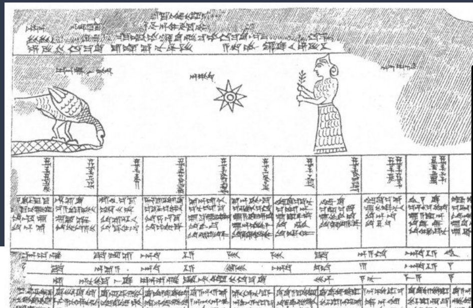
Digital Preservation

By Jean-Yves Le Meur

project leader of CERN Digital Memory

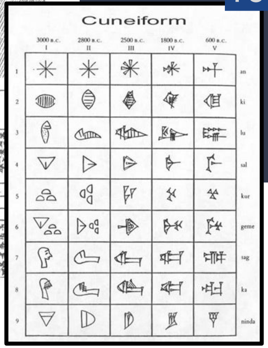
The quest for Earth's oldest ice p. 436
Origin of lung disease in cystic fibrosis p. 503
Dispensing with genes on the Y chromosome p. 524
Science
$15
29 JANUARY 2016
sciencemag.org
AAAS
Jupiter rising
Geometric techniques in Babylonian astronomy
pp. 435 & 482

29 Jan 2016



350-50 BC

Jupiter orbit 👍



100 AD

End of cuneiform on tablets 👎

1300-1400

Jupiter orbit (again) 👍

1800-

Tablets -> British Museum

2014

The missing 'rosetta' tablet

Digital preservation in a nutshell

- World wide Landscape
  - Rationale
  - Interesting initiatives
  - Good practices: OAIS
- The different Approaches

The Digital "Dark Age"

"*We are nonchalantly throwing all of our data into what could become an information black hole without realizing it*"

Vint Cerf (vice-president of Google in Feb 2015)

The Digital “Dark Age”

- Very large **community** worrying about the preservation of digital content
- Digital Preservation Coalition
- Open Preservation Foundation
- UNESCO PERSIST project, EU e-ARK project, National Libraries and Archives
- Many related **conferences**: iPRES series, etc.

“This is not about preserving bits, It is about preserving meaning, much like the Rosetta Stone.”

More than 70 major libraries destroyed over time: accidents, disasters, ethnocides

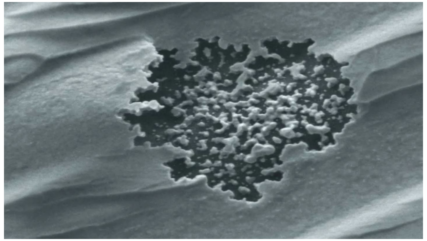
How digital data evaporates (I)

Ten

Major

New

Risks



1. Physical Obsolescence: **Bit rot**
2. **Redundancy** failure
3. **Technological** Obsolescence of readers, formats, OS, HWs
4. Lost in **migrations** !
5. Missing **context**: no codec !

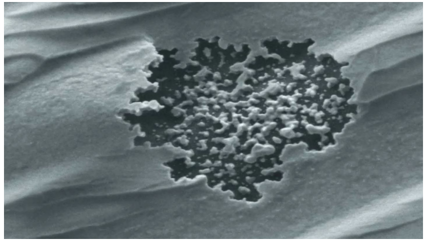
How digital data evaporates (II)

Ten

Major

New

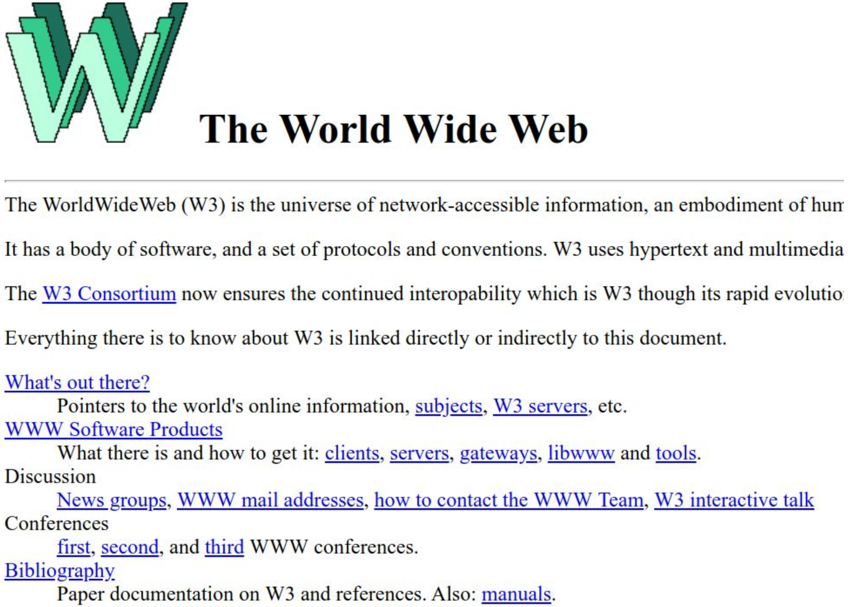
Risks



6. Redundancy failure
7. **Economical Failures**
8. Lost in transitions: people !
9. Corruption, mistake or attack
10. **Dissipation**: out of reach

Some examples at CERN

- The very first WWW pages
  - Reconstructed in 2013 - found again in 2018!
- Important emails
- Business Agreements
- A few scientific Datasets



The World Wide Web

The WorldWideWeb (W3) is the universe of network-accessible information, an embodiment of hum

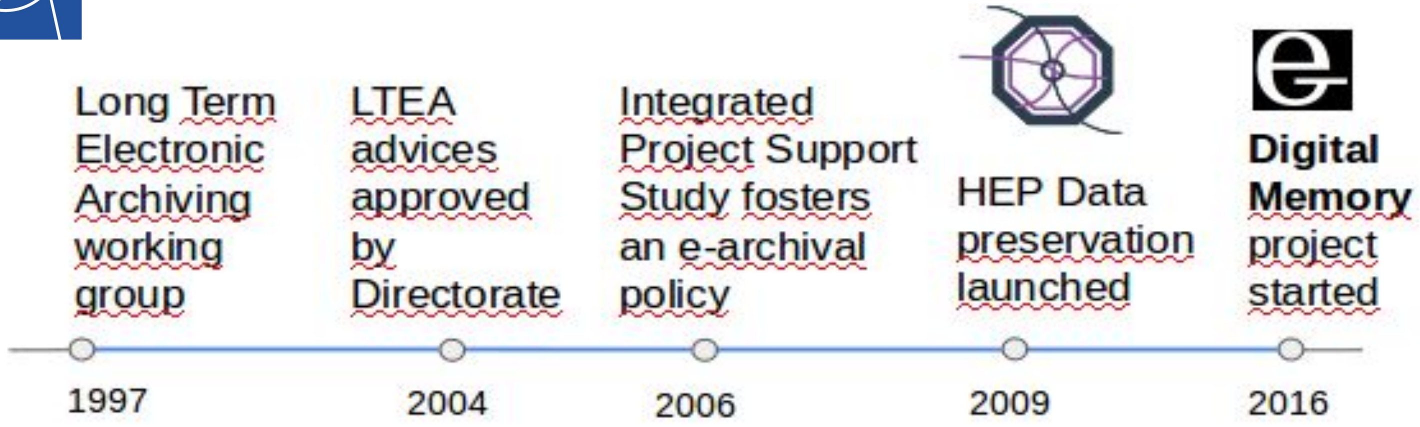
It has a body of software, and a set of protocols and conventions. W3 uses hypertext and multimedia

The W3 Consortium now ensures the continued interopability which is W3 though its rapid evolutio

Everything there is to know about W3 is linked directly or indirectly to this document.

What's out there?
  Pointers to the world's online information, subjects, W3 servers, etc.
WWW Software Products
  What there is and how to get it: clients, servers, gateways, libwww and tools.
Discussion
  News groups, WWW mail addresses, how to contact the WWW Team, W3 interactive talk
Conferences
  first, second, and third WWW conferences.
Bibliography
  Paper documentation on W3 and references. Also: manuals.

Initiatives to prevent loss of data at CERN



1997 — Long Term Electronic Archiving working group

2004 — LTEA advices approved by Directorate

2006 — Integrated Project Support Study fosters an e-archival policy

2009 — HEP Data preservation launched

2016 — **Digital Memory** project started

Interesting world wide initiatives

"Ignoring the problems raised by preserving information in digital forms would lead inevitably to the loss of this information."

Space Data System Recommended Practice

Interesting world wide initiatives

Policies review: more than 50 in 2015 listed by OPF

OAIS: The Open Archival Information System

- *The Coordinated Archive at NASA* (NSSDCA) http://nssdc.gsfc.nasa.gov/
- *The Digital curation at ETH Zurich* (Ex-Libris)
- *Cornell University Library* eCommons platform (DSpace)
- *German National Libraries* (TIB, ZB MED and ZBW) Goportis solution
- *Indian Institute of Geomagnetism (IIG)* preservation framework on top of the existing IR

The Open Archival Information System (OAIS)

- Strict and powerful **reference model:** Trustworthy Digital repo **ISO 16363**
- Information Producers, Consumers, Managers & Designated Communities
- Protection against contingencies: Organizational, Infrastructure, Digital Object
- Existing **software**, e.g.: Preservica, Rosetta, Archivematica, eARK, etc.
    - Conversions to Master formats
    - Fixity check
    - Workflow support

OAIS Good practices

"Preservation is a journey, not a destination."

Digital Preservation: Issues, Concepts & Tools

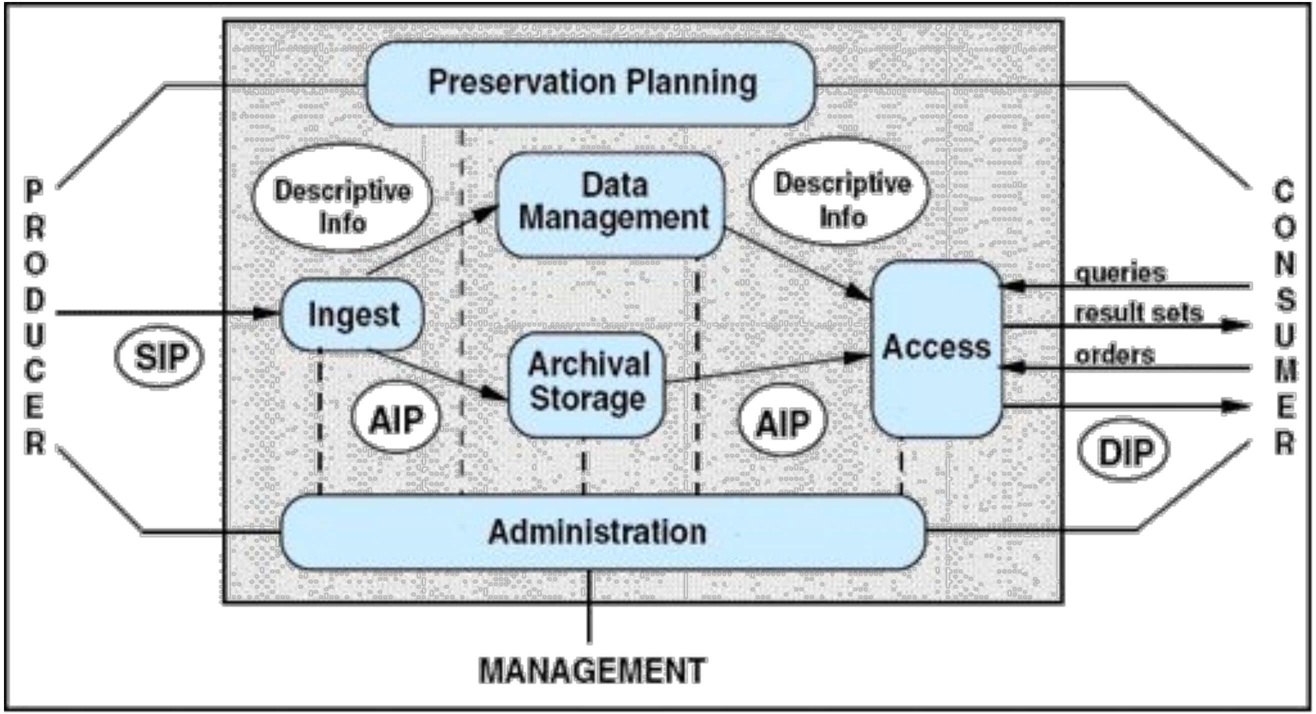
- Strategy
  - Establish a POLICY
  - Run updated Preservation Plans
- Access guarantees
  - Availability and Security
  - Authenticity and integrity
- Usability
  - Derived proxy formats
- Trust and Sustainability
  - ISO 16363 & Data Seal of Approval

The Open Archival Information System (OAIS)

SIP, AIP and DIP:

Information

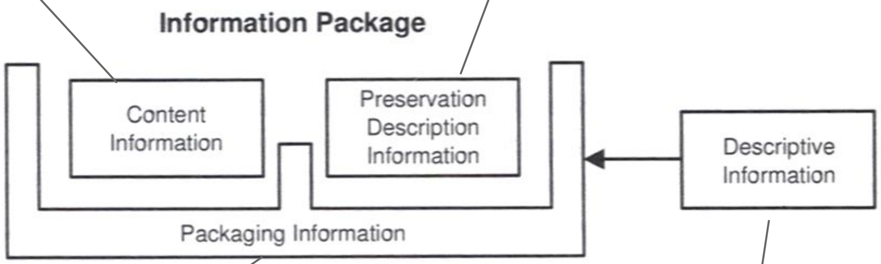
Packages



Document → Archival Information Packages (AIP)

E.g. title, author, abstract, etc

E.g. checksum, digital signature

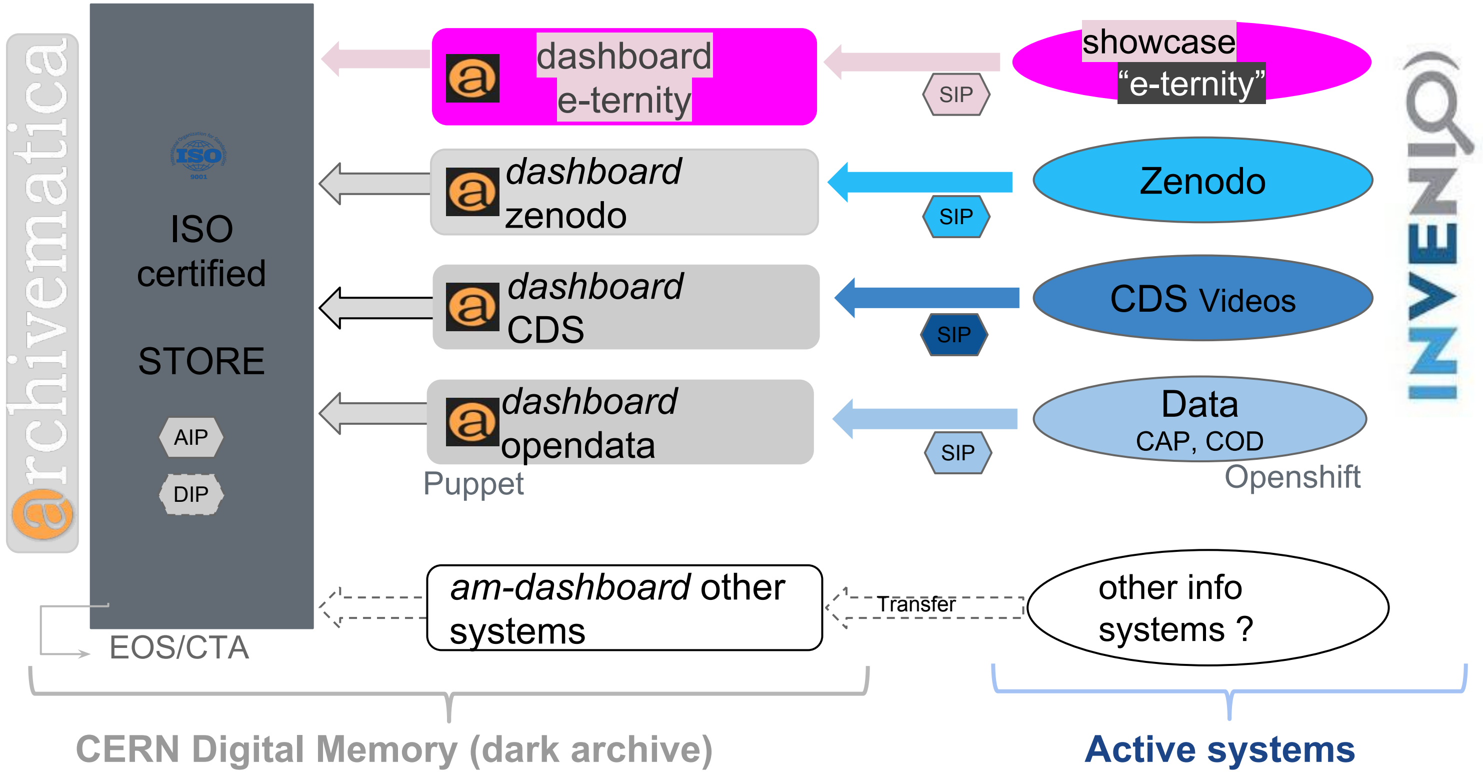


E.g. directory structure, filenames, tape marks

E.g. access control, finding aids

→ Must have redundant copies
→ Must be regularly checked
→ Must be supported by preservation plan
→ Must be sustained by an organizational policy

Conclusion: E-Ternity at CERN



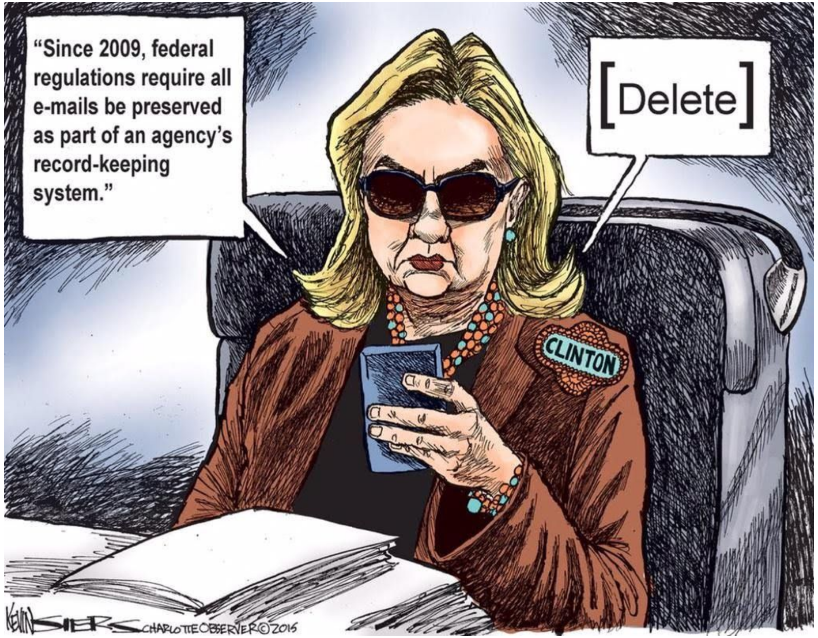
Different Approaches for different content types

© Carlos Perez Naval

Personal Digital Archiving (PDA)

- Family Memory: the “long tail” of Collective Memory
- “Memory services” to individuals
  - Associations: e.g.http://saa.archivists.org http://thedigitalbeyond.com
  - Commercial PaaS: http://forever.com   http://bcelebrated.com http://digi.me (social media)
- “Family” data services : QR-codes on gravestones, funerary IT services, genealogy trees with data, digitization...

Archiving Users’emails


“Since 2009, federal regulations require all e-mails be preserved as part of an agency’s record-keeping system.”
[Delete]
CLINTON
KEVIN SIERS CHARLOTTE OBSERVER ©2015

- In US, the CAPSTONE policy enforces email preservation to all federal agencies - usual embargo period of 50 years

often the best source of information to understand exactly what has happened!

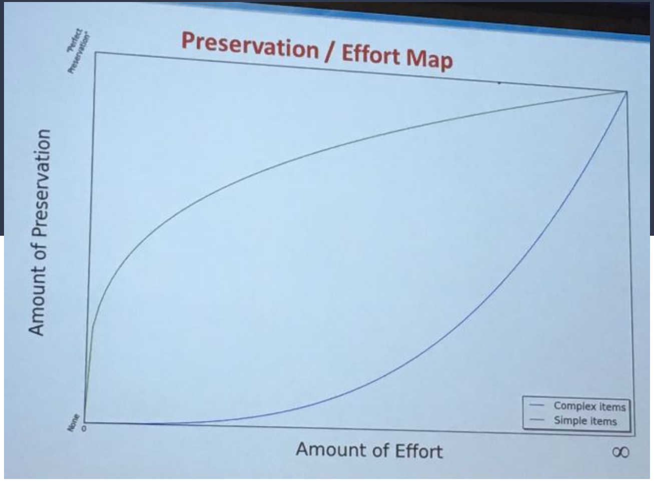
Emails: Born Digital with metadata out of the box

- High volumes, Disorganised, Non Uniform, Specific search, Human intervention, Depends on technology
- Projects: *E@Sy* at Harvard , *ePADD* at Stanford, *Darcmail* at Smithsonian, *TOMES* at Kansas univ etc.
- Weeding Tools: many free tools to help extract, convert, analyse. E.g. : Emailchemy / Aid4Mail / Ringtail
- Challenges: policy building ; scalability ; harvesting issues ; privacy concerns
- Workflow example: retire/volunteer → email analysis e.g. with ePADD→ cleaned mbox → Digital Archive Store

Archiving Dynamic Web Sites

- Standard format for Web archiving: **WARC** (previously ARC)
- Web Crawlers (e.g. Internet Archive) **limitations**:
    - Misses restricted pages - Relies on ‘robots.txt’ - Misses complex pages (with Javascript, audio, videos, etc)
- WebRecorder tool to capture site in WARC
- **WebPlayer** to display WARC in browsers, e.g Wayback machine, pywb extension
- Enable collecting Tweets, Facebook pages, etc

Archiving Multimedia

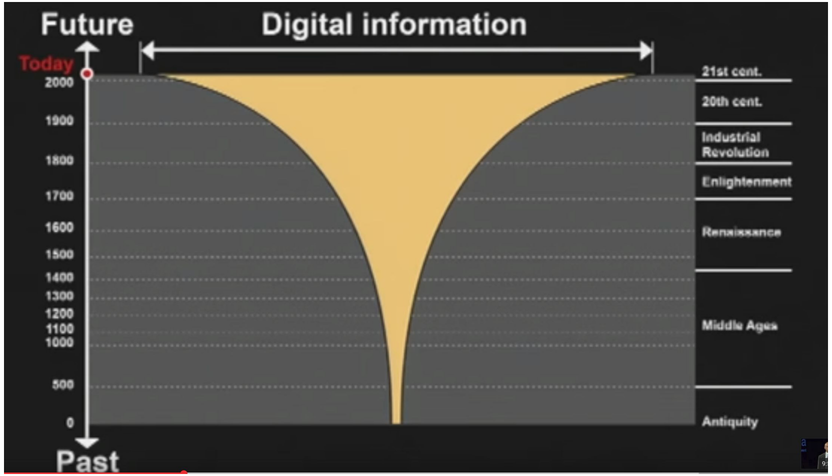


- UNESCO Memory of the World Programme
- Audiovisual lifetime = ~100 years only !
- Videos: 100 different formats in the 40 years since video-tape recording started
- Magnetic tape storage popular, but very fragile
- Photo: captioning is tricky
- Sound always existed but has never been archived : the most short-lived artistic asset produced in history !

Digitizing Paper Archives

- The Venice Time Machine project: 80 km of shelves
  - to enable navigating ~1000 years back in time!
- Historical and Scientific papers
- Administrative and Personnel papers
- The investment needed for these digitizations must be compared with the cost of managing the physical folders, the risk of losing content, plus the return-on-investment considering the added-values of having a fully electronic Archive available to authorized end users.

Preserving digitally-born Content

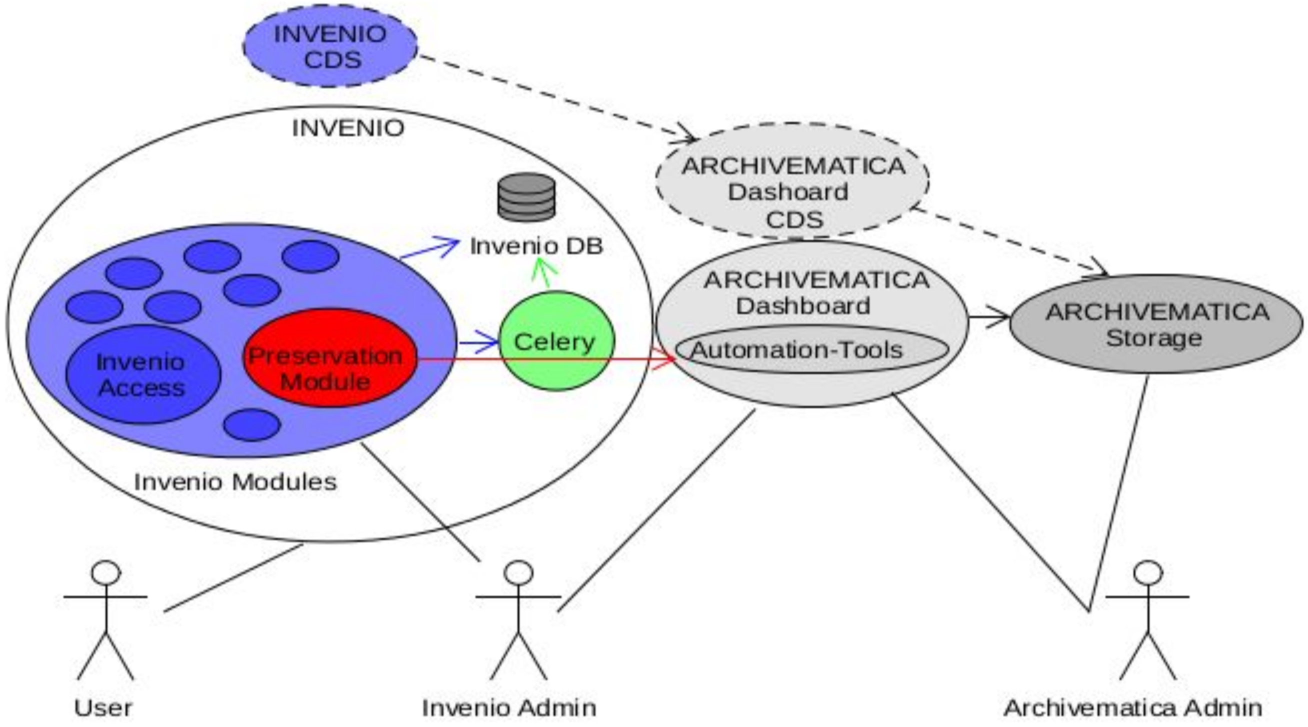


- Today’s trend: from the “capture-all” to the “select-few“
- E-journals: the perpetual access issue → the LOCKSS solution
- Blogs and social media supplanting letters, journals, etc. ; Facebook “Look Back” option
- Conferences, meetings, maps, equipments…

Conclusions

© Gennady Fedorenko

Conclusion: Digital Memory at CERN



Conclusion: evolution of Library Systems

DPC: Making the **obsolescence** obsolete ?

The upstream approach:

any IT system managing information worth-preserving should be designed with preservation in mind.