AAF - Groupe PIN

Les données numériques, une espèce en voie d'extinction ?

Digital data, an endangered species?

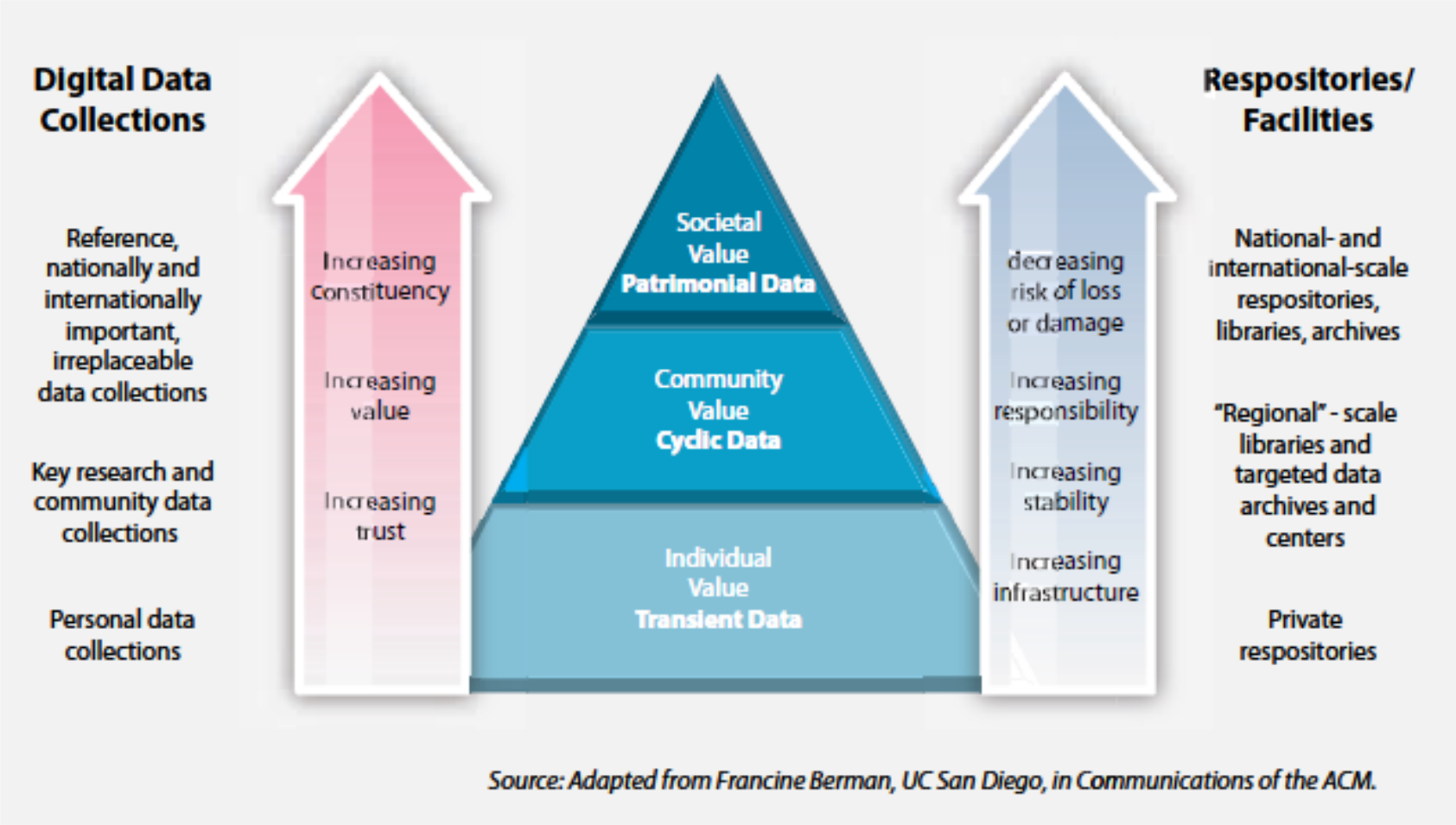
Digitally encoded information, an endangered species?

Dr David Giaretta
Director of PTAB
david@giaretta.org
http://www.iso16363.org

Outline

- Digital Preservation – a personal perspective
  - Looking back 30 years and forward 10+ years
  - Digital Data – which encodes Information – not “just the bits”
- What is likely to be preserved – value/volume?
- Data Lifecycle
- Why preserve?
- What is likely to be preserved - type
- Is data endangered?
  - What risks
  - How can it be saved?
- Who can/will save it?
- Is digital data an endangered species?

Information pyramid – a hierarchy of rising value and permanence



Source: Adapted from Francine Berman, UC San Diego, in Communications of the ACM.

Things change

1990

- Web not yet begun
- XML not yet begin
- Internet speeds kbps in universities and offices
- 300,000 internet hosts
- Data volume ??
- XXX researchers
- Few computer programming languages
- Transition from text to 2D image visualisation

2010

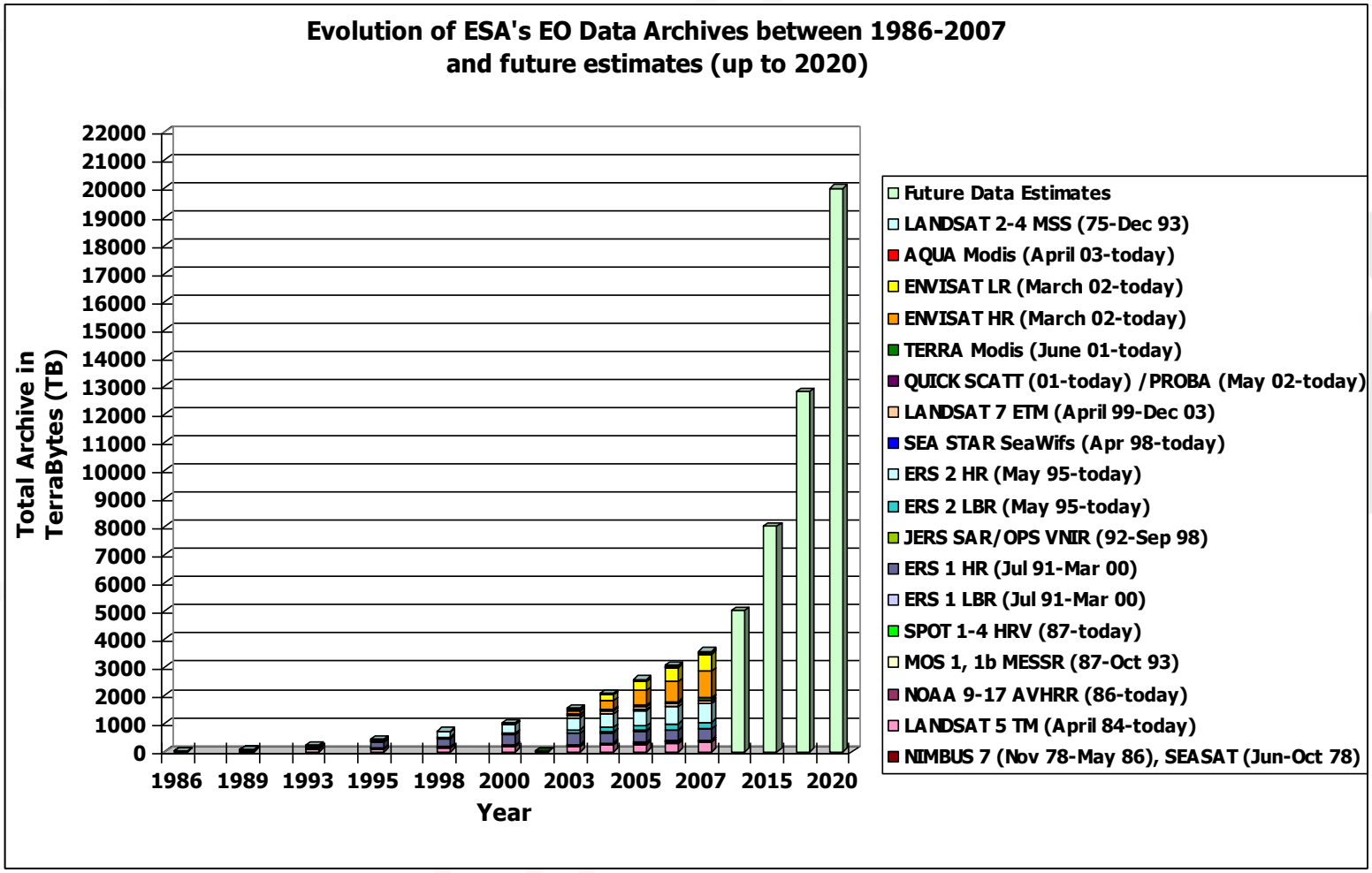
- Web 2.0 started
- XML widespread
- Internet speeds Mbps widespread
- 600,000,000 internet hosts
- 5.10^{18} bytes of data
- Millions of researchers
- Many new paradigms for programming languages
- 3-D and Virtual reality visualisation
- Cloud growing

2030

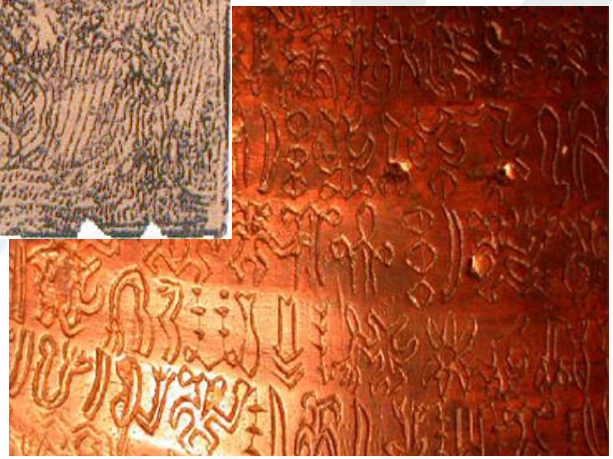
- Semantic Web?
- AI?
- XML forgotten?
- Java not used?
- Internet speeds Pbps widespread
- 2,000,000,000,000 hosts
- 5.10^{24} bytes of data
- Billions of citizen researchers
- Natural language programming for computers
- Virtual worlds (Meta?)

Work on OAIS started 1992 | OAIS v1 2002 | OAIS v2 2012 | OAIS v3 2022 | OAIS v4 2032?

Data grows


Evolution of ESA's EO Data Archives between 1986-2007
and future estimates (up to 2020)
Total Archive in TerraBytes (TB)
22000
21000
20000
19000
18000
17000
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1986
1989
1993
1995
1998
2000
2003
2005
2007
2015
2020
Year
Future Data Estimates
LANDSAT 2-4 MSS (75-Dec 93)
AQUA Modis (April 03-today)
ENVISAT LR (March 02-today)
ENVISAT HR (March 02-today)
TERRA Modis (June 01-today)
QUICK SCATT (01-today) / PROBA (May 02-today)
LANDSAT 7 ETM (April 99-Dec 03)
SEA STAR SeaWifs (Apr 98-today)
ERS 2 HR (May 95-today)
ERS 2 LBR (May 95-today)
JERS SAR/OPS VNIR (92-Sep 98)
ERS 1 HR (Jul 91-Mar 00)
ERS 1 LBR (Jul 91-Mar 00)
SPOT 1-4 HRV (87-today)
MOS 1, 1b MESSR (87-Oct 93)
NOAA 9-17 AVHRR (86-today)
LANDSAT 5 TM (April 84-today)
NIMBUS 7 (Nov 78-May 86), SEASAT (Jun-Oct 78)

Documents, Paper and Vellum



Last for hundreds of years
.....even if neglected

.....but no guarantee that the text can be understood

Rongorongo text on wooden tablets on Easter Island - not translated

Digital ….. Word and PDF

- The obvious analogy with paper documents … (just print and display)
  - WordStar, WordPerfect, MS Word , PDF (various version)
- But……
- Problems include
  - Availability of the software
    - Availability of the Operating systems on which the software runs
      - Availability of the hardware in which the Operating System runs
  - Backward compatibility
    - Later version cannot read file produced by earlier version of the software
    - Later version produces a different result from earlier version
  - …and what it means e.g. a document written in Chinese or using special abbreviations or technical terms – which the reader does not understand

More complicated things:

- Modern publications with embedded applications
- MS Word files with embedded links to spreadsheets and databases
- Engineering designs – Computer Aided Design
- Scientific data
  - High Energy Physics
  - Astronomy
  - Biology, Genomics, …
- Finance data
- Websites with embedded applications
- Massively distributed data systems such as globally interconnected linked data systems
- ….

Simply being able to printing the numbers or text in the future is not enough!

Digitally encoded information – 1's and 0's

- BITS: 01001110 01001101 01010001 01001101
  01010000 01001010 00100000 00100000
- HEX: 4e 4d 51 4d 50 4a 20 20
- Two IEEE 754 32 bit real numbers:
  8.6116461E8 1.35644119E10
- Two 32 bit integers 164211241 168379396
- Actually... ....
- ASCII Characters: NMQMPJ
- ……. Was my flight reference

Example:
"ca fe ba be" at start
indicates Java class file

Assuming "big-endian"

What does this mean?

...semantics ...

Could be encoded as Comma Separate Value (CSV) file in ASCII or Unicode

Can anyone guess what this table means?

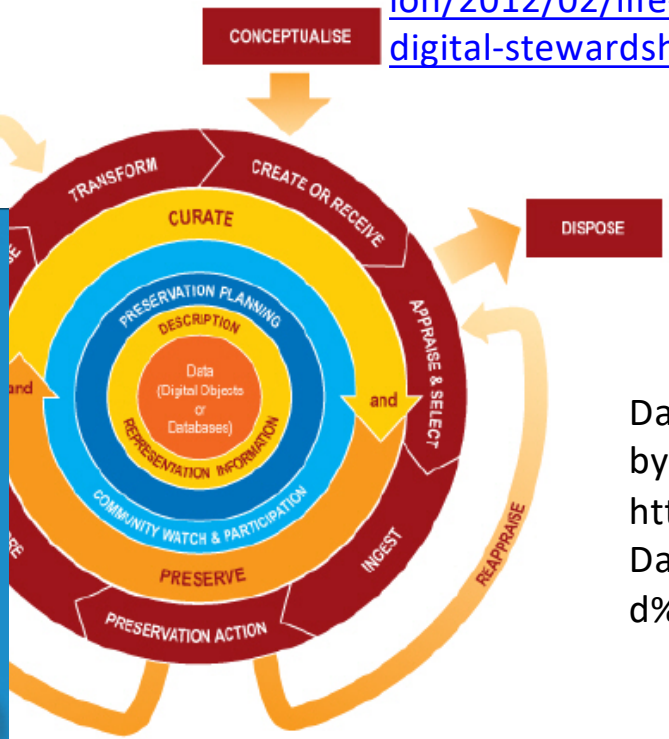
Latitude	Longitude	Ozone	Date
132	50	34.9	12/03/1999T17:20:43.1
178	50	45	12/03/1999T19:37:52.7
190	50	78	12/03/1999T21:16:23.9

So many threats to preservation ….

- The bits may be lost – accident or on purpose or floods or earthquakes
- Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.
- Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.
- Evidence may be lost because the origin and authenticity of the data may be uncertain.
- Access and use restrictions (e.g. Digital Rights Management) may not be respected in the future.
- Loss of ability to identify the location of data.
- The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
- The ones we trust to look after the digital holdings may let us down.

From PARSE.Insight global survey of researchers and data managers
(http://www.alliancepermanentaccess.org/index.php/community/current-projects/parse-insight/)

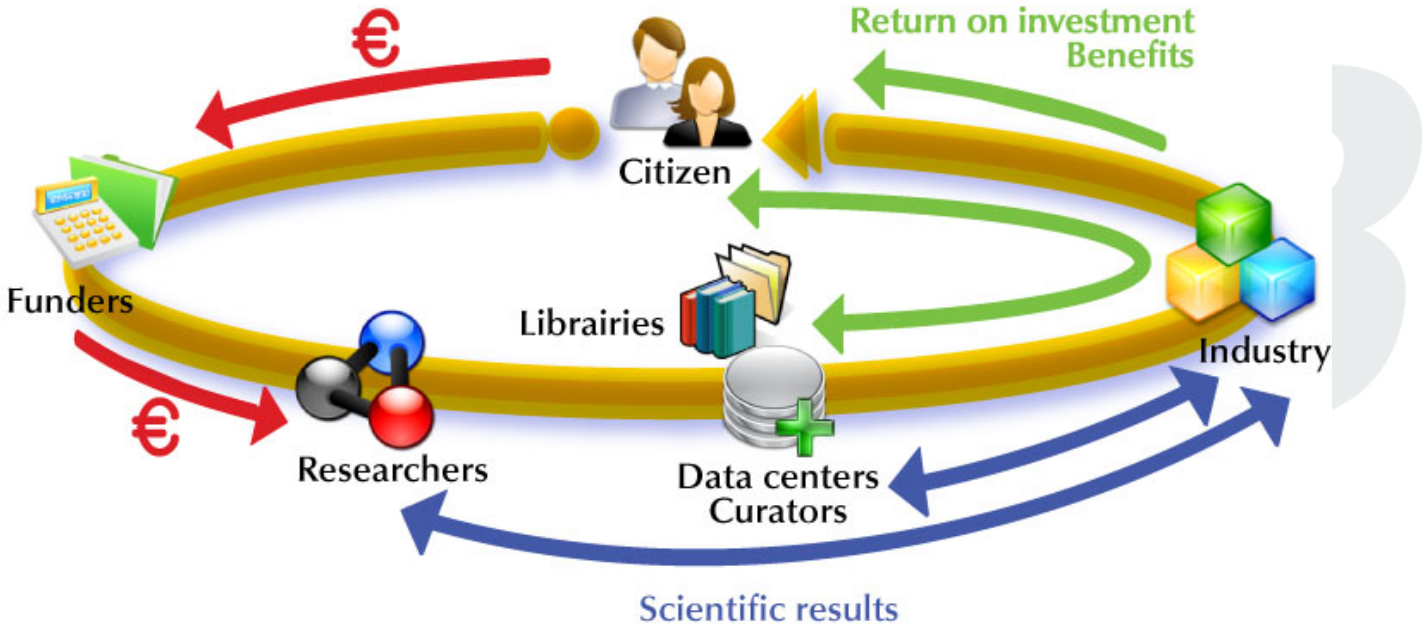
Many lifecycle models



See http://blogs.loc.gov/digitalpreservation/2012/02/life-cycle-models-for-digital-stewardship /

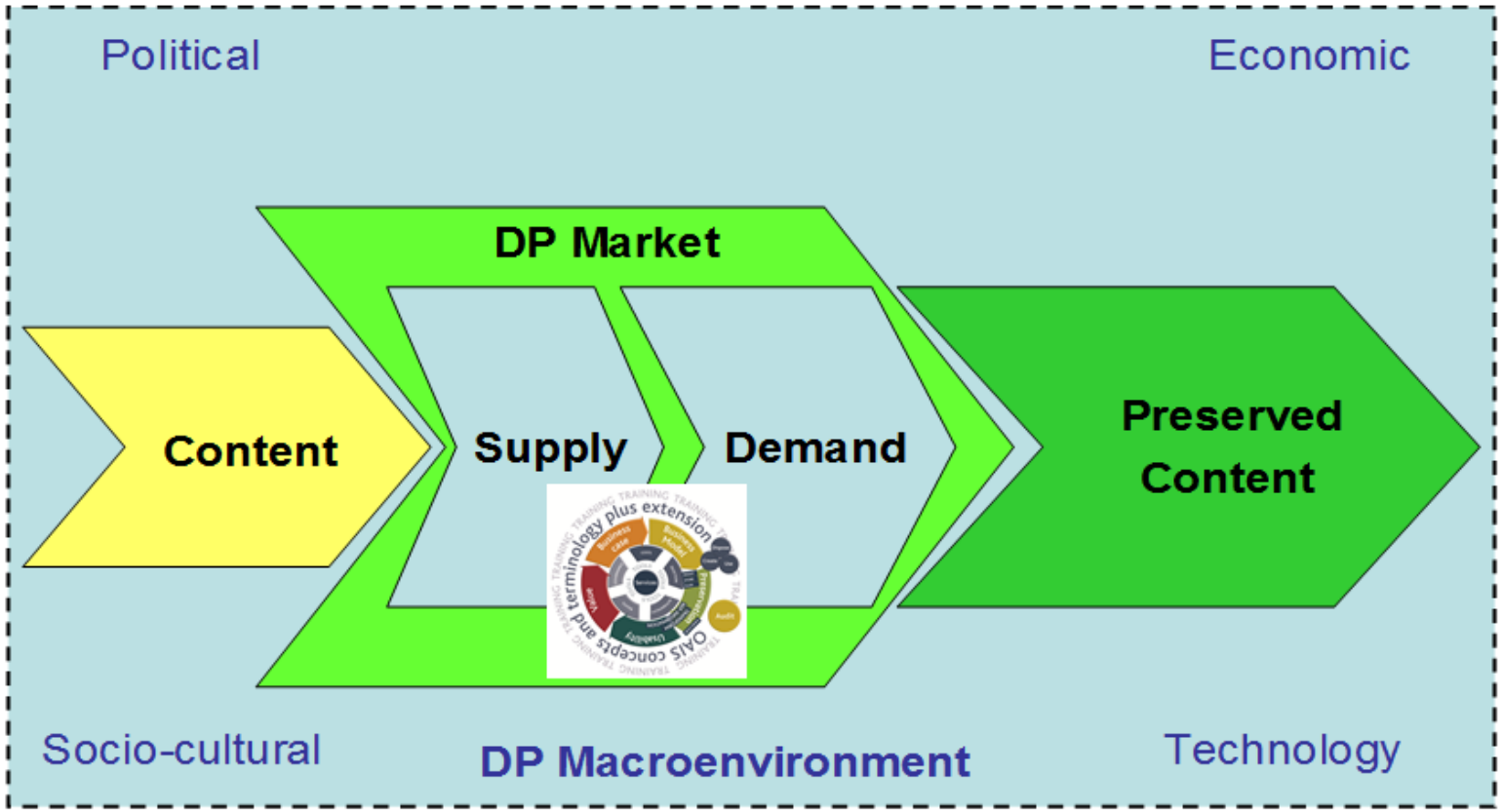
Data Lifecycle Models and Concepts by CEOS, 2012, see http://www.ceos.org/images/DSIG/Data%20Lifecycle%20Models%20and%20Concepts%20v13.docx

Value? Commerce? Industry? People? Market




TRAINING TRAINING TRAINING TRAINING TRAINING TRAINING TRAINING TRAINING
terminology plus extension
Business case
Business Model
Dispose
Use
Audit
Value
Preservation
Usability
Services
TOOLS
COSTS
SELECTION
STORE
PLAN
RIGHTS

Market model

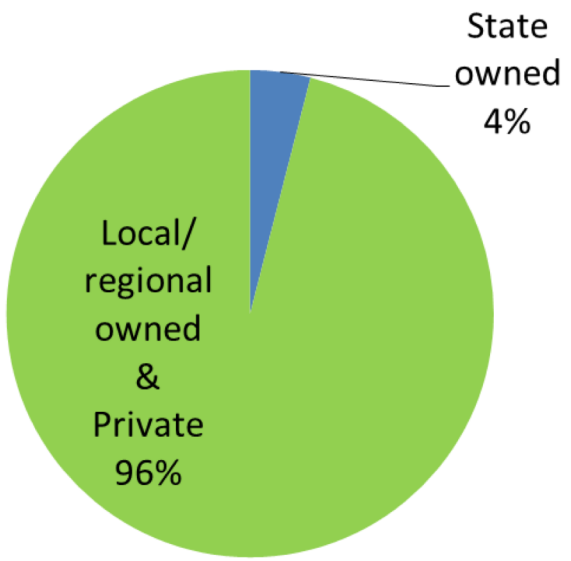


Some Customer segments

- ✓ Memory Institutions (MI)
- ✓ Scientific Research Institutions (SRI)
- ✓ Highly exposed industries

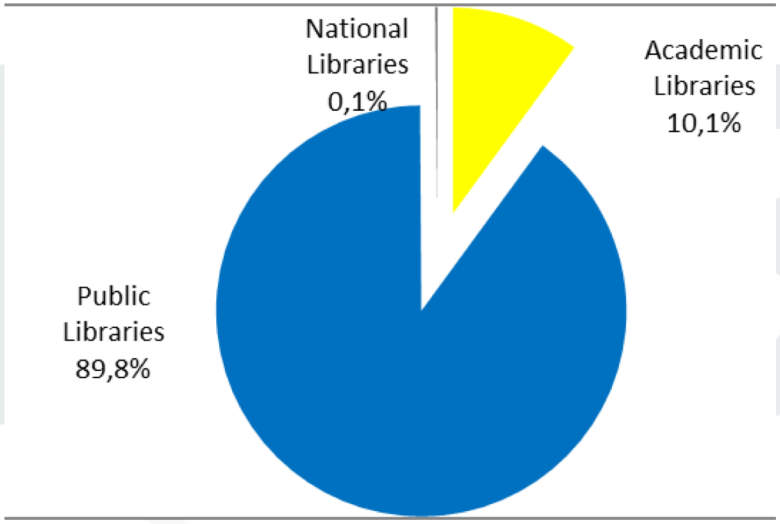
Potential demand – Memory Institutions

Museums: 151,361 EU



The European Group on Museum Statistics (EGMUS) 2013

Libraries: 120.000 EU



ALBA (Association of Leading Visitors Attractions -2013-)

Archives: at least 16,000→ about 2,5% national archives

APENET- Archives Portal of Europe on the Internet. 2008

Potential demand – **MI patterns of usage**

- A few MI lead the DP pioneering
- Digital content deluge and digitalization→ increasing demand for DP.
- DP challenges depend on: the size, mandate, and type of content
- Libraries with annual budget over 50K and more than 10TB → DP
- MI are understaffed for DP and there is need of skilled staff

Potential demand – **Scientific Research Institutions**

- About 31,000 SRI, comprising:
  - Government Scientific Research Institutions
  - University Research Institutions
- **Grow as communities around large research infrastructures**
  - ESFRI: 48 Infrastructures (10 under implementation – 38 to go);
  - EGI: 340 organisations, 34 countries, 2 EIROS, 212 Virtual Organisations, 22,000 Researchers, 373,800 CPUs and 190 PB storage.
  - OSG: 72 institutions, 115 countries, 25,000 computers, 43,000 procs.
  - GEANT: 40 million users (EU), 10,000 institutions, 65 countries
  - PRACE: HPC ecosystem, 20 members, 20 EU countries. Funding secured form FP7

- DP not in their culture

+ RI / DMP in H2020

Potential demand – **SRI Decision Making**

- Decision makers recognize importance of pubic value and data re-use
- Researchers believe government, publishers or research organisations should pay for DP

Incentives for decision making

- Results should become public if publicly funded
- Stimulate advancement of science
- Allows Re-analysis
- It is unique
- It might serve for validation in the future
- Stimulate inter-disciplinary collaborations
- Potential economic value

Potential demand – **Industry**

- Industry has challenges for preserving content
  - Information Lifecycle management
  - Information Governance and Risk management (IG&R)
  - Archiving

Potential demand – Cultural and Creative Industry

Sub-sectors	**Cultural Industries/Media & Entertainment**: Film & video, TV & Radio, Video Games, Music, Books & Press (Publishing) **Creative Industries**: fashion, graphic, interior, product design **Heritage**: Museums, Libraries, Archives & Archaeological sites **Other core arts**: visual and performing arts
Compliance	20-50Years for music, prototypes and designs, +100 for long tail (e.g. film, cultural heritage) IPRs (copy rights and trademarks). Activities based on massive reproduction
Challenges	Transition to the digital content era Extract, combine and manage external and internal data Manage a complex cooperation / collaboration environment with new entrants and social media How to provide value-added services to potential customers (based on content) Maintain quality as a competitive advantage Create dynamic and interactive experiences based on existent content

Potential demand – Energy Industry

Revenues	• **435 nuclear reactors in 30 countries and generate 14% of the globe's electricity** (World Nuclear Association) • Demand for electricity will grow faster than any other energy. • Demand growth: 70% (2010-2035) - 2.2% /year • Electricity industry the largest end-use sector through 2035 • The electricity sector's annual turnover of **€420 billion** • +3% of European GDP
Demand & Others	• To maintain safety, security, and compliance at power plants, • Management must have well-documented and highly visible information across the asset life cycle. • Photovoltaic and wind energy capacity increasing

Unsatisfied needs– Memory Institutions

- Increase the awareness of the fragility of digital content, and create consensus of what, when and how data have to be preserved
- Cohesive policies across departments
- Available financial resources: better “cost data”, for complex cost models and funders.
- Dedicated staff or team for DP activities and adequate staff
- Training for local government officials, especially archivists in the Archives case.
- Funding for training need to be increased

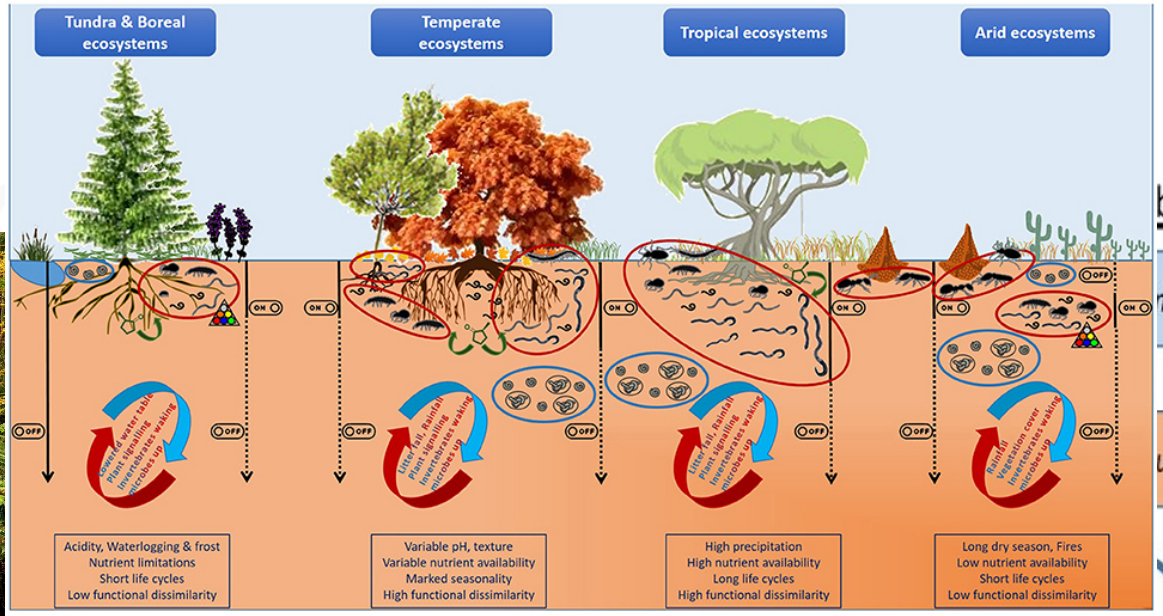
Unsatisfied needs– SRIs

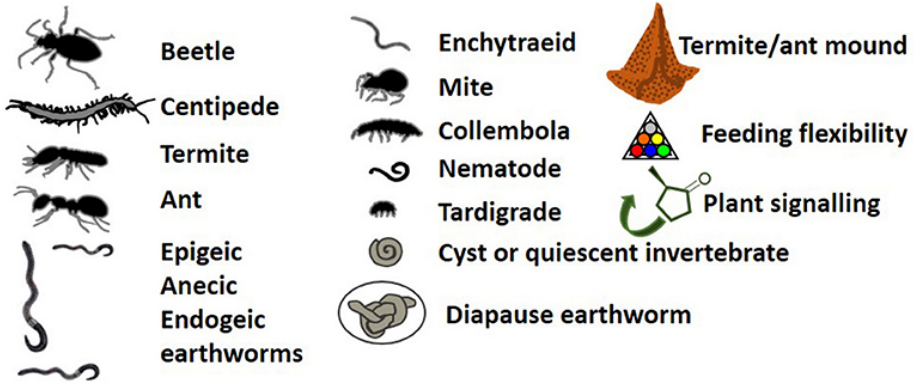
- Research institutions need to develop its DP policies in an integrated way, currently they are developing DP policies separately as policy makers and funders.
- Need for partnership among publishers and data producers.
- SRI expect publishers to add value to core journal content, including active content, visualization an analytics
- There is a need of semantic enrichment and linked data to make content smarter and improve discoverability.
- Define better, with publishers, the role of data as part of the research outputs and publishing of journals
- There is a need to train researchers
- Distrust towards digital archives generate that researchers keep their research data in personal computers at work (80%)

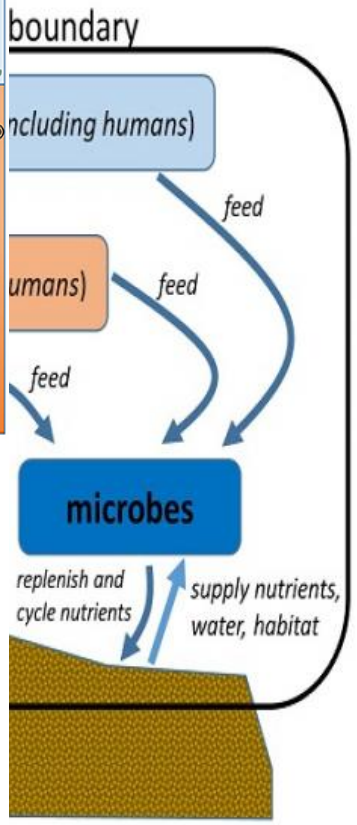
Unsatisfied needs - Industry

- Tools and deployment that support storage growth management, as well as designing needs for storage, virtual storage, cloud services and disaster recovery solutions.
- Tools and services for more efficient content access, analytics and content management (especially disposal of files).
- Tools and services for strategic and big picture analyses
- Tools and services to support increasing need of security and risk
- Getting future-proven products.

Ecosystems







This Photo by Unknown Author is licensed under CC BY-SA-NC

Nature – red in tooth and claw

By Frits Ahlefeldt

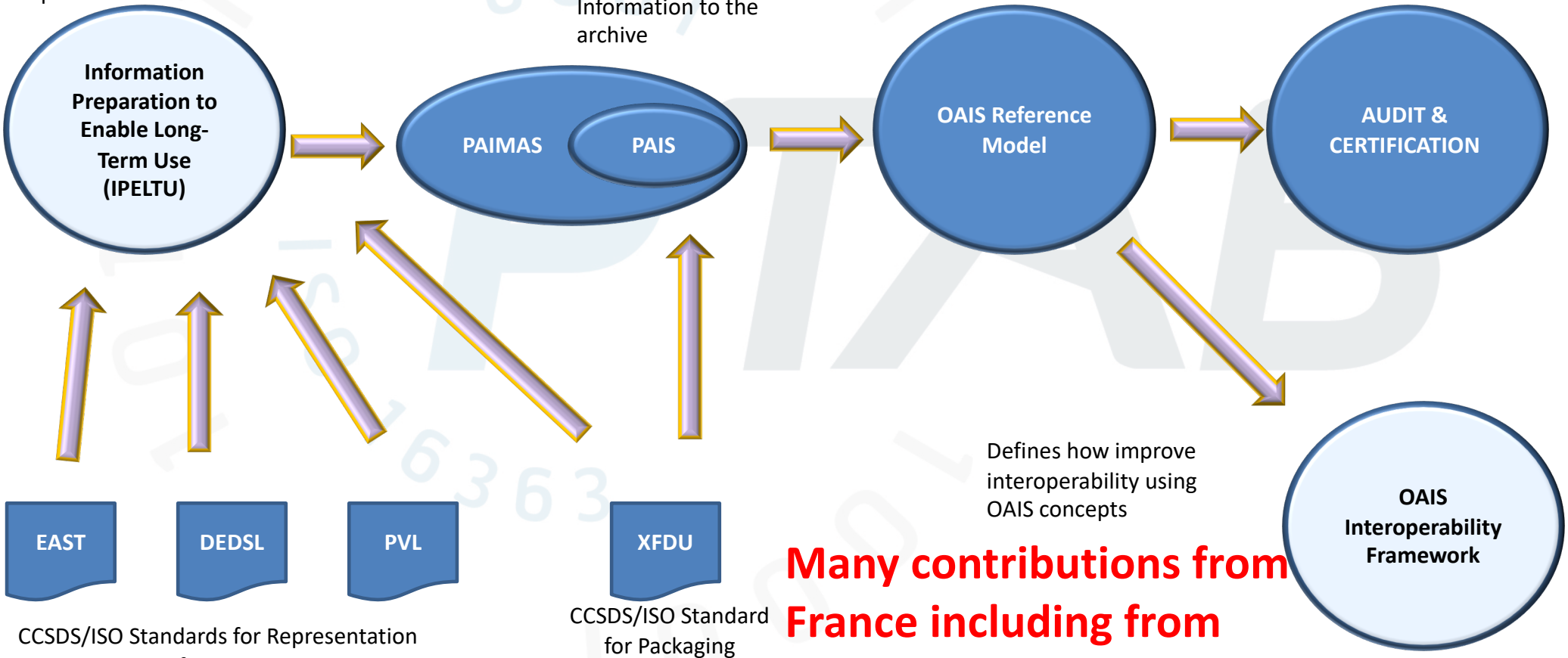
Relationship between OAIS family of standards

Guides the collection/creation of the Additional Information required

Defines a mechanism to transfer Data and the appropriate Additional Information to the archive

Defines how the information should be preserved

Defines how to check that the information is being preserved



CCSDS/ISO Standards for Representation Information

CCSDS/ISO Standard for Packaging

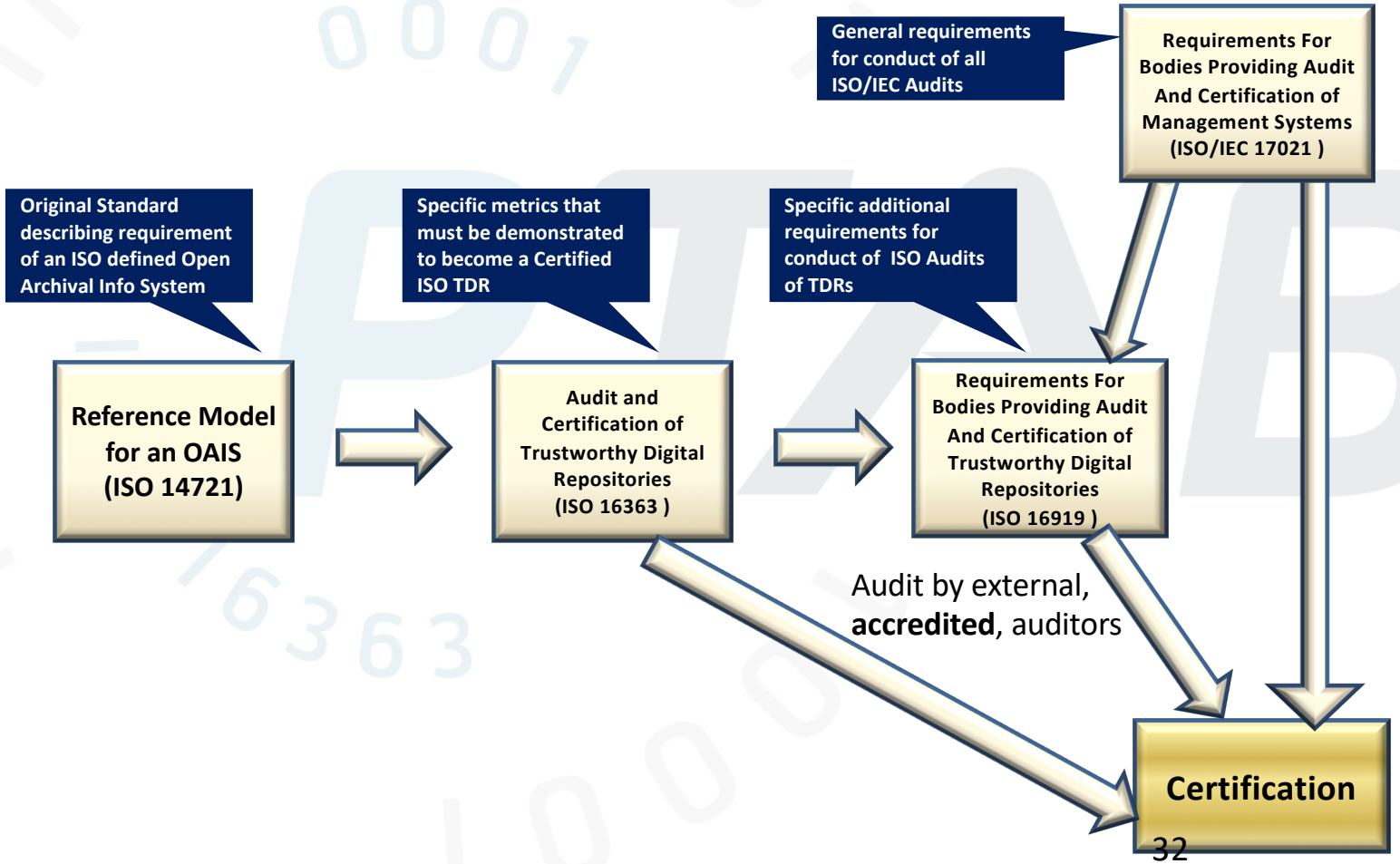
Defines how improve interoperability using OAIS concepts

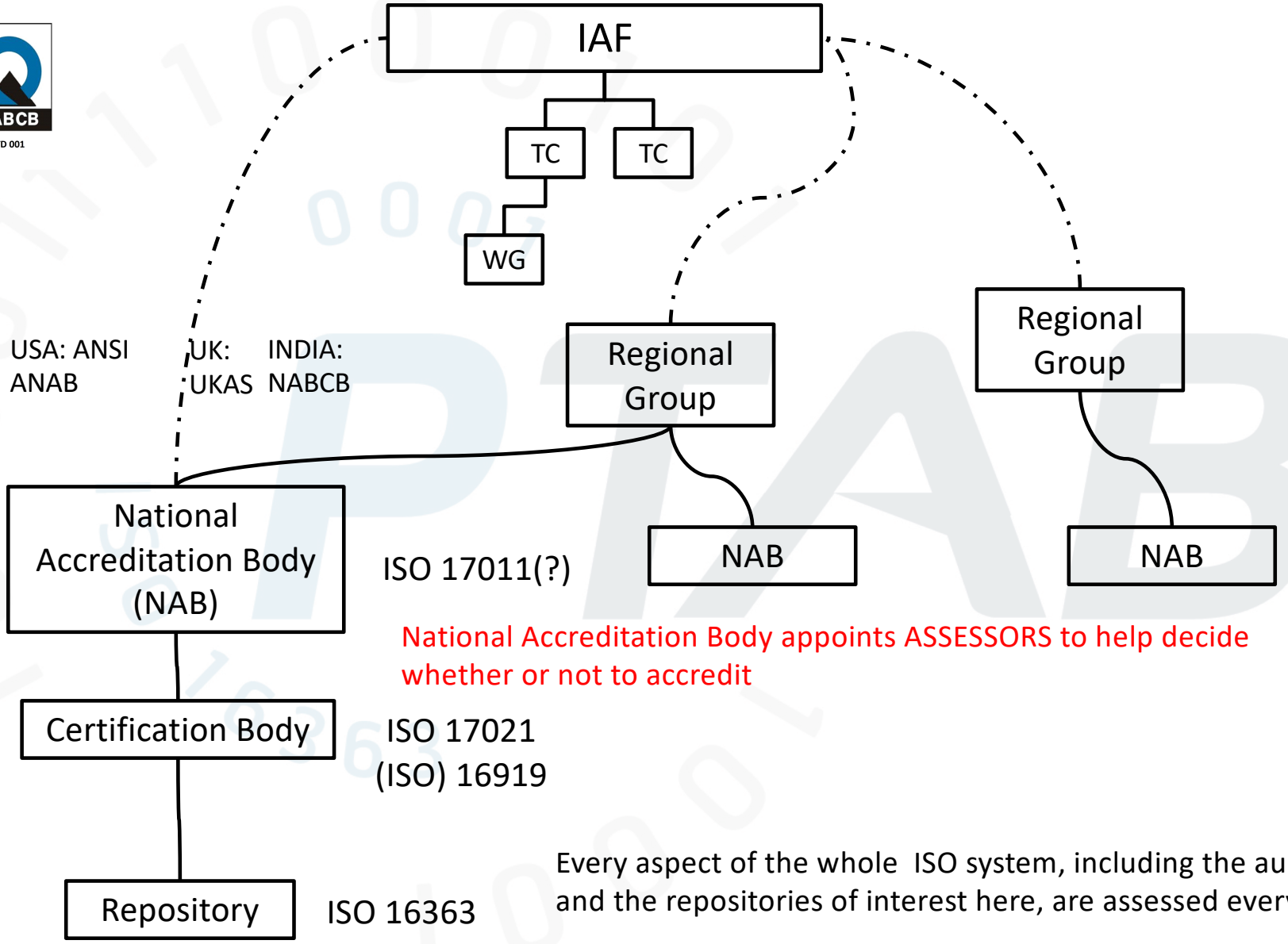
Many contributions from France including from CNES and GAEL

What is a Trusted/Trustworthy Digital Repository?

- Trusted for what?
  - Honesty?
  - Speed of transactions?
  - Security of information about its clients?
  - "Truth" of the contents of the repository?
  - Records Management?
  - Preservation of digitally encoded information?

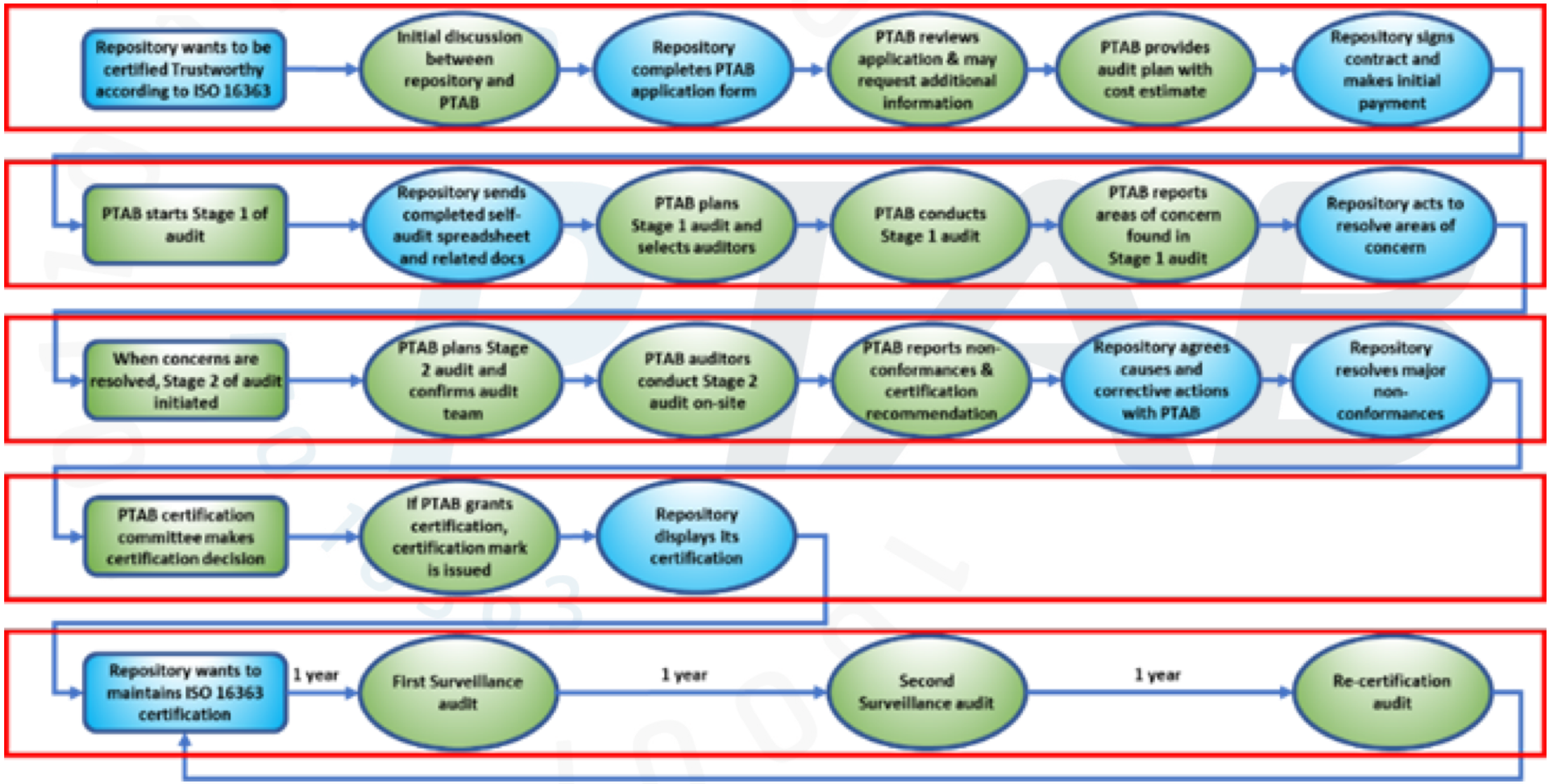
Relationship between standards



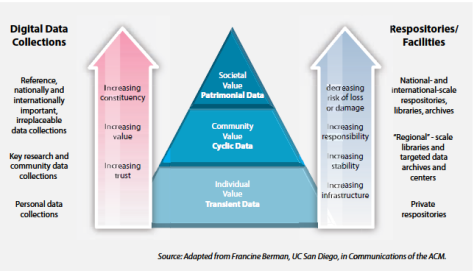


Every aspect of the whole ISO system, including the auditors and the repositories of interest here, are assessed every year.

ISO/PTAB process



Is digital data an endangered species?



- Some (perhaps most) data will disappear
- Some data is too valuable to lose – it will be preserved and used – for as long as it is recognised as being of value
    - Scientifically, culturally, legally, commercially ….
- The rest, not currently recognised as sufficiently valuable, will require careful consideration

The rest will be in danger ….

- **IF** “data” is/are preserved in the same way as “documents” and “images”
  - Misunderstand Rendered and Non-rendered
- **IF** the value of information is not recognised in time
  - Resources for preservation will be inadequate
- **IF** software vendors’ claims are believed
  - E.g. “we are OAIS compliant”/ “we create AIPs” / “we are certified”
- **IF** Archives are not assessed adequately
  - All will claim to be “trustworthy” – but are they? And for how long can they be trusted?

References

- All the following are being updated
  - Reference Model for an Open Archival Information System (OAIS). Magenta Book. Issue 2. June 2012, available from https://public.ccsds.org/Pubs/650x0m2.pdf also known as . ISO 14721:2012
  - Audit and Certification of Trustworthy Digital Repositories. Magenta Book. Issue 1. September 2011., available from https://public.ccsds.org/Pubs/652x0m1.pdf also known as ISO 16363:2012
  - Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories. Magenta Book. Issue 2. March 2014, available from https://public.ccsds.org/Pubs/652x1m2.pdf also known as ISO 16919:2014
- PTAB http://www.iso16363.org