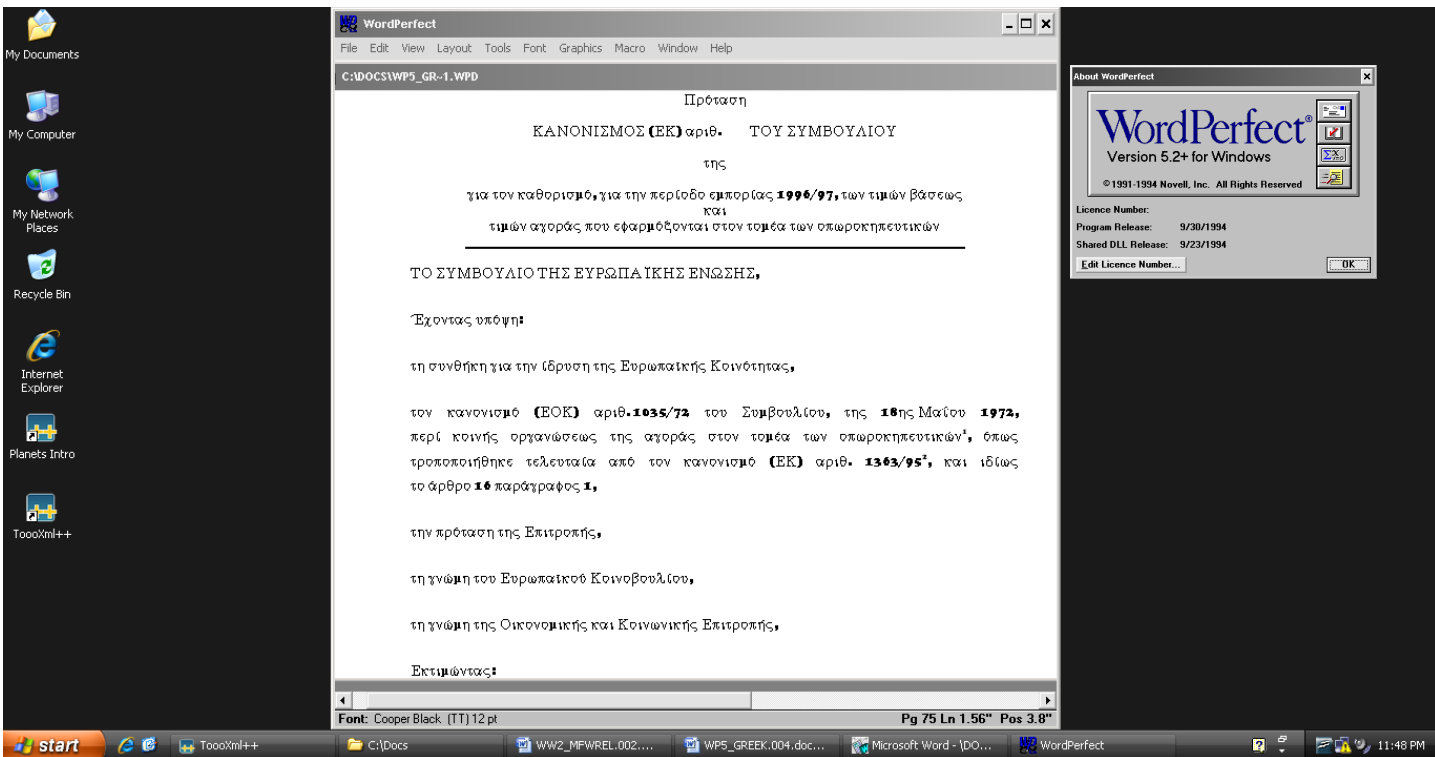


Word document shown in Microsoft Word 2.0 (from 1992)  
Running in the Virtual Machine with Windows XP

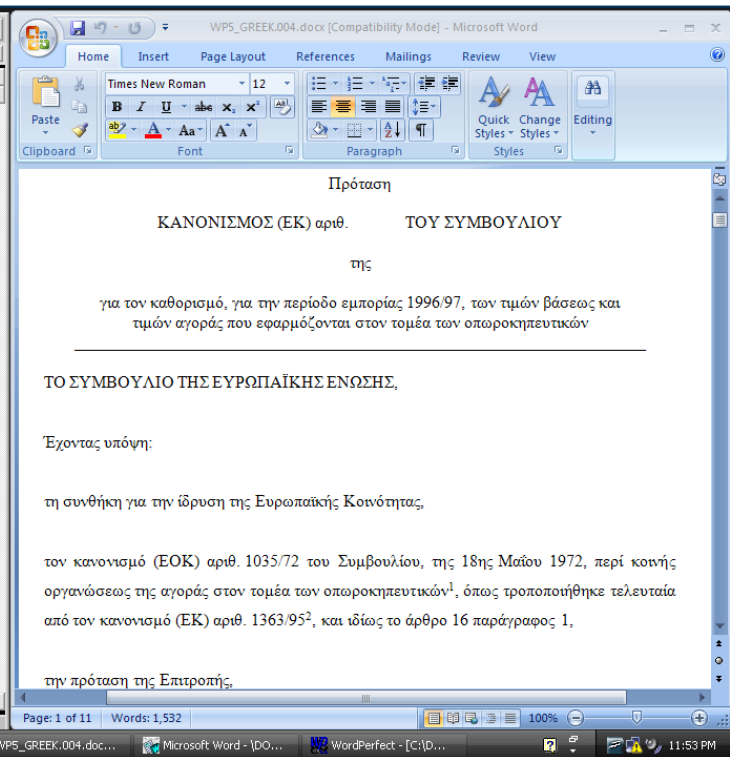
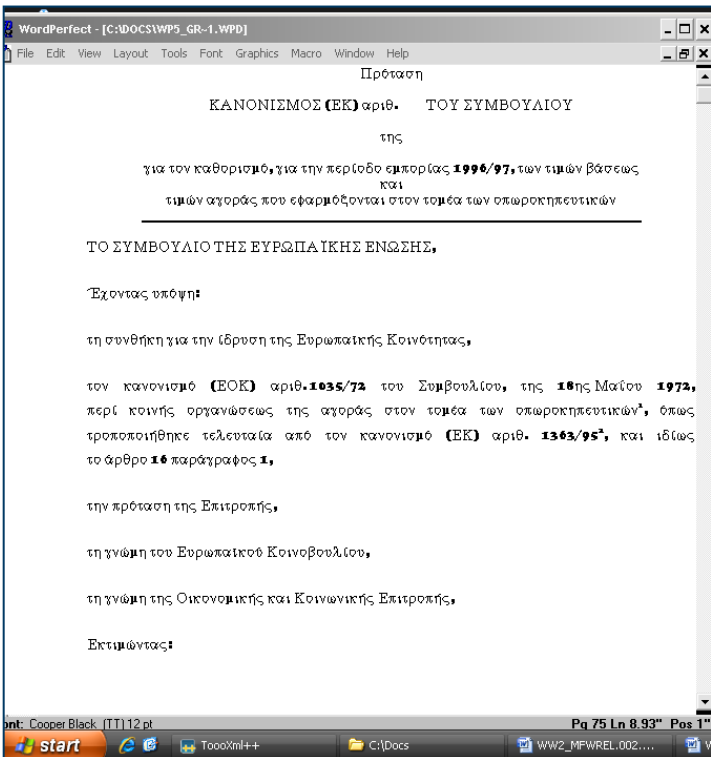


Word document in MS Word 2.0 (from 1992) and converted to Open XML format, shown in Office 2007 (right)





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