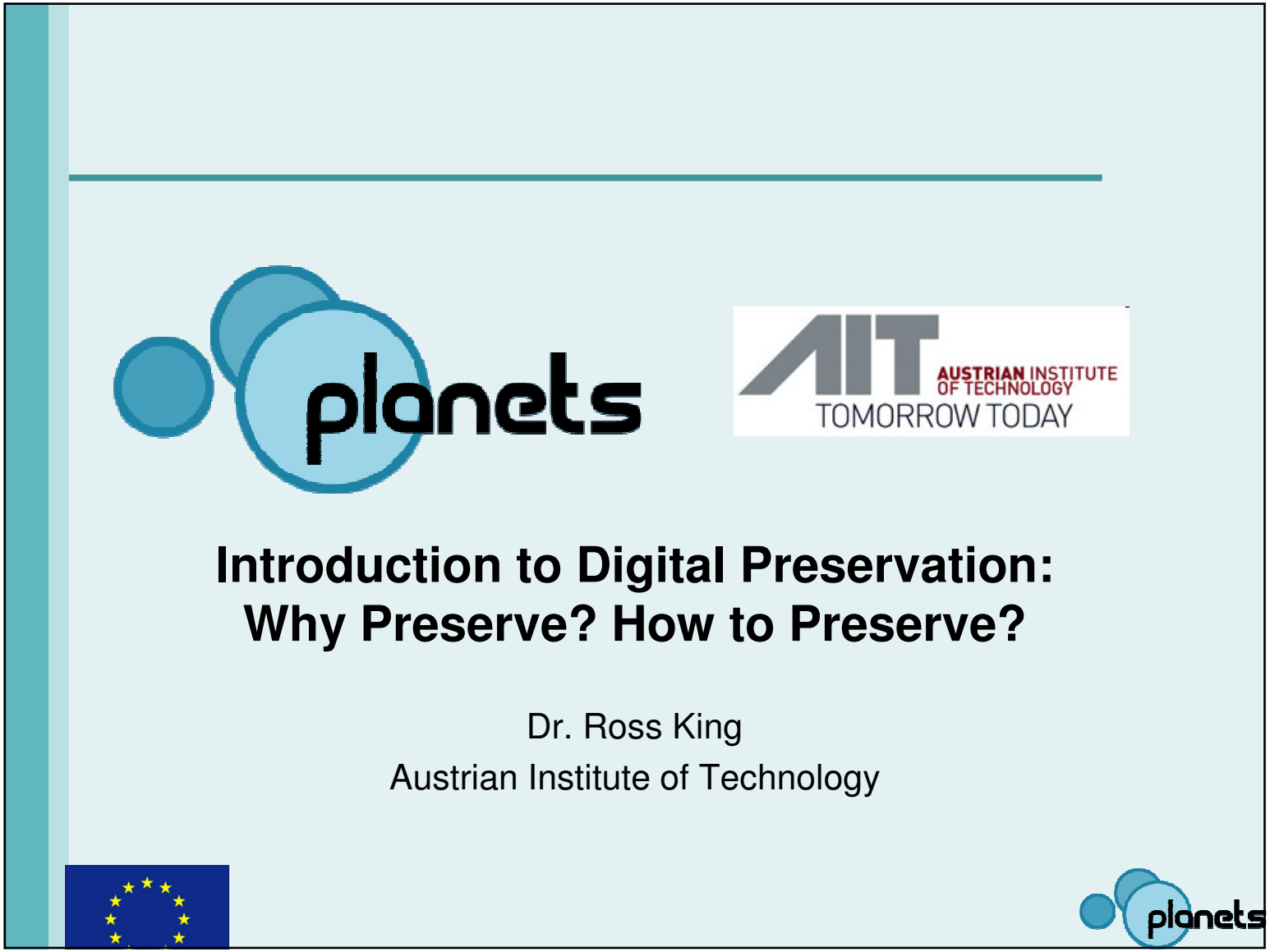

planets
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY
Introduction to Digital Preservation:
Why Preserve? How to Preserve?
Dr. Ross King
Austrian Institute of Technology
planets

Outline

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

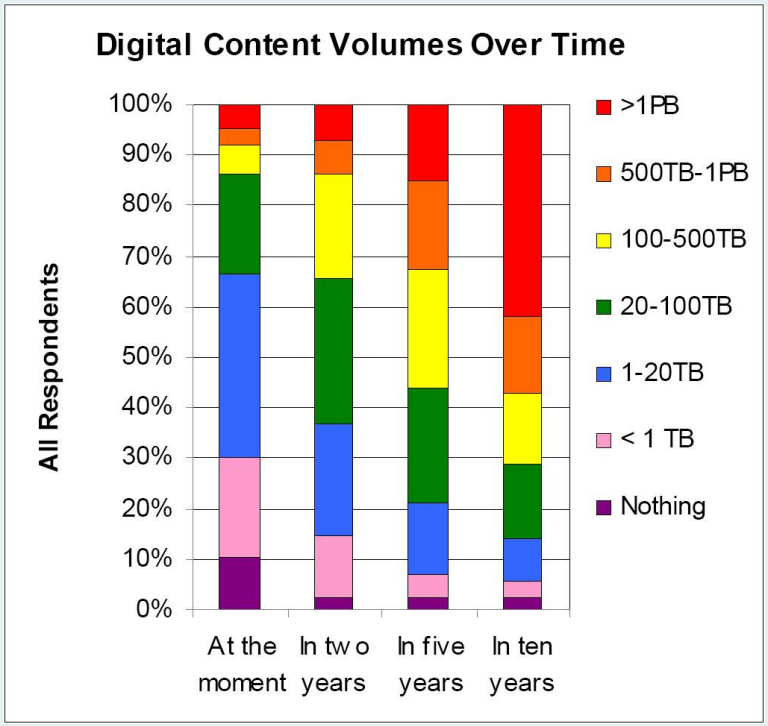
Introduction

The Digital Universe

The Digital Universe

- Between now and 2019, the volume of content that organisations need to hold will rise twenty-five-fold, from a median of less than 20TB now to over 500TB.

source:
"Are You Ready? Assessing European Organisations' Preparations forDigital Preservation"
Planets Deliverable D7b, November 2009

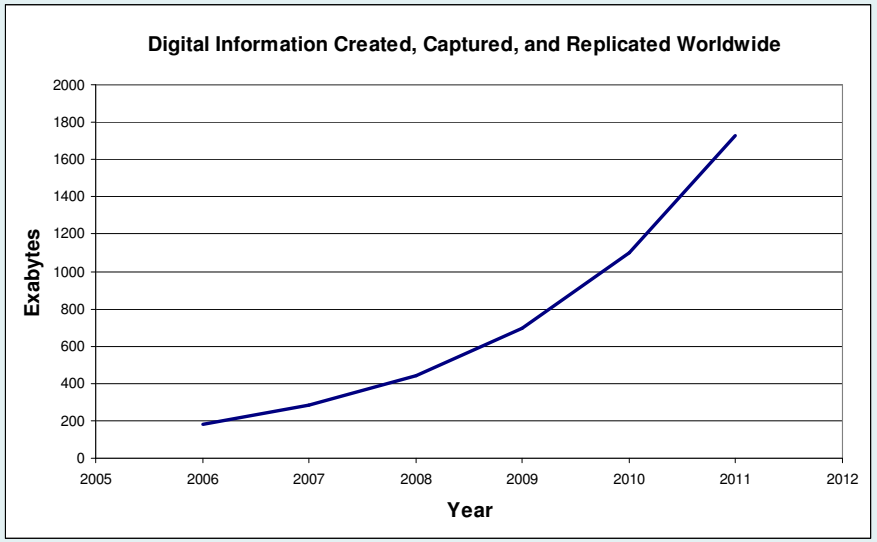


The Digital Universe

- 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

The Digital Universe

- In other words, in 2009, humanity produced approximately 100 GBytes of digital information for every person on Earth.
  - It would require one trillion (1.000.000.000) CD-ROMs to store all of this information (700 Exabytes)
  - This stack of CD-ROMs (in their jewel cases) would be ten million kilometers high
  - If we switch to DVDs, the stack would still be one-and-a-half a million kilometers high, over three times the distance between the Earth and the Moon
- For the first time, information creation is beginning to exceed storage capacity, although 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?

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

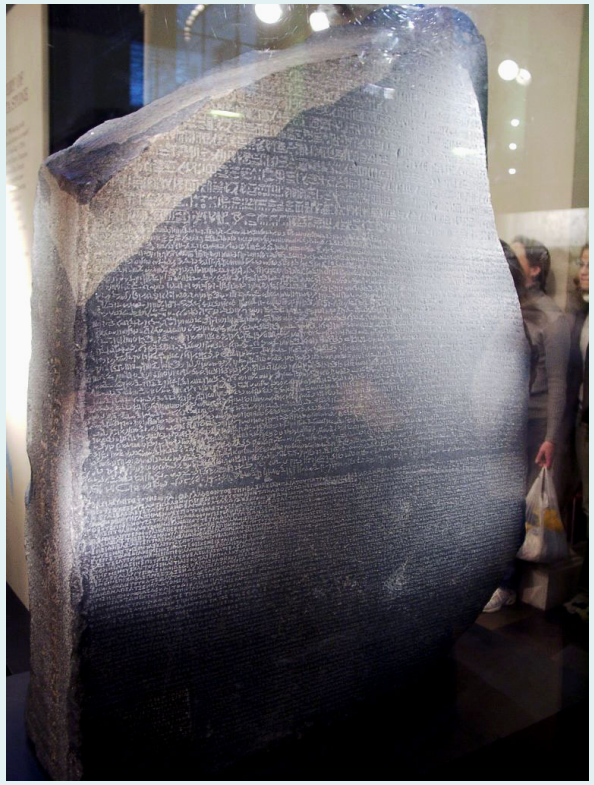
Digital Preservation

- 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."

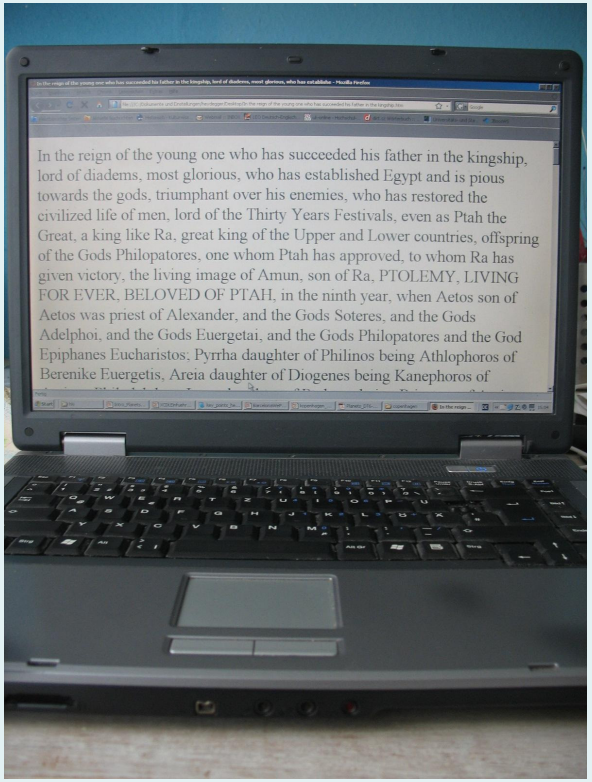
- http://www.clir.org/pubs/reports/rothenberg/introduction.html

Information Retrieval – 196 BC



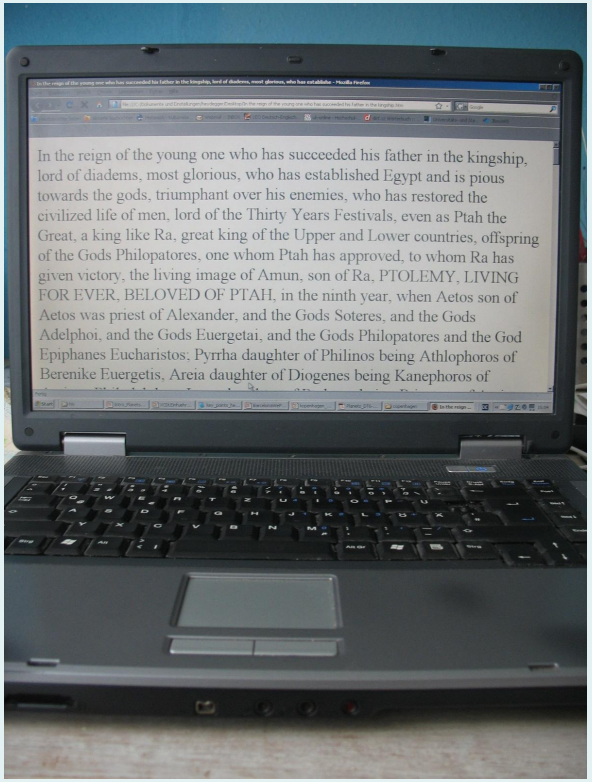
- Carrier
  - Solid material (granodiorite)
  - 114 x 72 x 28
  - 760 kg
- Encoding
  - Human-readable characters
  - Three language scripts (hieroglyphic, demotic, ancient greek)
- How to get the information?
  - Human, capable of reading (at least) one of the scripts

Information Retrieval – 2009 AD



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

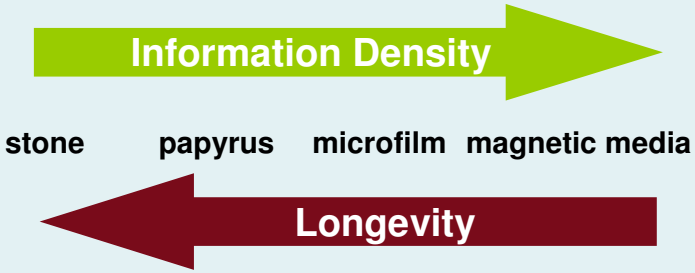
Information Retrieval – 2009 AD



- 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

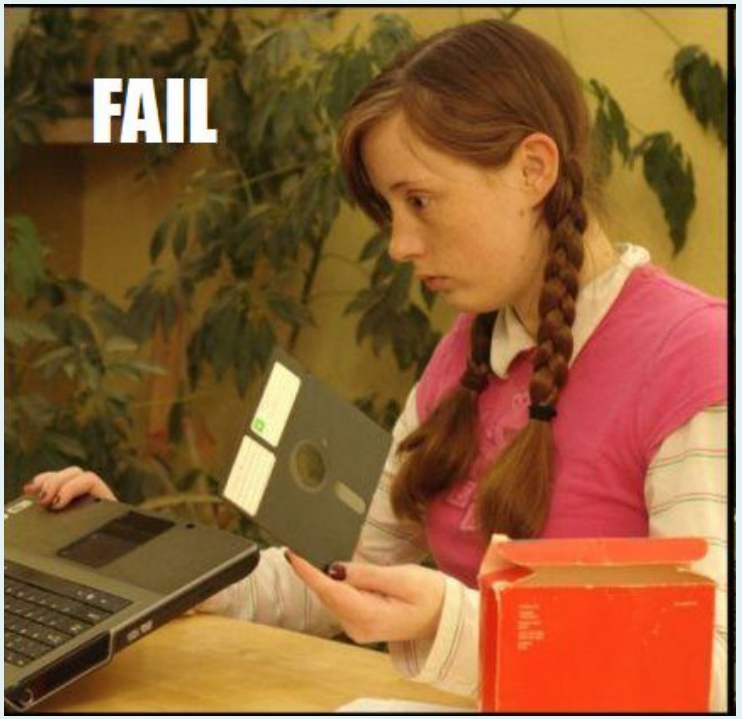
Digital Preservation Challenges

- First challenge: preserving the bits
  - also known as *bit-stream preservation*
- this challenge relates to the storage medium itself, which is subject to decay over time
  - also known as *media obsolescence*
- over history we have generally traded information density and reproducibility against media longevity



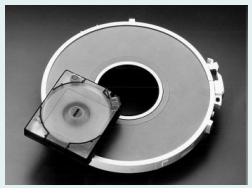
Digital Preservation Challenges

- First challenge:
  preserving the bits
  - also known as
    bit-stream preservation
- presently solved by continuous media migration
- this must be accompanied by concurrent hardware migration



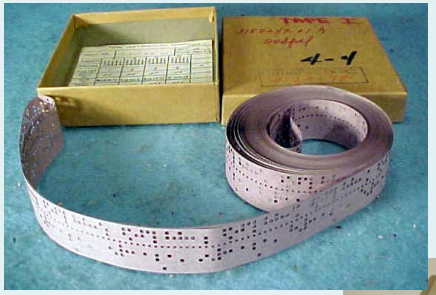
Media Obsolescence

- 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



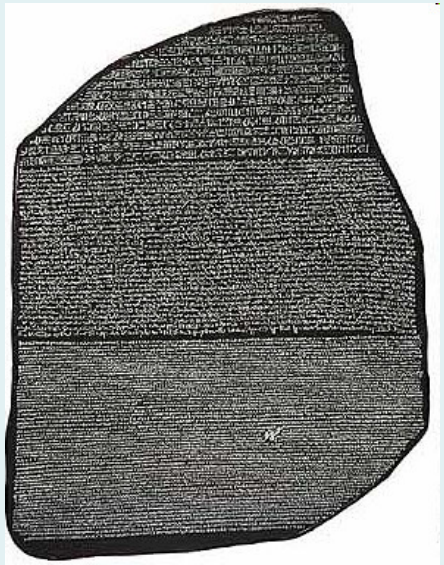
Is this progress?

Hardware Obsolescence



Digital Preservation Challenges

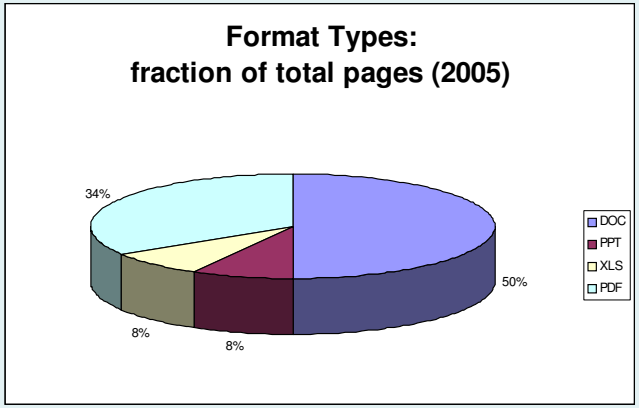
- Second challenge: understanding the bits over time
  - also known as *logical preservation*
- this challenge relates to the contextual information about the bit-stream, which can be lost over time
  - also known as *format obsolescence*
- format *migration* or context *emulation* are two approaches to this challenge



Format Obsolescence

- Typical knowledge workers produce at least two-thirds of their documents in proprietary formats.
- Such formats have high preservation risks related to
  - limited support over time
  - limited backwards compatibility
  - dependency on third parties

- How can one be sure what file formats really exist in institutional repositories? For example, here are all the different formats with the .DOC extension:



source: "Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents," Michael K. Bergman, 2005

Format Name	Format Version	Pronom ID
Interleaf Document		x-fmt/329
Microsoft Word for Macintosh Document	6.0	x-fmt/2
Microsoft Word for Macintosh Document	X	x-fmt/129
Microsoft Word for MS-DOS Document	3.0	x-fmt/273
Microsoft Word for MS-DOS Document	4.0	x-fmt/274
Microsoft Word for MS-DOS Document	5.0	x-fmt/275
Microsoft Word for MS-DOS Document	5.5	x-fmt/276
Microsoft Word for Windows Document	1.0	fmt/37
Microsoft Word for Windows Document	2.0	fmt/38
Microsoft Word for Windows Document	6.0/95	fmt/39
Microsoft Word for Windows Document	97-2003	fmt/40
Stationary for Mac OS X		x-fmt/131
Wordperfect Secondary File	5.0	x-fmt/42
Wordperfect Secondary File	5.1/5.2	x-fmt/43
WordPerfect for MS-DOS/Windows Document	6.0	x-fmt/44

source: http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx

Digital Preservation Challenges

- Media obsolescence
- Hardware obsolescence

Bit-stream Preservation

- Software obsolescence
- Format obsolescence
- Loss of context

Logical Preservation

How can Planets help?

- Logical Preservation
  - Solution: Emulation
    - use representation information to re-create the environment necessary to access the preserved bit-stream
    - Planets GRATE tool
  - Solution: Migration
    - use representation information to identify endangered bit-stream formats and convert them to accessible (open, standardized) formats
    - Planets Preservation Services Suite

How can Planets help?

- We must profile collections and identify risks
  → Pronom Technical Registry
- We must plan preservation to mitigate risks, which means making decisions regarding what to preserve and how to preserve it
  → Planets Preservation Tool (Plato)
- We must perform concrete preservation actions on digital objects
  → Planets Preservation Service Suite
  → Planets Service Developers Guidelines

How can Planets help?

- We must quantitatively measure how preservation services act on digital objects - in a controlled environment

  → Planets Testbed

  – We must characterise objects and control quality after migration
    → XCL Tool Suite

- We must be able to combine different preservation services in an orchestrated way in order to carry out different preservation workflows

  → 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
    - …

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
- …

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/

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

Thank you for your attention!

Contact information:

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