

Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Preserving Digital Content

A Short Introduction to Digital Information in the Preservation Context

Digital Preservation – The Planets Way
Copenhagen, 22 – 24 June 2009

Volker Heydegger

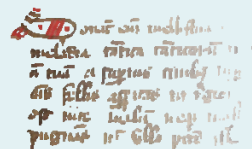
Overview

- ❑ Two ways of information representation
(what is digital content?)
- ❑ Shapes of digital content
(Which logical and technical units of digital content are important in the preservation context?)
- ❑ Preservation issues and challenges
(How can we assure preservation of digital content?)

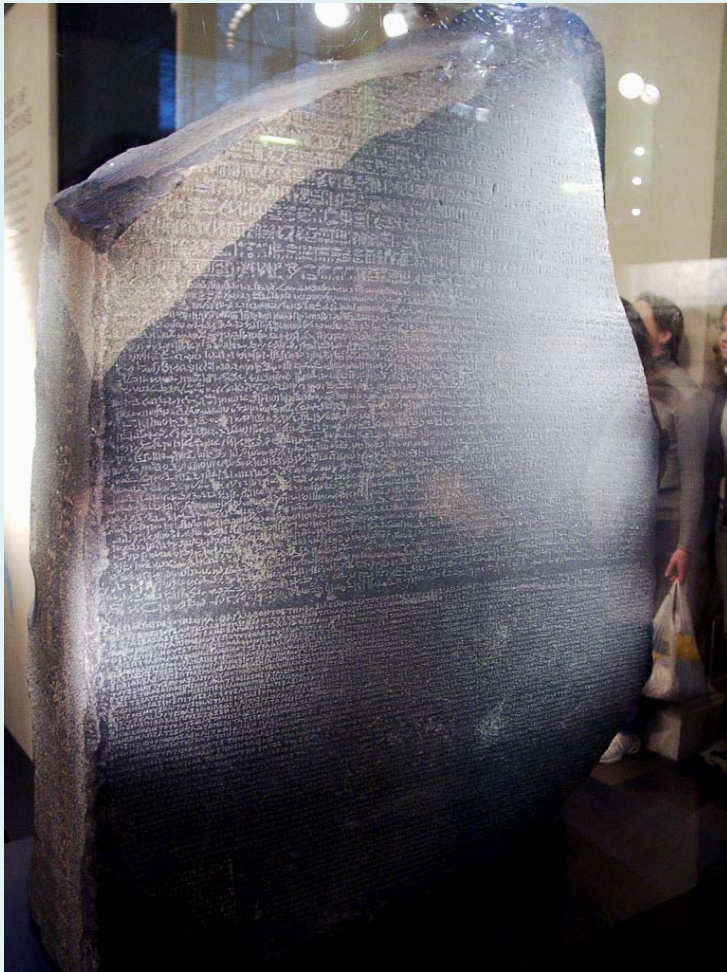


Two Ways of Information Representation

„In the reign of the young one who has succeeded his father in the kingship, lord of diadems, most glorious, who has established Egypt and is pious towards the gods, triumphant over his enemies, who has restored the civilized life of men, lord of the Thirty Years Festivals, even as Ptah the Great, a king like Ra, great king of the Upper and Lower countries, offspring of the Gods Philopatores, one whom Ptah has approved, to whom Ra has given victory, the living image of Amun, son of Ra, Ptolemy, Living or ever, beloved of Ptah, in the ninth year, when Aetos son of Aetos was priest of Alexander, and the Gods Soteres, and the Gods Adelphoi, and the Gods Euergetai, and the Gods Philopatores and the God Epiphanes Eucharistos; Pyrrha daughter of Philinos being Athlophoros of Berenike Euergetis, Areia daughter of Diogenes being Kanephoros of Arsinoe Philadelphos; Irene daughter of Ptolemy being Priestess of Arsinoe Philopator; the fourth of the month of Xandikos, according to the Egyptians the 18th Mekhir. „



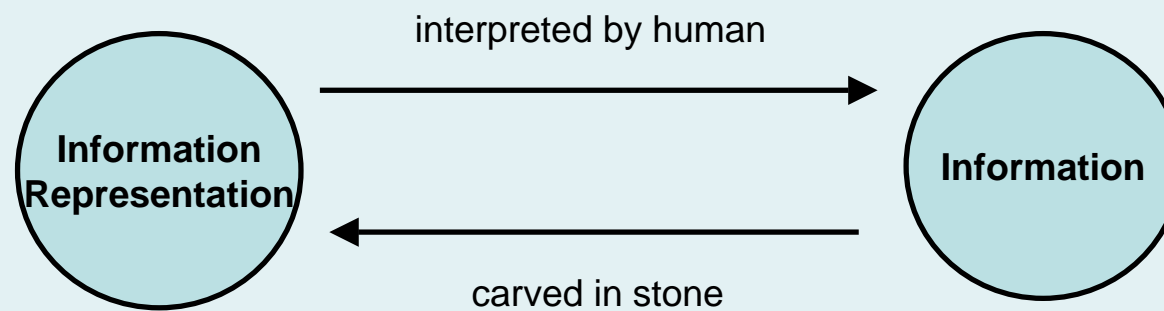
Information representation – 2205 years ago



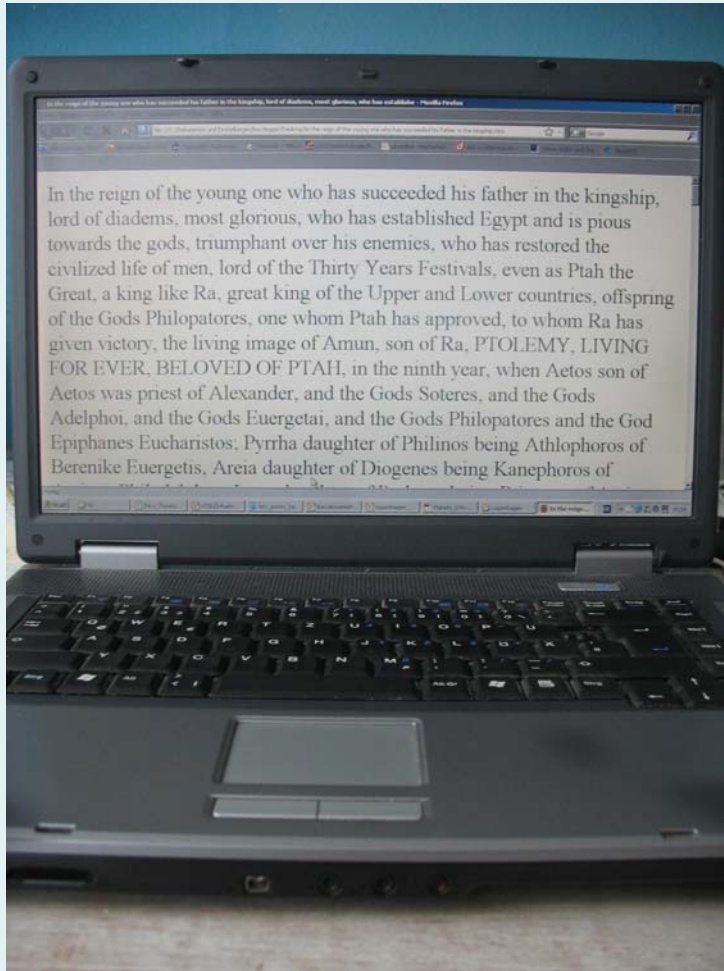
- ❑ Carrier
 - Solid material (granodiorite)
 - 114 x 72 x 28
 - 760 kg
- ❑ Encoding
 - Human-readable characters
 - Three language scripts (hieroglyphic, demotic, ancient greek)
- ❑ How to get the information?
 - Human, capable of reading (at least) one of the scripts



Information cycle – 196 BC



Information representation – 2009



❑ Hardware

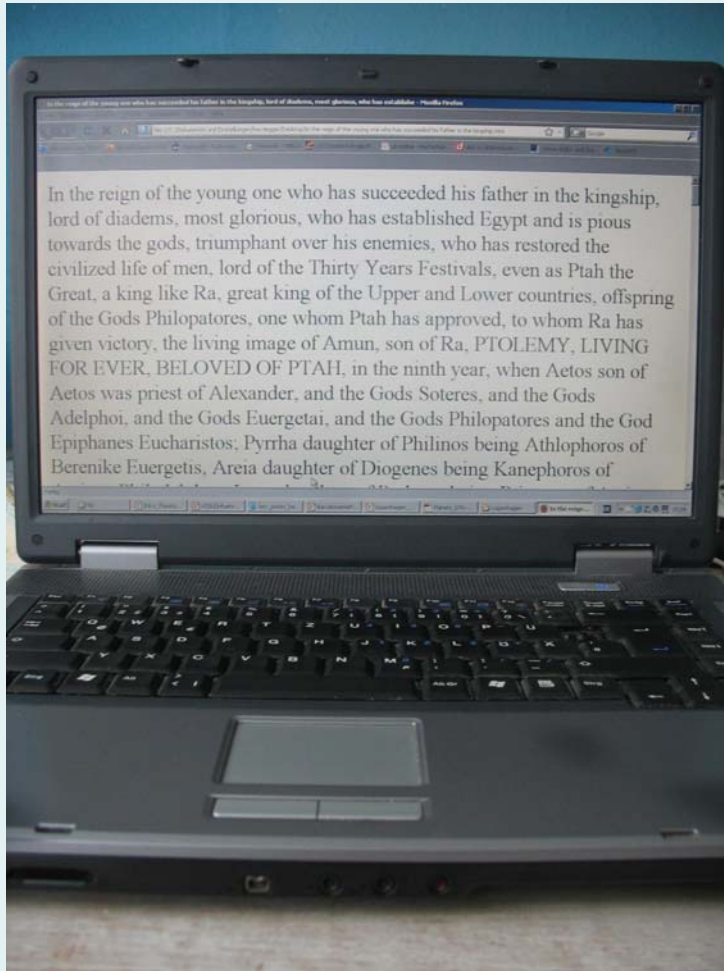
- Storage medium (hard disk, optical disc, ...)
- Rendering environment (display, printer, ...)

❑ Software

- Low level software (operating system)
- Application software (webbrowser, texteditor, ...)



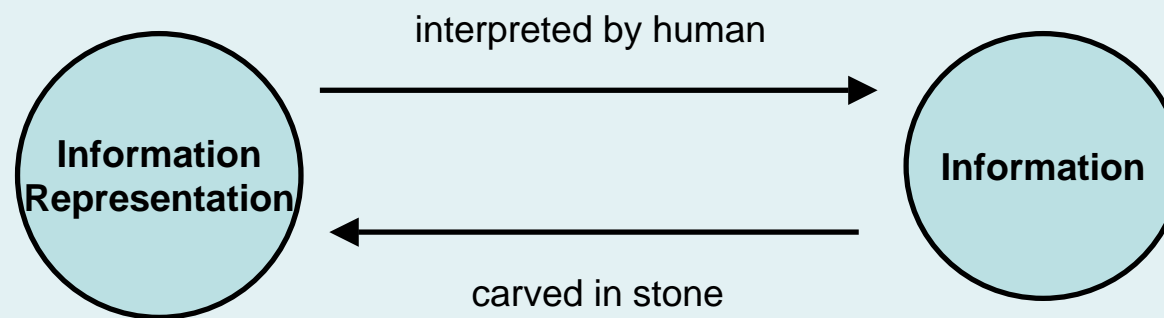
Information representation – 2009



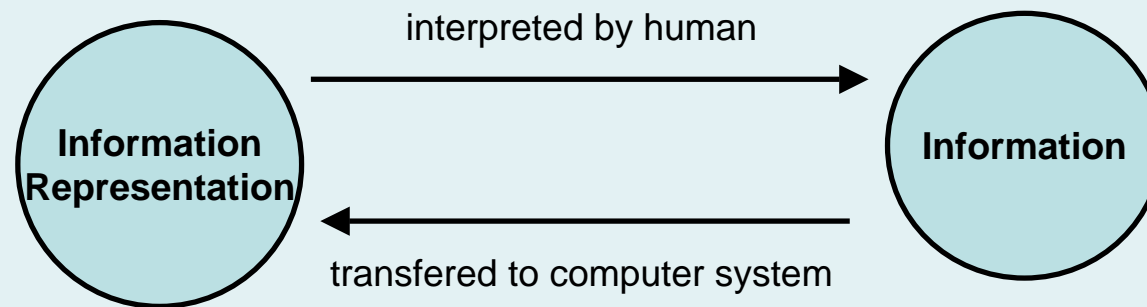
- ❑ Encoding
 - Machine-readable: Binary data
 - Human-readable: Characters
- ❑ How to get the information?
 - Human, capable of understanding english language
 - We need software
 - We need representation facilities



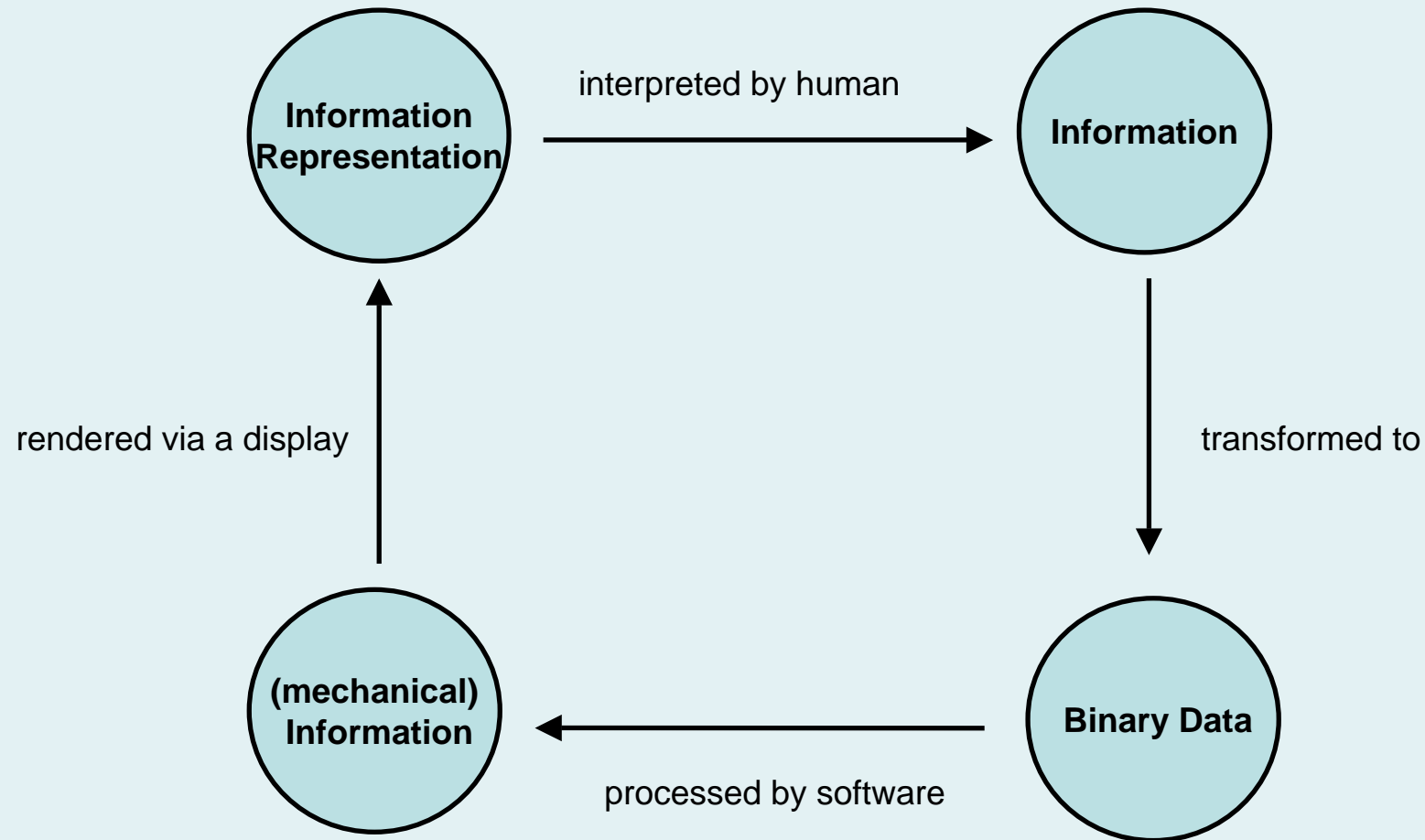
Information cycle – 196 BC



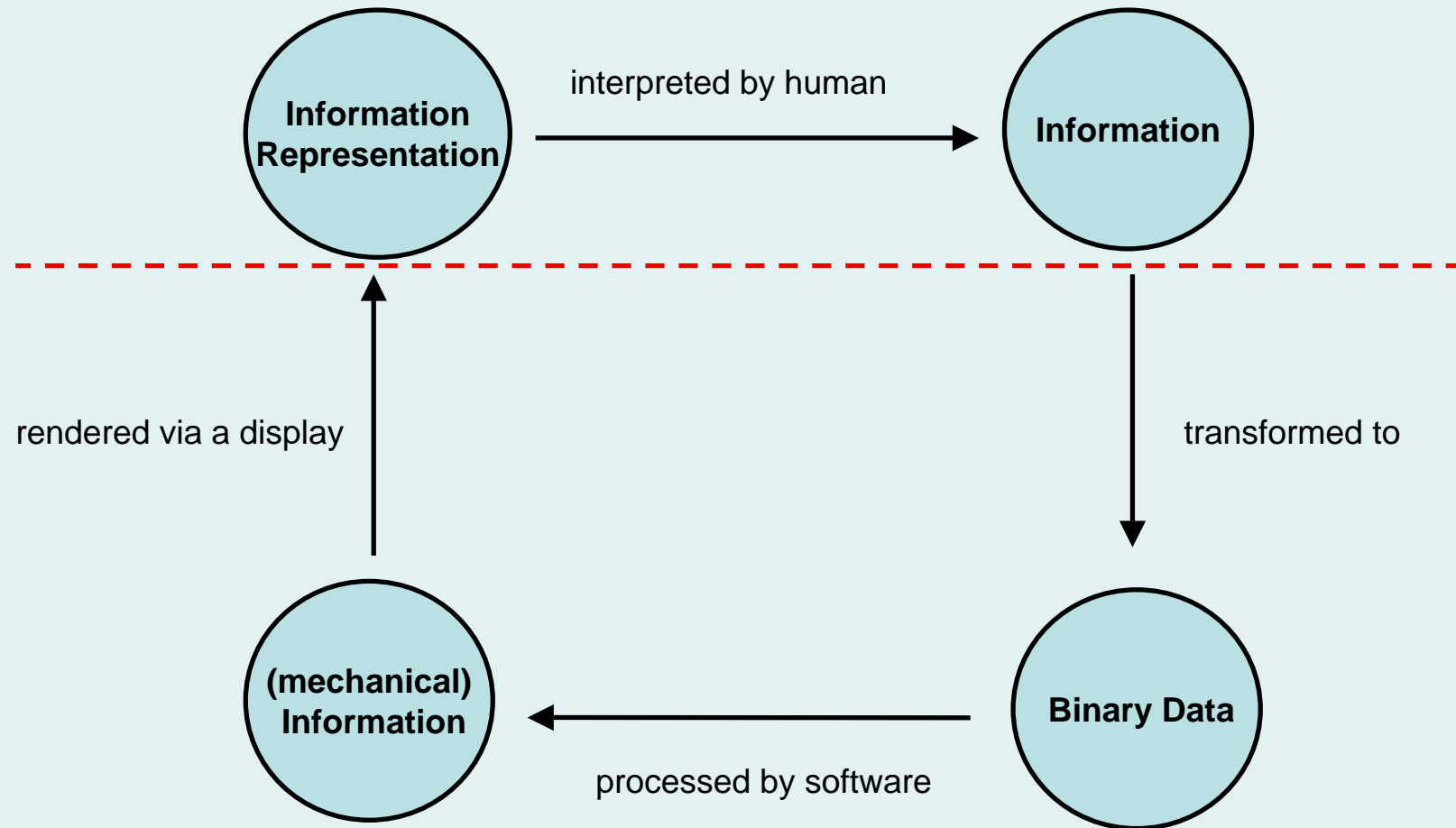
Information cycle – 2009 AC (simplified)



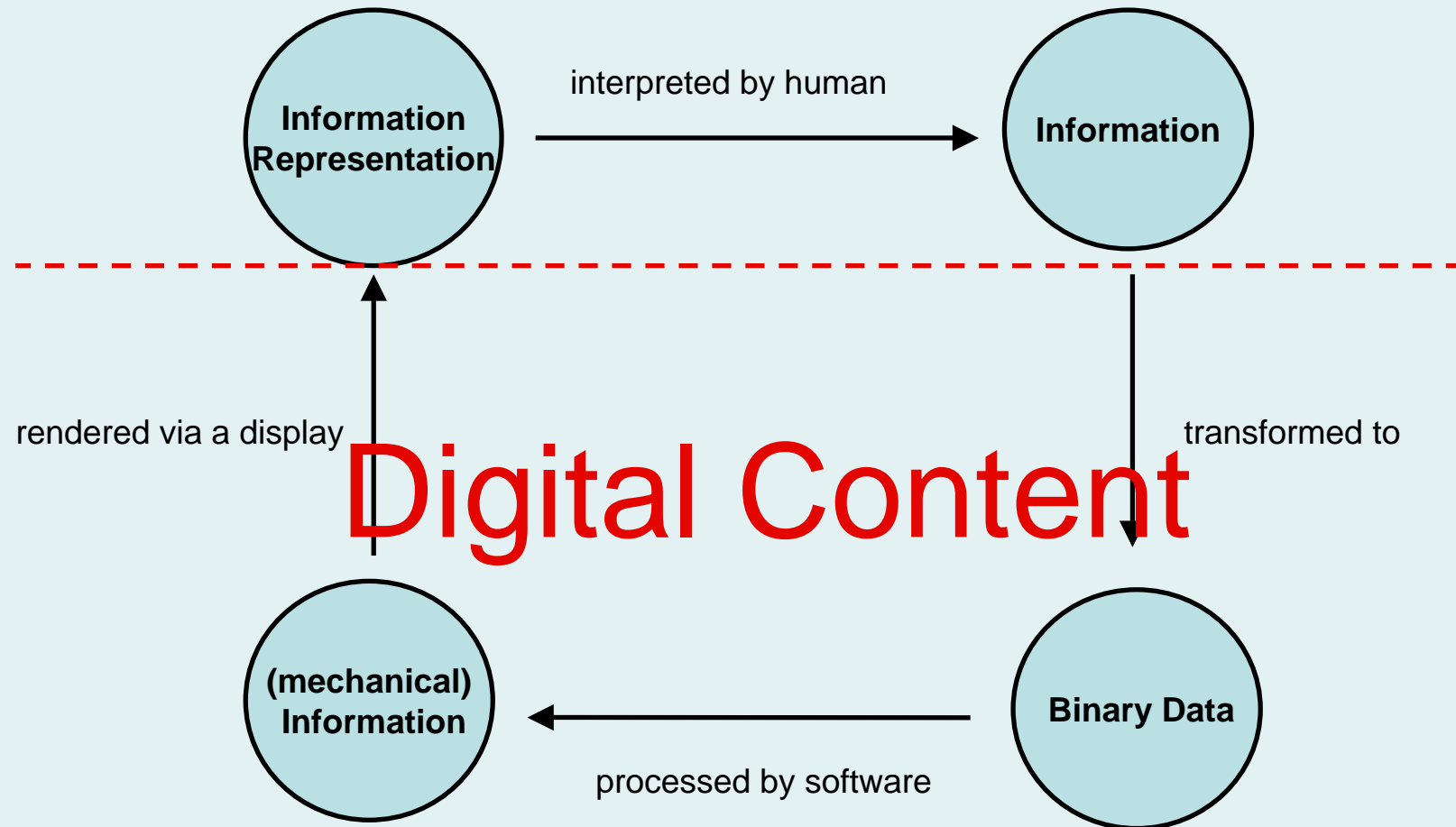
Information cycle – 2009 AC



Information cycle – 2009 AC



Information cycle – 2009 AC



What is digital content ?

- ❑ Information which
 - is encoded in a specific way
 - must be processed by the means of computers
 - must be presented by technical means



Shapes of Digital Content

- ❑ On a the storage level digital content is just **binary data**

011100110001110100011010...

- ❑ ???

- ❑ On the most human-perceivable level digital content appears in a rendered form



Shapes of Digital Content

- ❑ On a very basic level (storage level) digital content is just **binary data**
- ❑ On the software level: File
Operating system: Sequence of bytes
Application program: *meaningful* sequence of bytes
→ **File (Data) Format**
- ❑ On the most human-perceivable level it appears in a rendered form

011100110001110100011010...

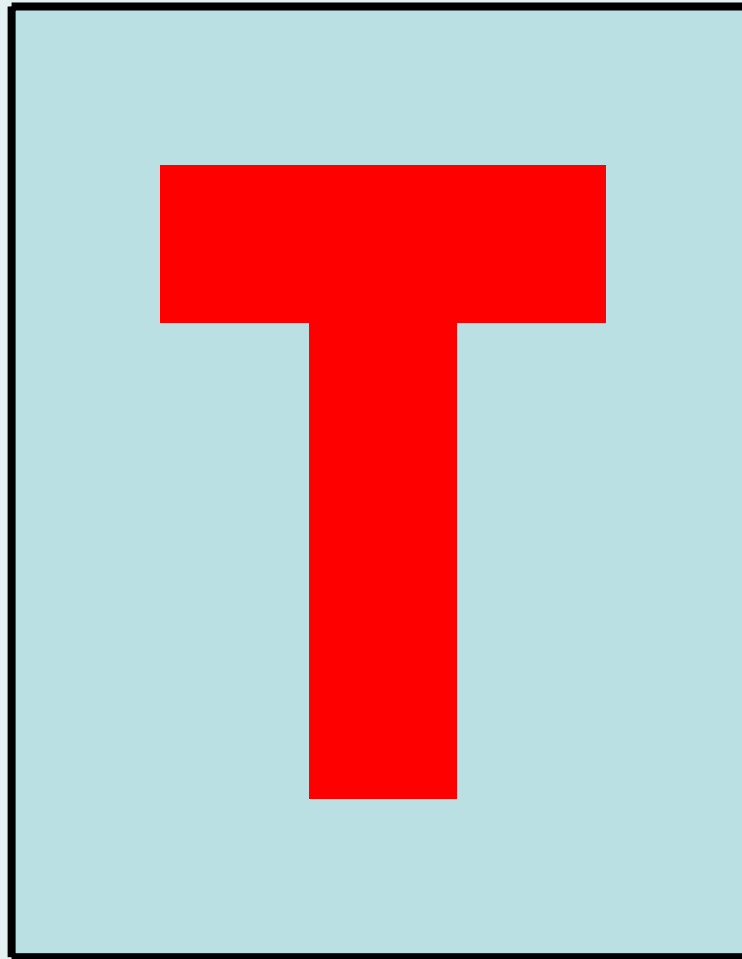


Shapes of Digital Content: File Format

What is in a format?

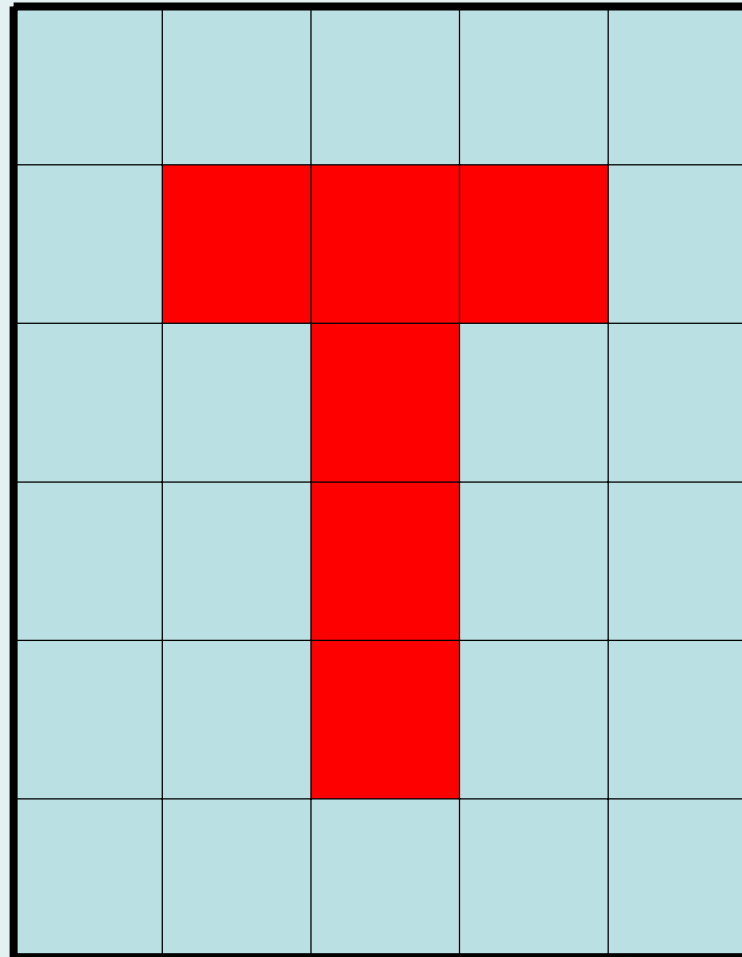


(Significant) characteristics/ properties

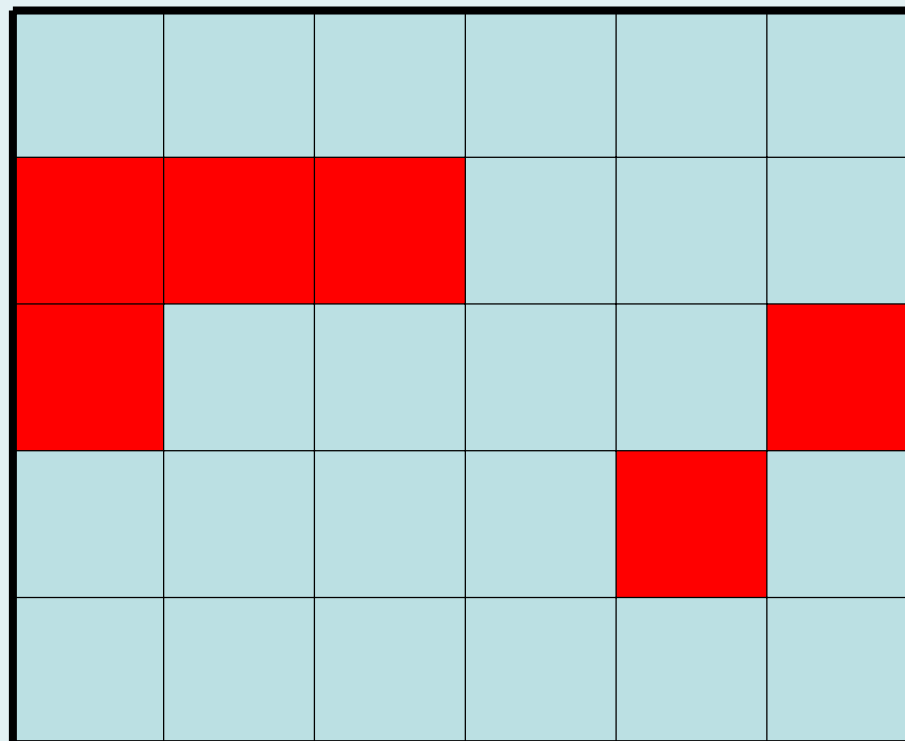


6 rows
5 columns

→ dimensions



5 rows
6 columns



1 == blue

0 == red

→ photometric
Interpretation

1	1	1	1	1
1	0	0	0	1
1	1	0	1	1
1	1	0	1	1
1	1	0	1	1
1	1	1	1	1



1 == green
0 == yellow

1	1	1	1	1
1	0	0	0	1
1	1	0	1	1
1	1	0	1	1
1	1	0	1	1
1	1	1	1	1



Store:

1,1,1,1,1,1,0,0,0
,1,1,1,0,1,1,1,1,
0,1,1,1,1,0,1,1,1
,1,1,1,1

uncompressed

1	1	1	1	1
1	0	0	0	1
1	1	0	1	1
1	1	0	1	1
1	1	0	1	1
1	1	1	1	1



Store:

6,1,3,0,3,1,
1,0,4,1,1,0,
4,1,1,0,7,1

compressed
(Run Length
Encoding)

1	1	1	1	1
1	0	0	0	1
1	1	0	1	1
1	1	0	1	1
1	1	0	1	1
1	1	1	1	1



Properties (Characteristics)

6 rows

5 columns

→ dimensions

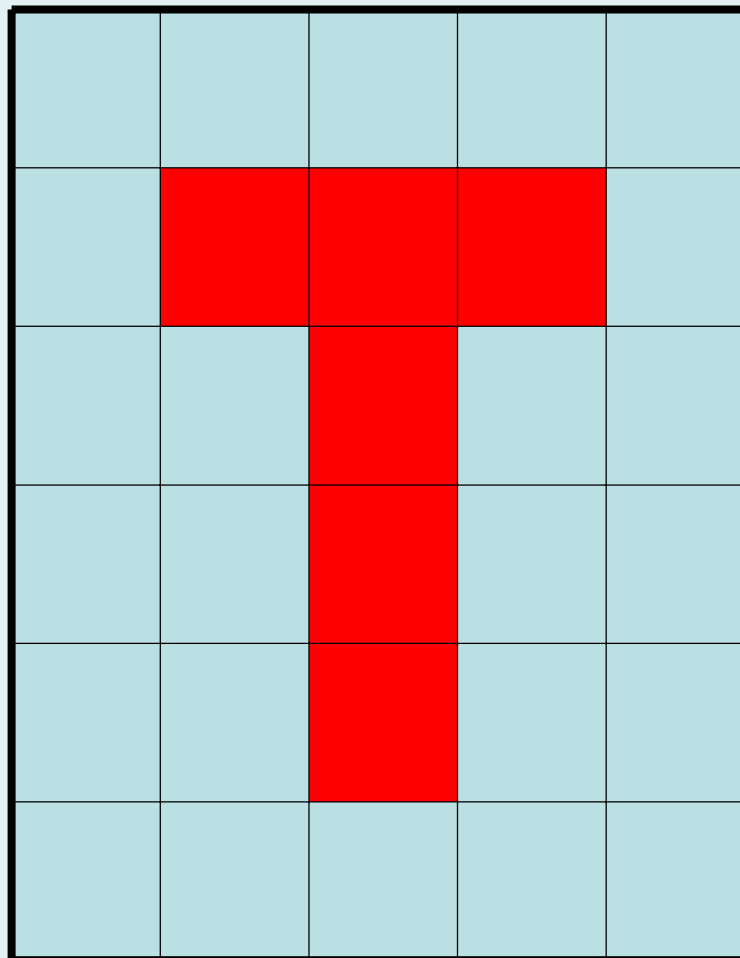
1 == blue

0 == red

→ photometric
interpretation

uncompressed

→ compression



Properties in files (example: tiff)

The screenshot shows a hex editor window with the following data:

Offset	Hex	ASCII
0:	49 49 2A 00 18 00 00 00 7D 00 00 00 01 00 00 00	II*.....}
16:	7D 00 00 00 01 00 00 00 0E 00 FE 00 04 00 01 00	}.....p.....
32:	00 00 00 00 00 00 00 01 03 00 01 00 00 00 15 01S.....
48:	00 00 01 01 03 00 01 00 00 00 53 01 00 00 02 01E.....
64:	03 00 01 00 00 00 08 00 00 00 03 01 03 00 01 00S.....
80:	00 00 01 00 00 00 06 01 03 00 01 00 00 00 01 00A.....
96:	00 00 11 01 04 00 01 00 00 00 C6 00 00 00 15 01%.....
112:	03 00 01 00 00 00 01 00 00 00 16 01 04 00 01 00&.....
128:	00 00 53 01 00 00 17 01 04 00 01 00 00 00 CF 6E\$.....
144:	01 00 1A 01 05 00 01 00 00 00 08 00 00 00 1B 01
160:	05 00 01 00 00 00 10 00 00 00 28 01 03 00 01 00
176:	00 00 02 00 00 00 41 01 03 00 02 00 00 00 CB 00
192:	08 00 00 00 00 00 1A 19 25 19 17 16 17 18 18 15
208:	17 15 17 17 18 19 17 18 19 1C 19 15 17 15 15 17
224:	15 15 16 15 19 17 17 18 18 16 19 18 18 19 19 19
240:	26 19 19 19 1A 1A 18 16 15 16 15 13 13 13 15 16
256:	18 1A 1A 1A 19 18 17 17 18 19 18 17 18 15 18 17 1A
272:	1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 18
288:	16 16 15 16 14 15 14 16 16 18 1A 18 16 1A 20 1F
304:	1C 1C 1C 1D 1F 18 1A 1C 19 18 1A 1D 1E 18 19 1A
320:	18 1C 1F 1E 18 18 1C 1A 18 18 16 18 17 17 18 18
336:	19 19 18 17 18 1C 1F 24 1F 1C 1C 1B 1A 17 18 1A
352:	1D 1A 18 18 16 18 17 19 18 18 17 18 18 1C 18 18
368:	17 19 18 18 17 18 15 16 16 18 19 1B 1B 1C 18 19
384:	1D 1B 18 17 18 17 17 18 16 19 17 16 17 16 16 18
400:	17 17 18 18 19 17 16 17 18 17 17 18 17 17 16 15
416:	13 13 14 15 16 16 18 1C 18 18 15 15 14 17 18 17
432:	17 17 16 16 17 17 15 16 16 14 13 15 15 15 15 16
448:	17 16 16 15 15 15 16 15 15 15 15 17 16 16 18 1C
464:	19 1C 19 15 14 14 15 15 15 19 15 1A 1A 1C 1A 1A
480:	1A 1B 1C 1B 1A 1A 19 1A 1C 1D 1C 1C 1C 1C 1E 1B
496:	1A 19 19 19 19 19 19 19 19 1A 1A 1A 1A 1A 1A 1A
512:	1A 1A 1A 1A 1B 1E 1A 19 1A 1A 1A 1B 1A 19 18 18
528:	18 18 18 18 19 1B 1C 1C 1C 1B 1A 1A 1B 1C 1C 1B
544:	1B 1A 1B 1D 1E 1B 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A
560:	1A 1A 1A 1A 1A 19 19 19 19 19 19 1A 1A 1A 1A 1B
576:	1A 19 1B 1E 1D 1D 1F 20 22 22 1D 1C 1D 1D 1D 1E
592:	20 1E 1C 1C 1F 1F 1F 1E 1D 1A 1B 1C 1C 1C 1C 1B

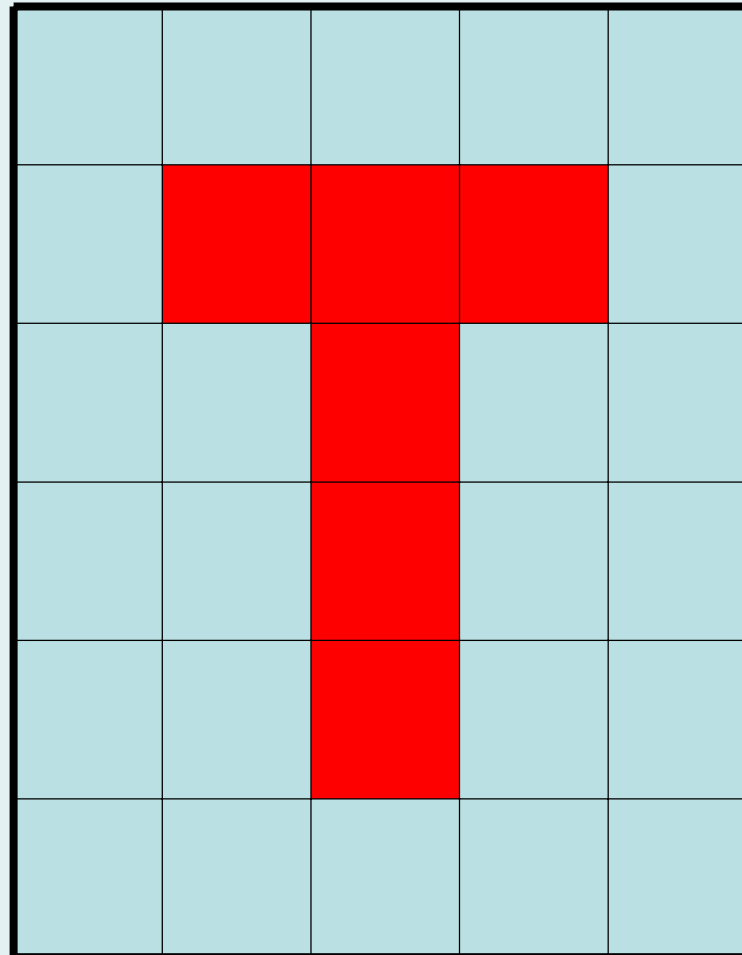
Annotations:

- image width (277) points to the value 01 03 at offset 32.
- image length (339) points to the value 53 01 at offset 48.
- compression (uncompressed) points to the value 08 at offset 64.
- image data points to the start of the image data block at offset 304.

*<basic
information>*

*<rendering
information>*

*<storage
information>*



Categories of digital content in a file format

<basic information>

What to do?

<rendering information>

How to do it?

<storage information>

How to move it from persistent to deployed form?

<data>

What to deploy?



Which categories are significant for keeping the essence of information?

<basic information>

Mandatory

<rendering information>

Useful

<storage information>

Historical

<data>

Mandatory



Shapes of Digital Content

- ❑ On a very basic level (storage level) digital content is just **binary data**
- ❑ On the software level: File
Operating system: Sequence of bytes
Application program: *meaningful* sequence of bytes
→ **File (Data) Format**
- ❑ On the most human-perceivable level it appears in a rendered form
→ **(rendered) Digital Object**

011100110001110100011010...



Shapes of digital content: Digital Object

- ❑ What is a digital object?

OAIS* Definition

“An object composed of a set of bit sequences.”

011100110001110100011010...



OAIS= Open Archival Information System



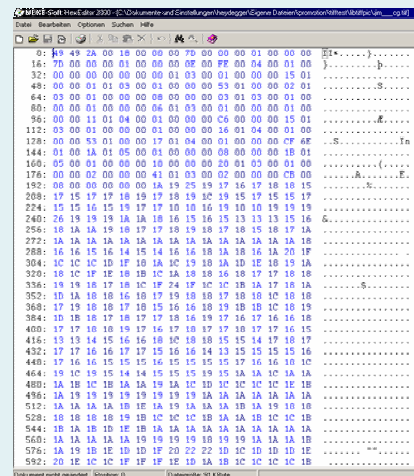
Shapes of digital content: Digital Object

- ❑ What is a digital object?

SB* Definition

„A sequence or stream of bits.“

Example: the content of a file or a network package.“



SB= State and University Library, Denmark



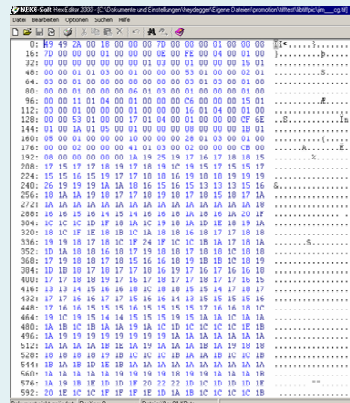
Shapes of digital content: Digital Object

□ What is a digital object?

PREMIS* Definition

„Discrete unit of information in digital form. A Digital Object can be a Representation, File, Bitstream, or Filestream. Note that the PREMIS definition of Digital Object differs from the definition commonly used in the digital library community, which holds a digital object to be a combination of identifier, metadata, and data.“

0111001100011101
00011010...



PREMIS= Preservation Metadata Implementation Strategies

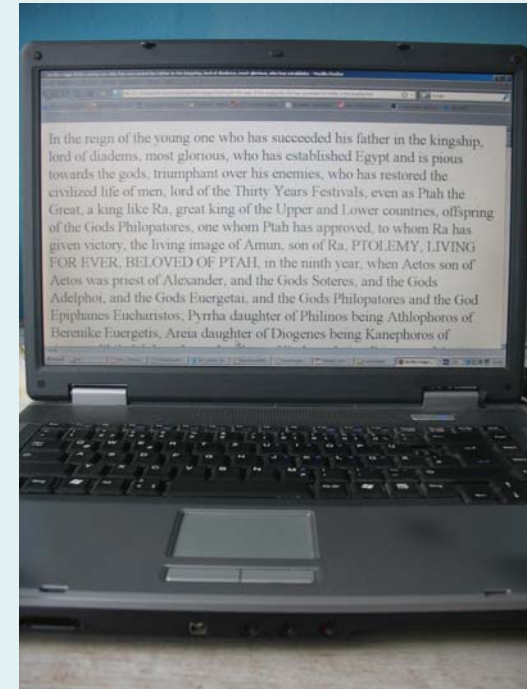


Shapes of digital content: Digital Object

□ What is a digital object?

KB-N / NANETH* Definition

„A digital object consists of a particular combination of hardware and software: the hardware platform, system and application software and one or multiple files. Each file consists of a series of ones and zeroes that is interpreted by a certain combination of hard- and software. The result of that interpretation is a unique representation that is called the digital object. Each digital object can be experienced by its structure, content, context, appearance and behaviour. Example: the representation of a PDF document, interactive multimedia application, database.“



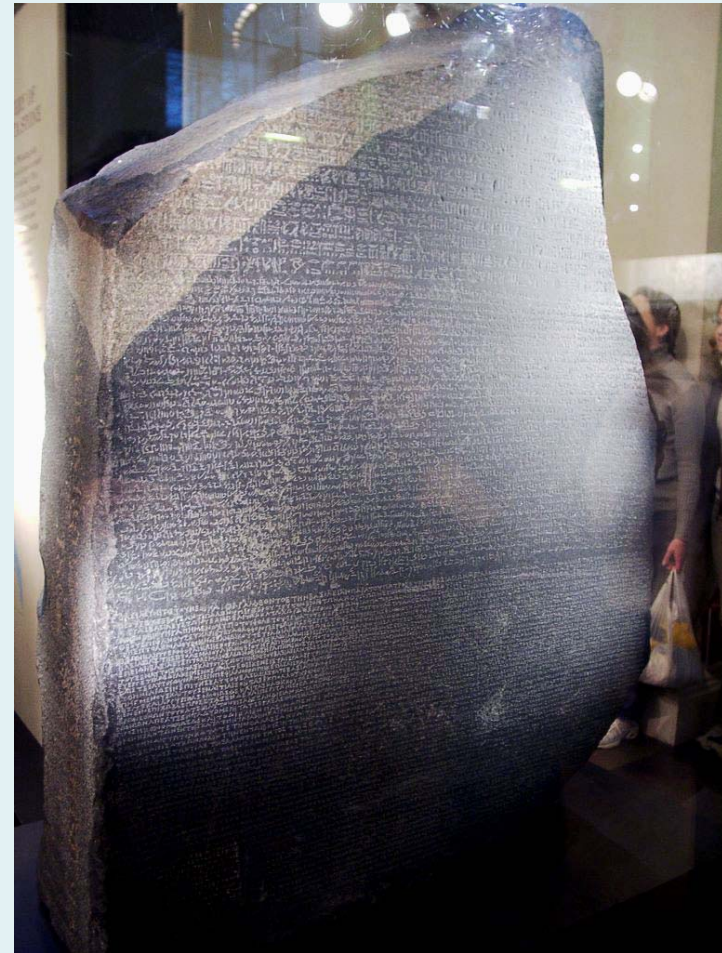
KB-N= Royal Library of the Netherlands,
NANETH= National Archives of the Netherlands



Digital Content: Preservation Issues and Challenges

- ❑ How can we preserve this information?

Traditionally stored information can be maintained by 'passive' preservation

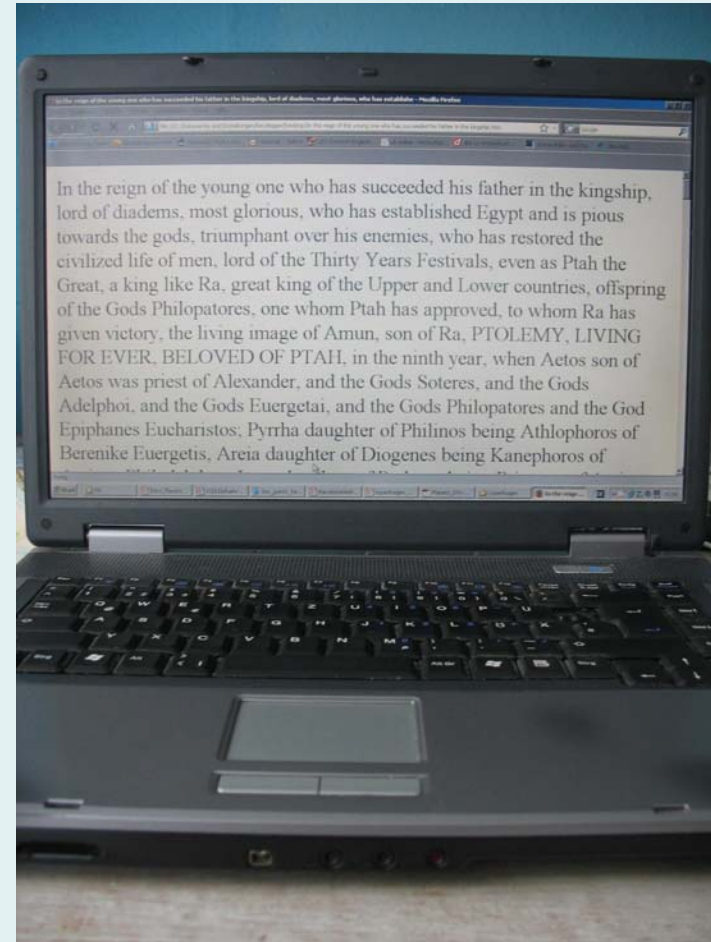


Digital Content: Preservation Issues and Challenges

- ❑ How can we preserve this information?

(As you all know 😊)

Even putting the whole machine in a computer museum would not be enough since...



(Some) risks for digital information

- ❑ Hardware obsolescence
- ❑ Software obsolescence
- ❑ Format obsolescence
- ❑ Context can get lost



How can we cope with it?

- ❑ We need to plan preservation: find out and make decisions on what to preserve and how to do it the best way
- ❑ We need to evaluate and perform concrete actions on digital objects
- ❑ We need to identify objects, extract, evaluate and register their characteristics, and profile collections
- ❑ We need to test - in a controlled environment - how objects behave under certain circumstances
- ❑ We need to combine different preservation tasks in an orchestrated way



How can Planets support these issues?

- ❑ We need to plan preservation: find out and make decisions on what to preserve and how to do it the best way
- ❑ → Planets Preservation Tool (Plato)

- ❑ We need to evaluate and perform concrete actions on digital objects
 - evaluate preservation action tools, provide services for action tools

- ❑ We need to identify objects, extract, evaluate and register their characteristics and profile collections
 - XCL tools, Pronom



How can Planets support these issues?

- ❑ We need to test - in a controlled environment - how objects behave under certain circumstances
→ Testbed
- ❑ We need to combine different preservation tasks in an orchestrated way
→ Planets Interoperability Framework



Thank you for your attention!

