Introduction to Digital Preservation: Why Preserve? How to Preserve?

Dr. Ross King

Austrian Institute of Technology

Outline

- Introduction: Digital Universe
- Digital Preservation Challenges
  - Information Retrieval past and present
  - Bit-stream Preservation
  - Logical Preservation
- Digital Preservation Incentives
  - Markets
  - Incentives
  - Risks
- Conclusions

Introduction

The Digital Universe

CENTRAL EUROPEAN INITIATIVE

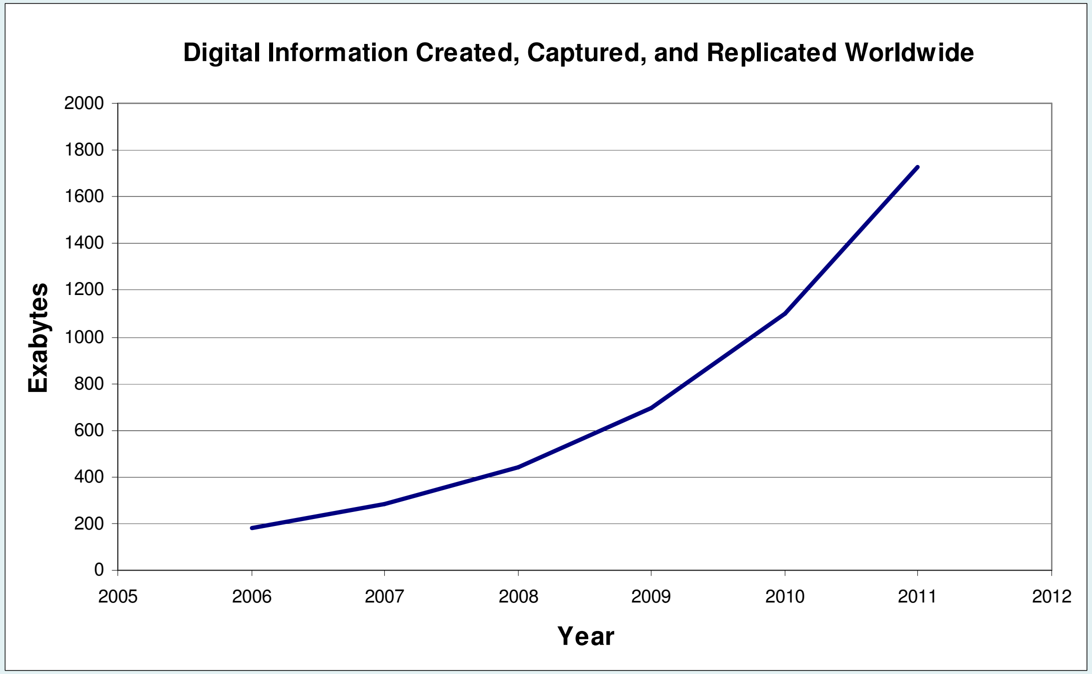
planets

The Digital Universe

- Estimated volume of digital information worldwide in 2007: 281 Exabytes

1000	k	kilo
1000^2	M	mega
1000^3	G	giga
1000^4	T	tera
1000^5	P	peta
1000^6	E	exa
1000^7	Z	zetta
1000^8	Y	yotta

- Estimated growth rate: ca. 60%
- → 700 Exabytes in 2009!



source:
"The Diverse and Exploding Digital Universe"
IDC White Paper, March 2008
http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

CEI
CENTRAL EUROPEAN INITIATIVE

planets

The Digital Universe

- Information creation is beginning to exceed storage capacity
- Much of this information is
  - transient
  - redundant

The Digital Universe

Issues:

- What is worth preserving?
- How to preserve?
- How to preserve so much?
- How to ensure quality?
- How to create incentives to preserve?

CENTRAL EUROPEAN INITIATIVE

planets

The Digital Universe

Issues:

- What is worth preserving?
- How to preserve?
- How to preserve so much?
- How to ensure quality?
- How to create incentives to preserve?

Part 1

Digital Preservation Challenges

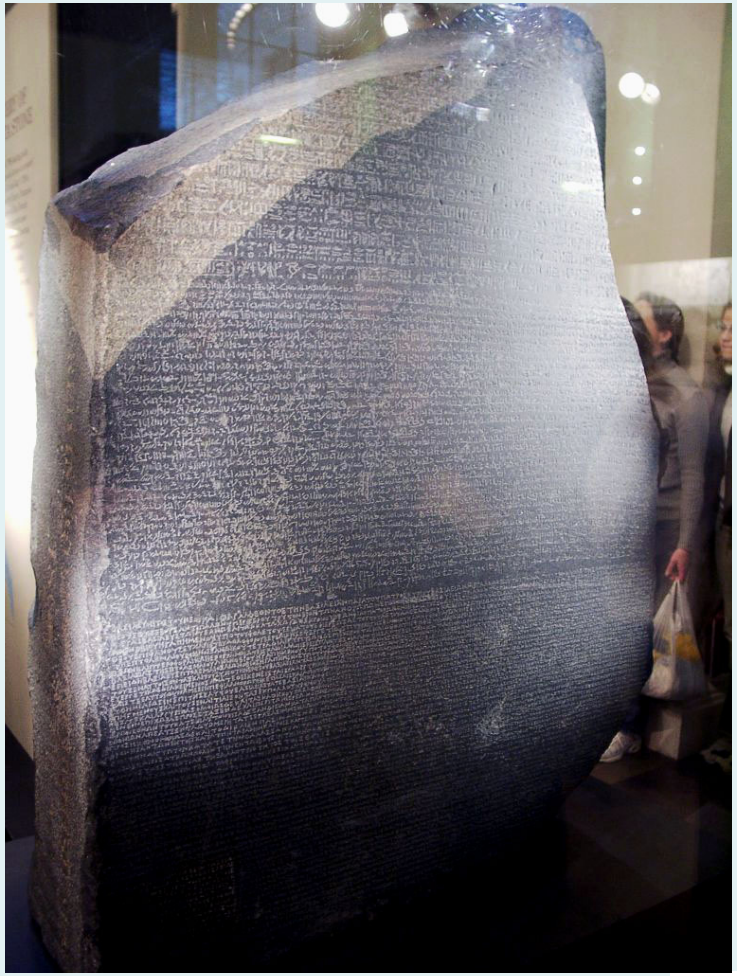
Digital Preservation

- standards, best-practices, and technologies utilized in order to ensure access to digital information over time

"Digital documents last forever – or five years, whichever comes first."

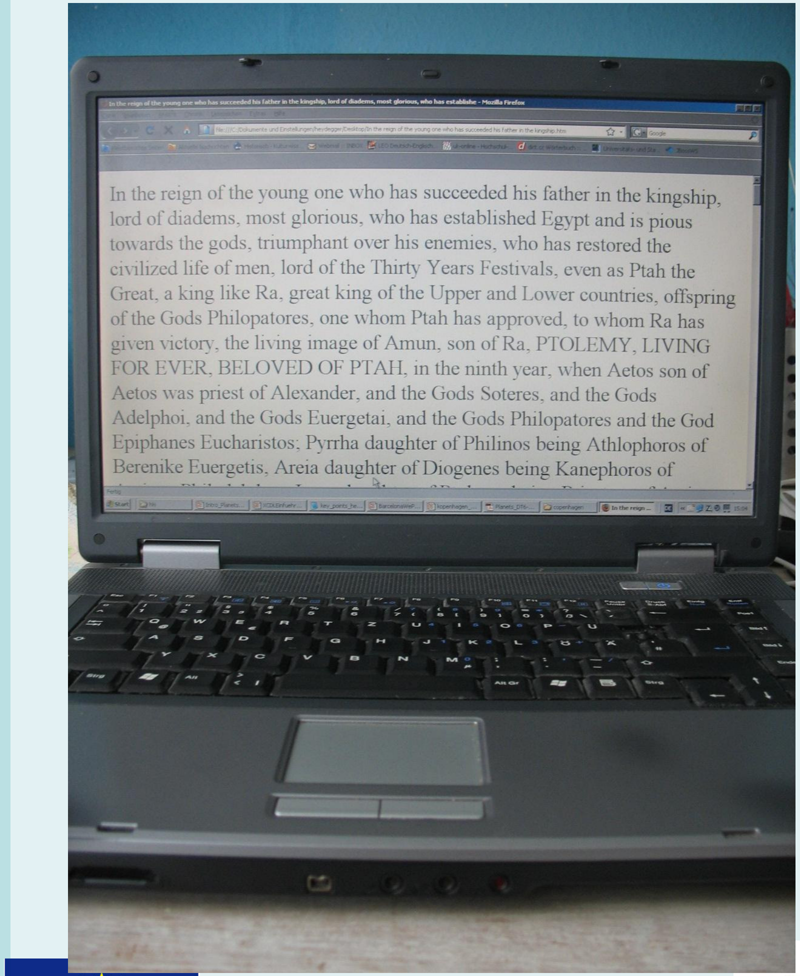
– http://www.clir.org/pubs/reports/rothenberg/introduction.html

Information Retrieval – 196 BC



- 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

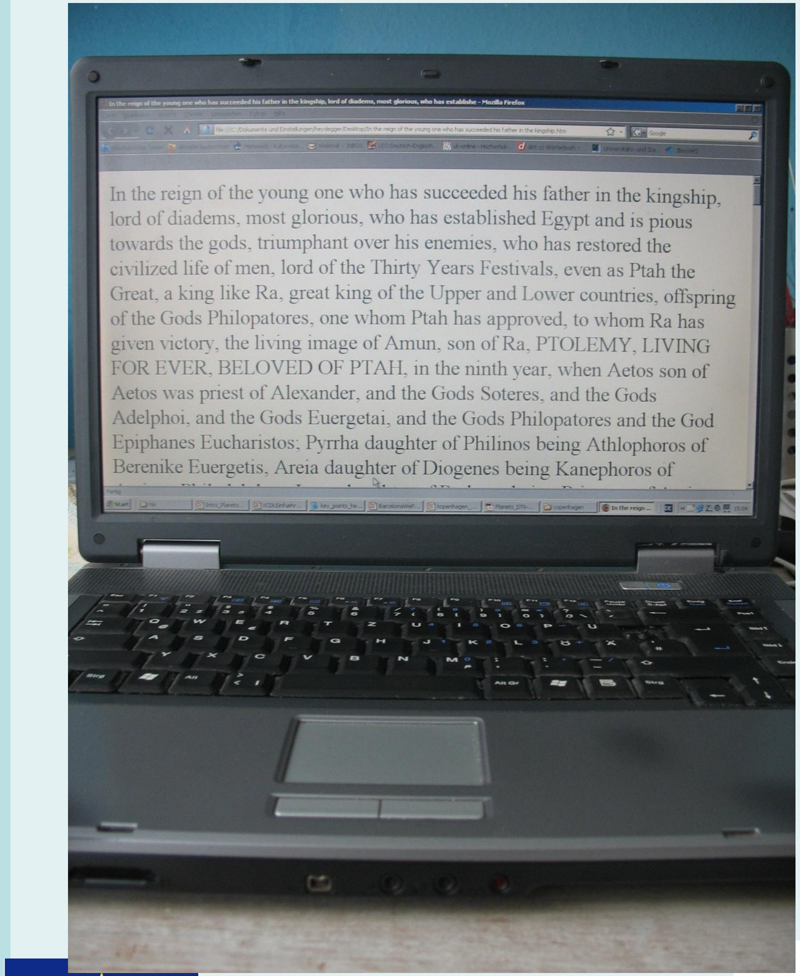
Information Retrieval – 2009 AD


In the reign of the young one who has succeeded his father in the kingship, lord of diadems, most glorious, who has establishe - Mozilla Firefox
file:///C:/Dokumente und Einstellungen/heydegger/Desktop/In the reign of the young one who has succeeded his father in the kingship.htm
Google
Start
In the reign ...
15:04

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 FOR 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

- Hardware
  - Storage medium (hard disk, optical disc, …)
  - Rendering environment (display, printer, …)
- Software
  - Low level software (operating system)
  - Application software (webbrowser, texteditor, …)

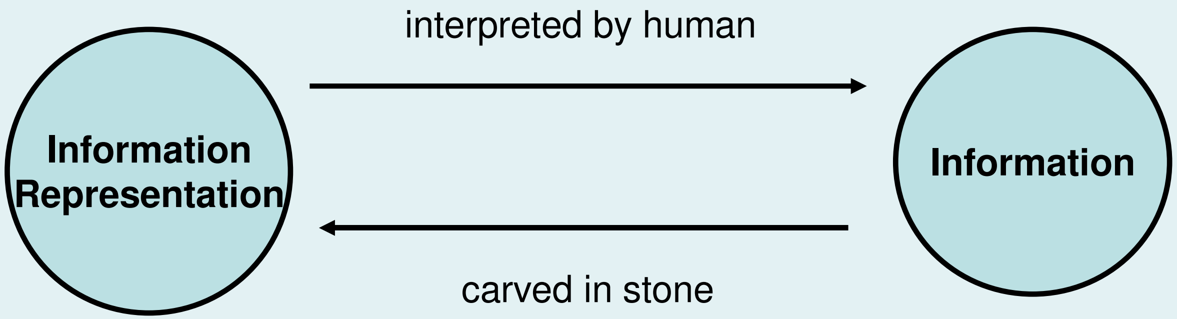
Information Retrieval – 2009 AD


In the reign of the young one who has succeeded his father in the kingship, lord of diadems, most glorious, who has establishe - Mozilla Firefox
file:///C:/Dokumente und Einstellungen/heydegger/Desktop/In the reign of the young one who has succeeded his father in the kingship.htm
Google
Start
In the reign ...
15:04

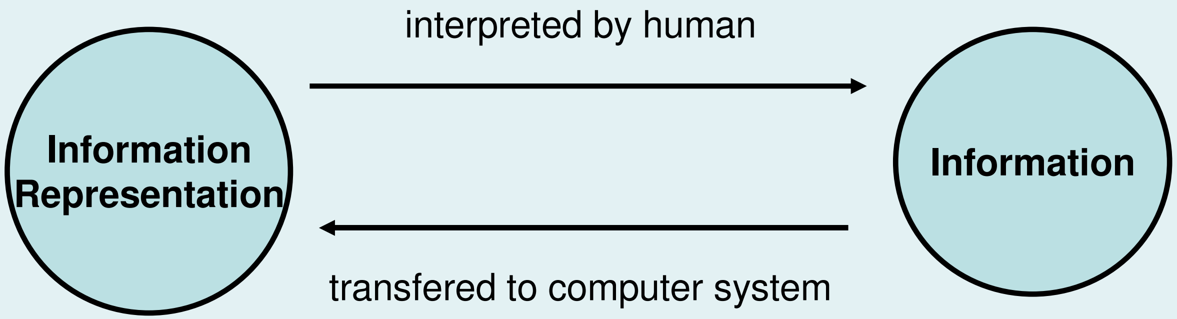
In the reign of the young one who has succeeded his father in the kingship, lord of diadems, most glorious, who has established Egypt and is pious towards the gods, triumphant over his enemies, who has restored the civilized life of men, lord of the Thirty Years Festivals, even as Ptah the Great, a king 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 FOR 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

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

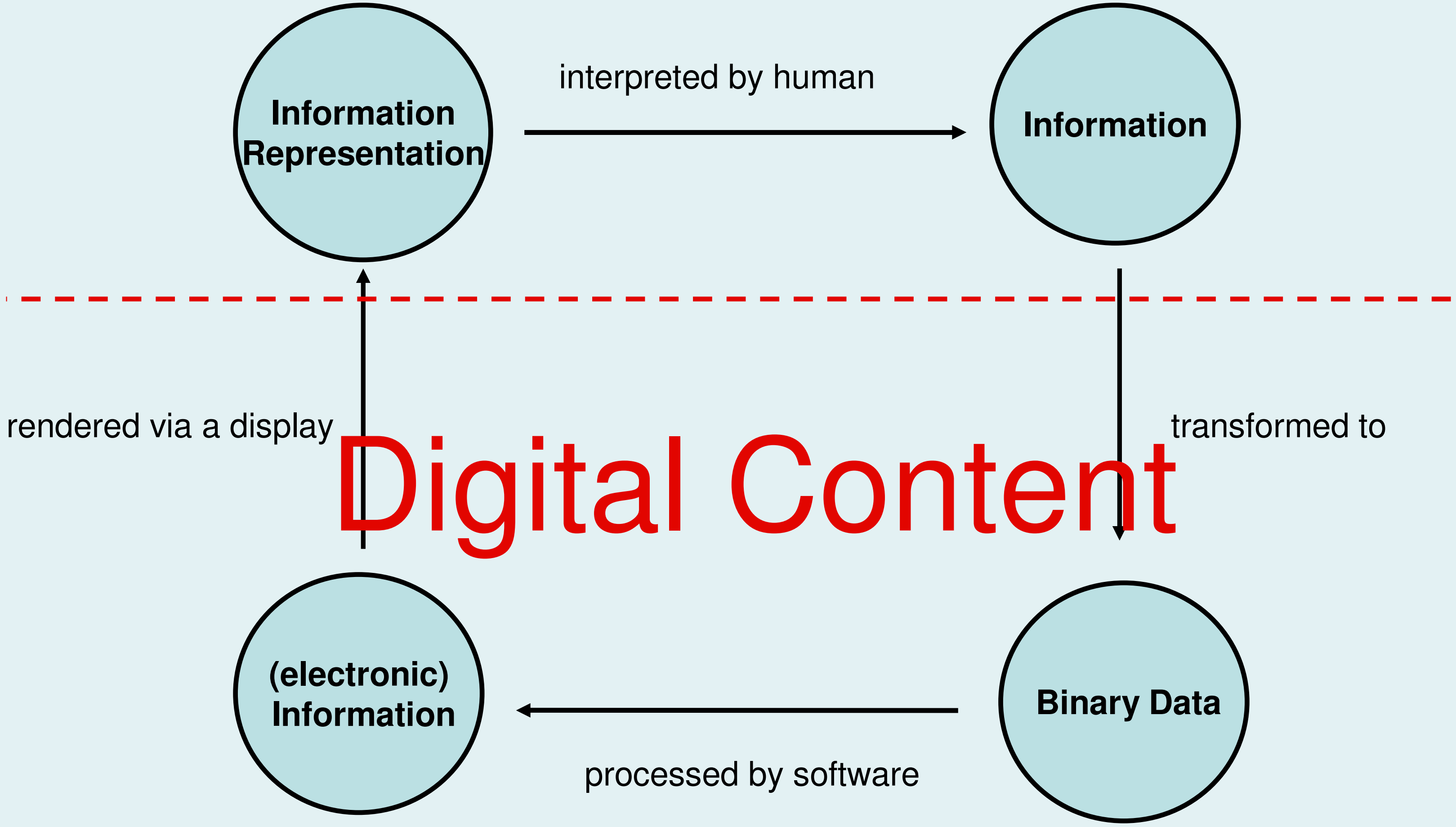
Information cycle – 196 BC



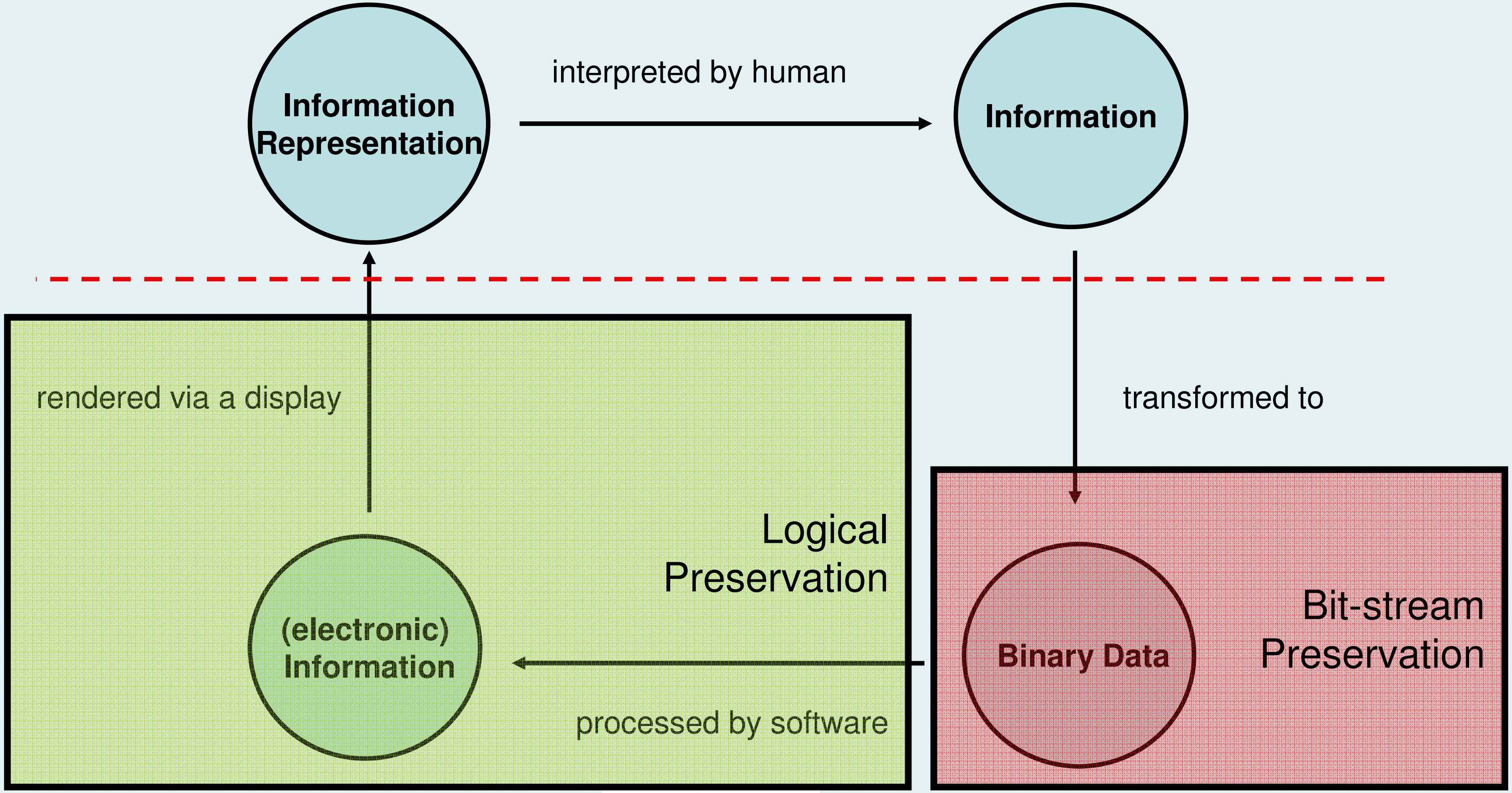
Information cycle – 2009 AD (simplified)



Information cycle – 2009 AD



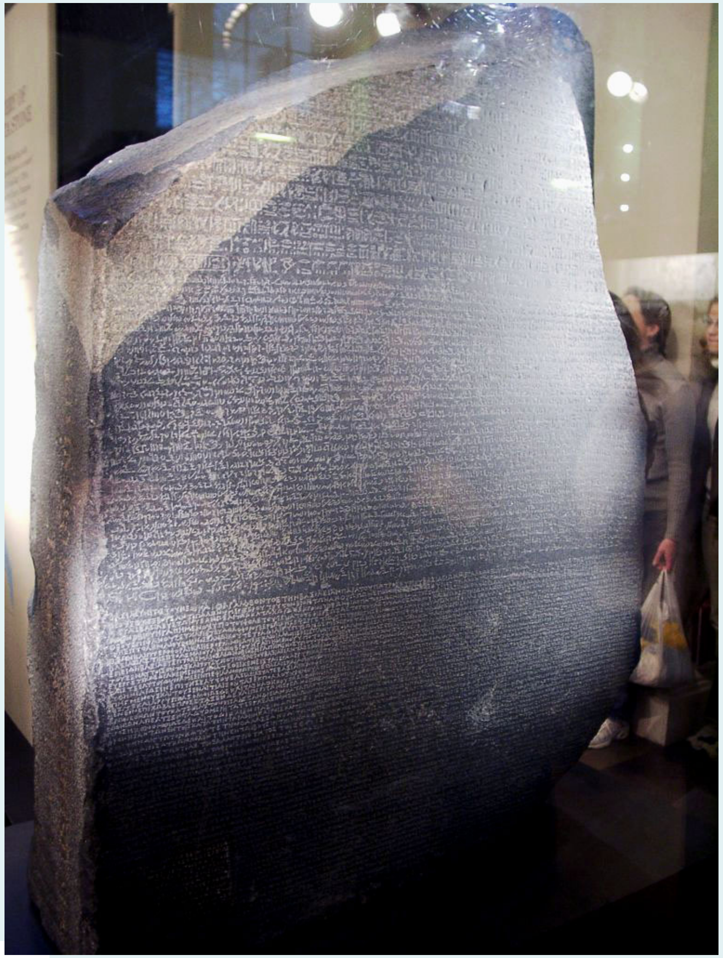
Information cycle – 2009 AD



Digital Content: Preservation Issues and Challenges

- How can we preserve this information?

  Traditionally stored information can be maintained by “passive preservation”

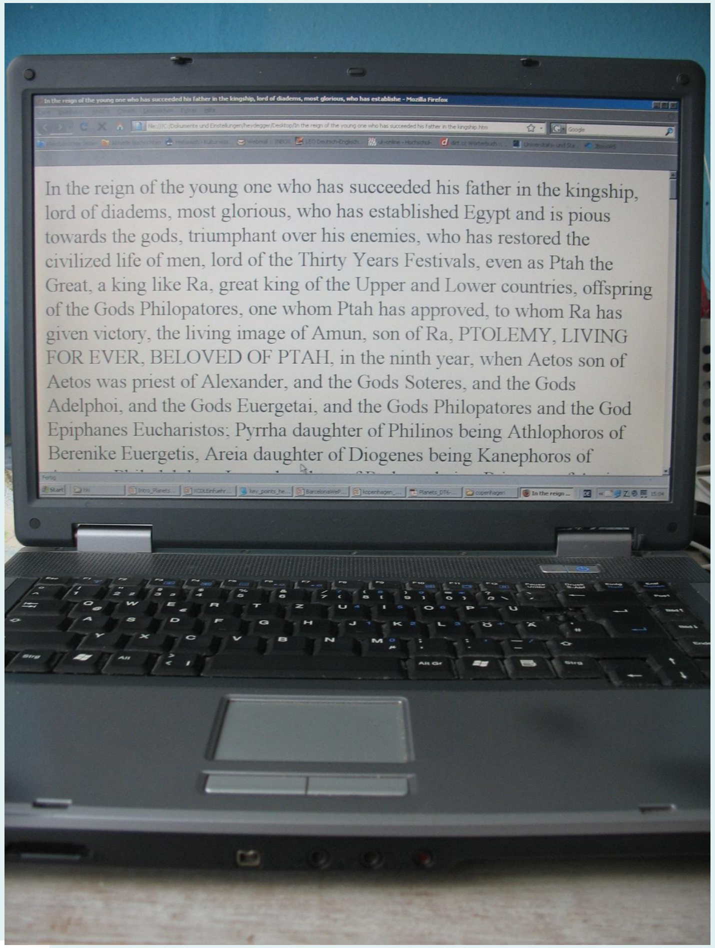


CENTRAL EUROPEAN INITIATIVE

Digital Content: Preservation Issues and Challenges

- How can we preserve this information?

  Even putting the whole machine in a computer museum would not be enough due to…


In the reign of the young one who has succeeded his father in the kingship, lord of diadems, most glorious, who has establishe - Mozilla Firefox

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 FOR 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

(Some) risks for digital information

- Media obsolescence
- Hardware obsolescence

Bit-stream Preservation

- Software obsolescence
- Format obsolescence
- Loss of context

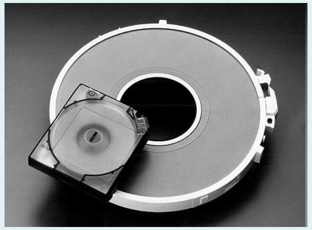
Logical Preservation

Bit-stream Preservation

- Problem: digital media do not last forever
  - media deterioration
- Problem: hardware for accessing digital media does not last forever
  - hardware obsolescence

Bit-stream Preservation: Lifespan of Media

- Parchment: 1000 years
- Microfilm: 500 years
- Paper: 50 – 200 years
  - high levels of acid can cause paper to disintegrate

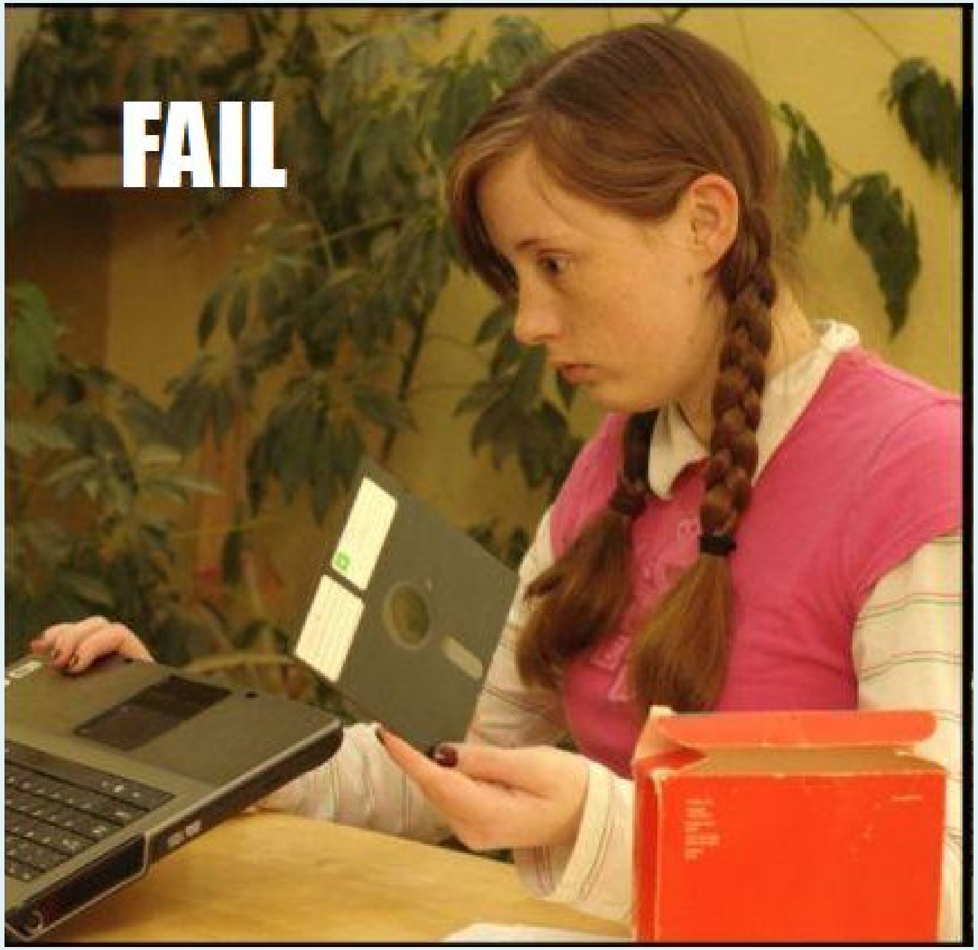


- Magnetic Tape: 100 years
  - the binder that holds magnetic particles to the tape can decompose and cause the layers of tape to stick together in a reel
- CD-ROM: 10 years
  - poor manufacturing processes allow the reflective aluminum layer to oxidize

Is this progress?

Bit-stream Preservation: Lifespan of Hardware

Bit-stream Preservation: Lifespan of Hardware



Bit-stream Preservation

- Solution: migration
  - regular copying to new media
    - requires automation to handle volume and reduce expense

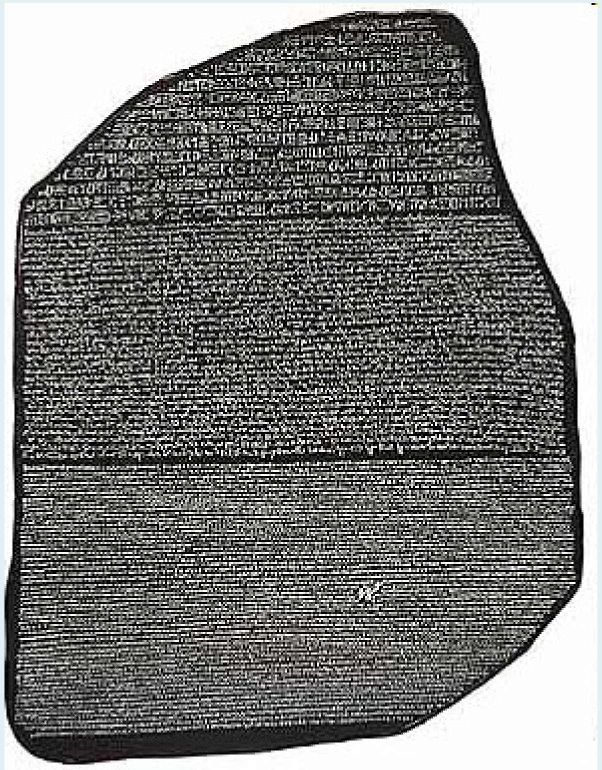
- Solution: hardware museums
  - expensive
  - depends on other obsolete hardware

Logical Preservation

- Problem: data formats are not supported forever
  - format obsolescence
- Problem: how to interpret preserved bit-streams

Logical Preservation

- Solution: Representation Information



From the DCC Digital Curation Manual: “Representation Information (RI) refers to all information required to access the information stored within a digital object. The term can be applied to all levels of abstraction and refers to both the structural and semantic composition. The use of RI is often recursive: using one element of RI in a meaningful manner requires further RI. This recursion continues until the contents of the original object are displayed in a form the user can understand.”

Logical Preservation

- Solution: Emulation
  - – use representation information to re-create the environment necessary to access the preserved bit-stream
- Solution: Migration
  - – use representation information to identify endangered bit-stream formats and convert them to accessible (open, standardized) formats

How can Planets help?

- We must plan preservation: find out and make decisions on what to preserve and how to do it the best way

  → Planets Preservation Tool (Plato)

- We must evaluate and perform concrete actions on digital objects

  → evaluate preservation action tools, provide services for action tools

- We must identify objects, extract, evaluate and register their characteristics and profile collections

  → XCL tools, Pronom

CEI
CENTRAL EUROPEAN INITIATIVE

planets

How can Planets help?

- We must test - in a controlled environment - how objects behave under certain circumstances

  → Testbed

- We must be able to combine different preservation tasks in an orchestrated way

  → Planets Interoperability Framework

Part 2

Digital Preservation Incentives

What is the market for Digital Preservation?

- Memory Institutions
- Governments
- Software Manufacturers
- ... and everybody else!
  - Companies and Individuals
  - All kinds of digital content
    - Documents
    - Photographs
    - Audio/Video
    - Databases
    - Emails
    - Spreadsheets
    - Websites
    - CAD
    - Simulations
    - …

CEI
CENTRAL EUROPEAN INITIATIVE

planets

Reasons to implement Digital Preservation

- compliance with legislation, for example on freedom of information, Sarbanes-Oxley, environmental information – and, of course, legal deposit
- providing the long-term guarantees of access to digital content needed to sustain the transition from paper to digital information societies and business processes
- where enforced by regulatory organisations, for example the European Medicines Agency and the US Food and Drug Administration in the case of pharmaceutical companies
- protecting the interests of the organisation and the rights of all present and future stakeholders
- providing evidence of IPR or patent rights
- providing evidence of good practice to defend against litigation
- protecting business critical information or allow data mining and analysis
- providing business continuity in the event of catastrophic data loss
- maintaining information of historical or scientific value
- maintaining life-long medical information
- maintaining information of personal value, such as e-mails, music and photographs
- ...

CEI
CENTRAL EUROPEAN INITIATIVE

planets

What is stopping us?

- Business decisions are made based on the short-term, whereas preservation is a (relatively) long-term problem.
- Business decisions are made based on return on investment. How to calculate return on investment?

Perhaps preservation should not be about “return on investment”, but rather about **risk management**.

What is the financial risk?

In order to answer that question, we must ask:

- how many digital objects are produced?
- what are these objects worth?
- how long do digital objects retain their value?
- how many objects are in danger of digital obsolescence?

and then

- what does it cost to preserve?

If we can estimate the financial risk, we can justify the preventative investment in digital preservation...

[1] M.K. Bergman, “Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents,” BrightPlanet Corporation White Paper, July 2005, 42 pp
[2] P. Lyman, and H. Varian, "How Much Information", Technical Report 2003
[3] LIFE^1 and LIFE^2 projects: http://www.life.ac.uk/

CEI
CENTRAL EUROPEAN INITIATIVE

planets

Conclusions

- The volume of digital information being produced is staggering
- There are multiple challenges, some solutions, many open questions
- Planets can offer solutions for some aspect of the digital preservation challenge
- There are many incentives for digital preservation, but the long-term nature of the problem is a hindrance
- A risk management approach might serve to involve industry stakeholders and decision-makers

CENTRAL EUROPEAN INITIATIVE

Thank you for your attention!

Contact information:

Dr. Ross King
AIT Austrian Insitute of Technology GmbH
ross.king@ait.ac.at