An assessment of costs and risks in the operation of long-term digital archive infrastructures

Jens Klump
German Research Centre for Geosciences

GFZ Helmholtz-Zentrum Potsdam

HELMHOLTZ GEMEINSCHAFT

Objective

- Digital information has become part of our cultural heritage.
- Scientific findings are increasingly presented in electronic form – often exclusively so.
- At the same time the underlying technology is changing rapidly.
- Many data in the earth and environmental sciences are unique and cannot be produced again.

- Whom do we entrust our digital heritage?
- How can we organise long-term data preservation?
- What are the costs of long-term data preservation?

Dealing with research data

Introduction

Research Data – The Challenge

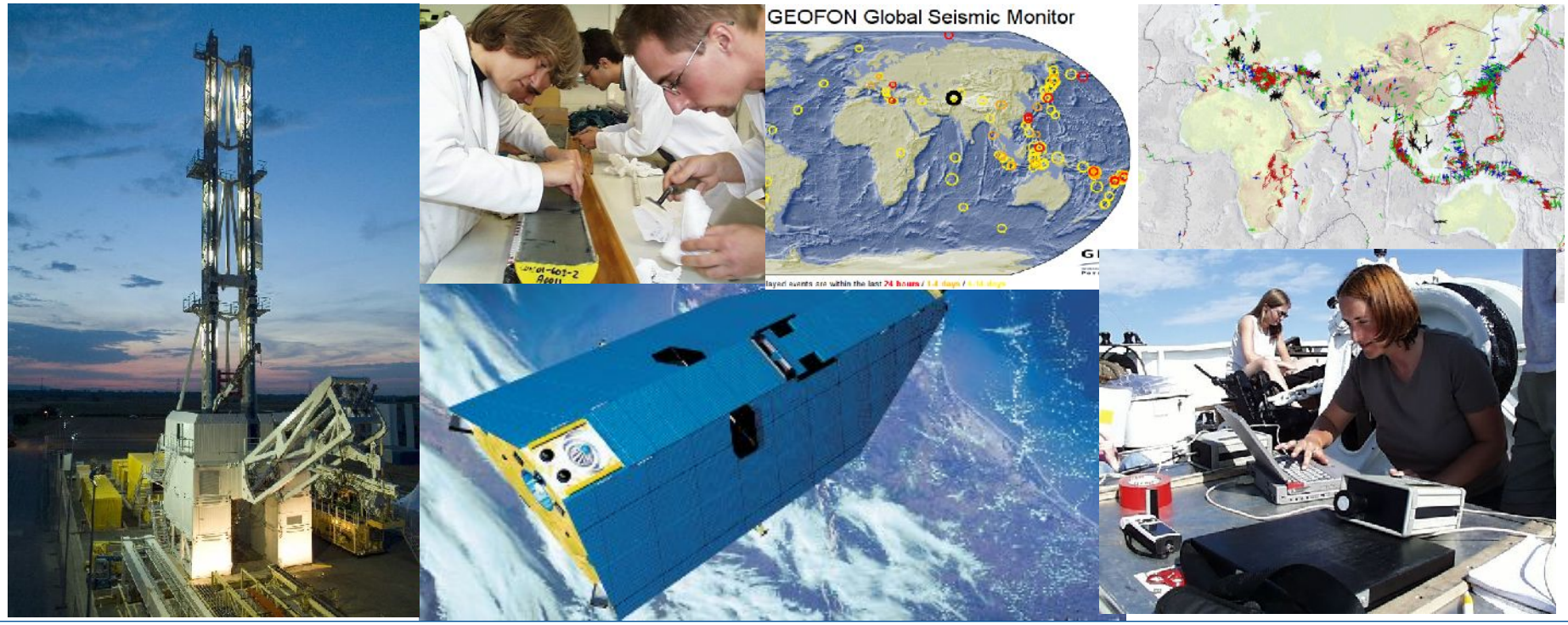
Curating research data is like herding cats.

The challenges:

- Large volumes of data
- Heterogeneous data structures
- Changing user requirements
- Short-term projects

Data Diversity

- The challenges of digital data curation are both the vast amounts of data and their diversity.



Data Types

- Data from sensor networks and automated processes
  - Large volumes
  - Homogeneous structures
  - Structured workflows
  - Data management plans exist
- Model data
  - Large volumes
  - Homogeneous structures
  - Partly structures workflows
- Individual data
  - Small volumes
  - Heterogeneous structures
  - Ad hoc workflows
  - No data management plan

Long-term preservation challenges specific to digital objects

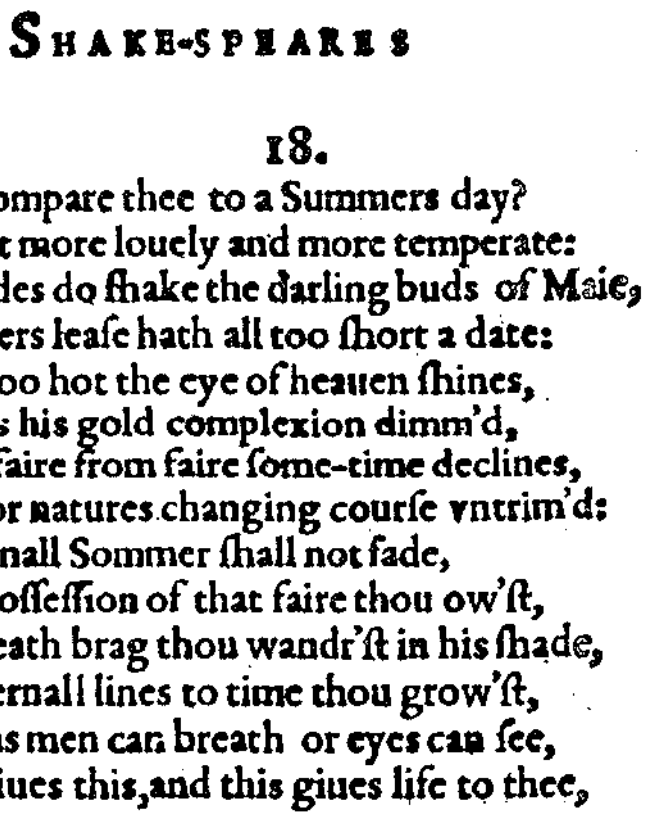
From: Funk (2010), In: nestor Handbuch 2.3, Ch. 7.2

Specific Challenges

- Bit stream preservation
  - Making copies is a centuries old cultural technique
- Representation information
  - Secure data formats, migration, emulation
- Content information
  - Cataloguing contents, metadata
- Significant properties
  - What property of the object is it, that is to be preserved?

Thibodeau (2002), Rothenberg (1997, 2001)

Significant Properties



SHAKE-SPEARES

18.

SHall I compare thee to a Summers day?
Thou art more louely and more temperate:
Rough windes do ſhake the darling buds of Maie,
And Sommers leaſe hath all too ſhort a date:
Sometime too hot the eye of heauen ſhines,
And often is his gold complexion dimm'd,
And euery faire from faire ſome-time declines,
By chance,or natures changing courſe vntrim'd:
But thy eternall Sommer ſhall not fade,
Nor looſe poſſeſſion of that faire thou ow'ſt,
Nor ſhall death brag thou wandr'ſt in his ſhade,
When in eternall lines to time thou grow'ſt,
So long as men can breath or eyes can ſee,
So long liues this,and this giues life to thee,

SHALL I COMPARE THEE TO A SUMMERS DAY? THOU ART MORE LOVELY AND MORE TEMPERATE: ROUGH WINDES DO SHAKE THE DARLING BUDS OF MAIE, AND SOMMERS LEASE HATH ALL TOO SHORT A DATE: SOMETIME TOO HOT THE EYE OF HEAVEN SHINES, AND OFTEN IS HIS GOLD COMPLEXION DIMM'D, AND EVERY FAIRE FROM FAIRE SOME-TIME DECLINES, BY CHANCE, OR NATURES CHANGING COURSE UNTRIM'D: BUT THY ETERNALL SOMMER SHALL NOT FADE, NOR LOOSE POSSESSION OF THAT FAIRE THOU OW'ST, NOR SHALL DEATH BRAG THOU WANDR'ST IN HIS SHADE, WHEN IN ETERNALL LINES TO TIME THOU GROW'ST, SO LONG AS MEN CAN BREATH OR EYES CAN SEE, SO LONG LIVES THIS, AND THIS GIVES LIFE TO THEE,

Outlining the challenge ahead

Long-term preservation

Definition of Long-Term

- Computer Science: > 5 yrs.
- DFG, MPG: > 10 Jahre.
- Engineering: > 30 Jahre.
- Linguistics: > 100 Jahre.

What does long-term mean for research data?

Cf. DFG Collaborative Research Centres: 15 years run-time plus ten years of archiving after end of project means 25 years of data curation.

Defition of Long-Term

"A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information held in a repository. This period extends into the indefinite future." (ISO 14721:2003)

In short: Well beyond the end of the project.

Organising long-term preservation

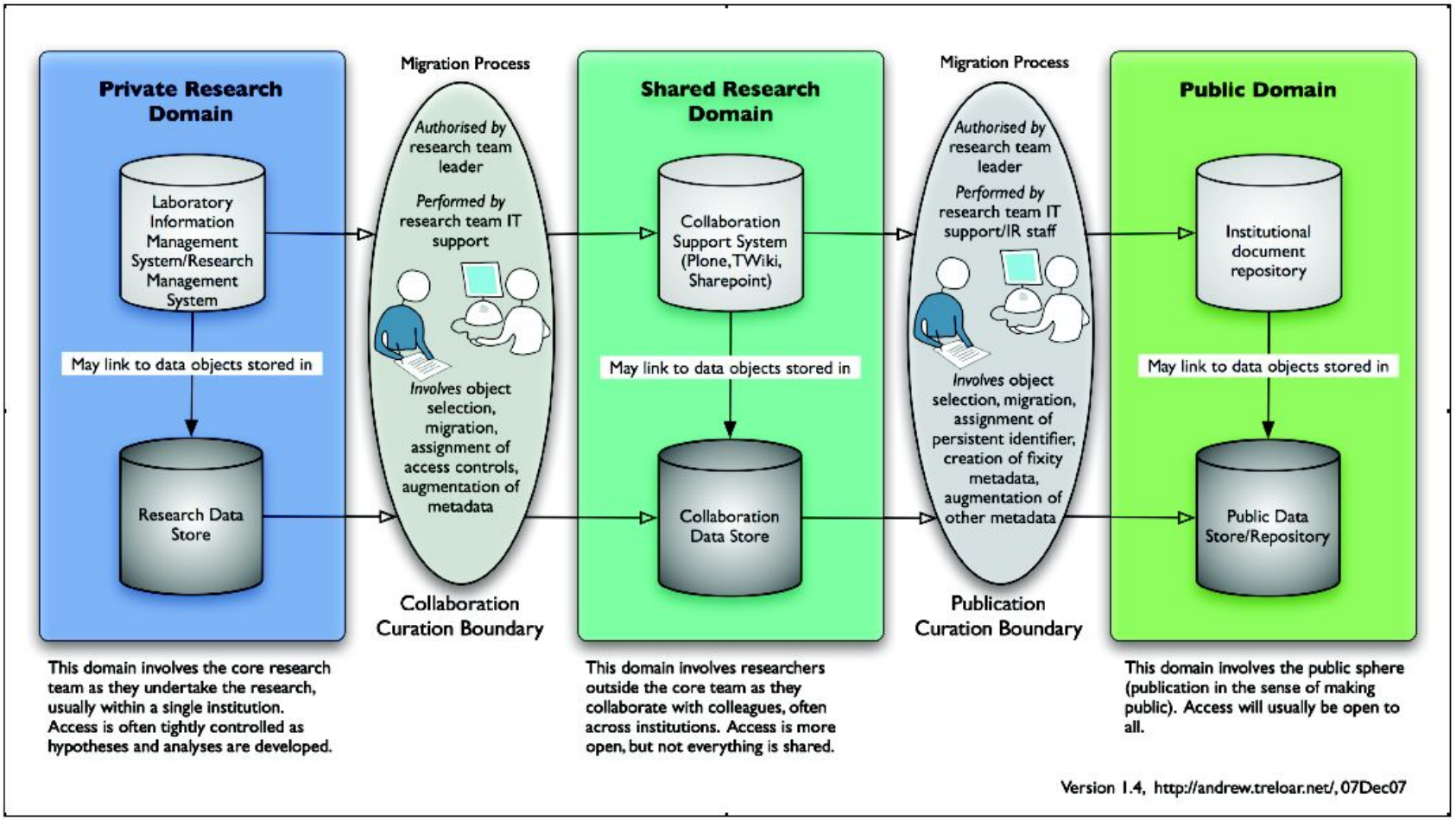
Data Curation

Data Curation Continuum

Object:	Less Metadata	⟷	More Metadata
	More Items	⟷	Fewer Items
	Larger Objects	⟷	Smaller Objects
	Objects continually updated	⟷	Objects static/derived snapshots
Management:	Researcher Manages	⟷	Organisation Manages
	Less Preservation	⟷	More Preservation
Access:	Mostly Closed Access	⟷	Mostly Open Access
	Less Exposure	⟷	More Exposure

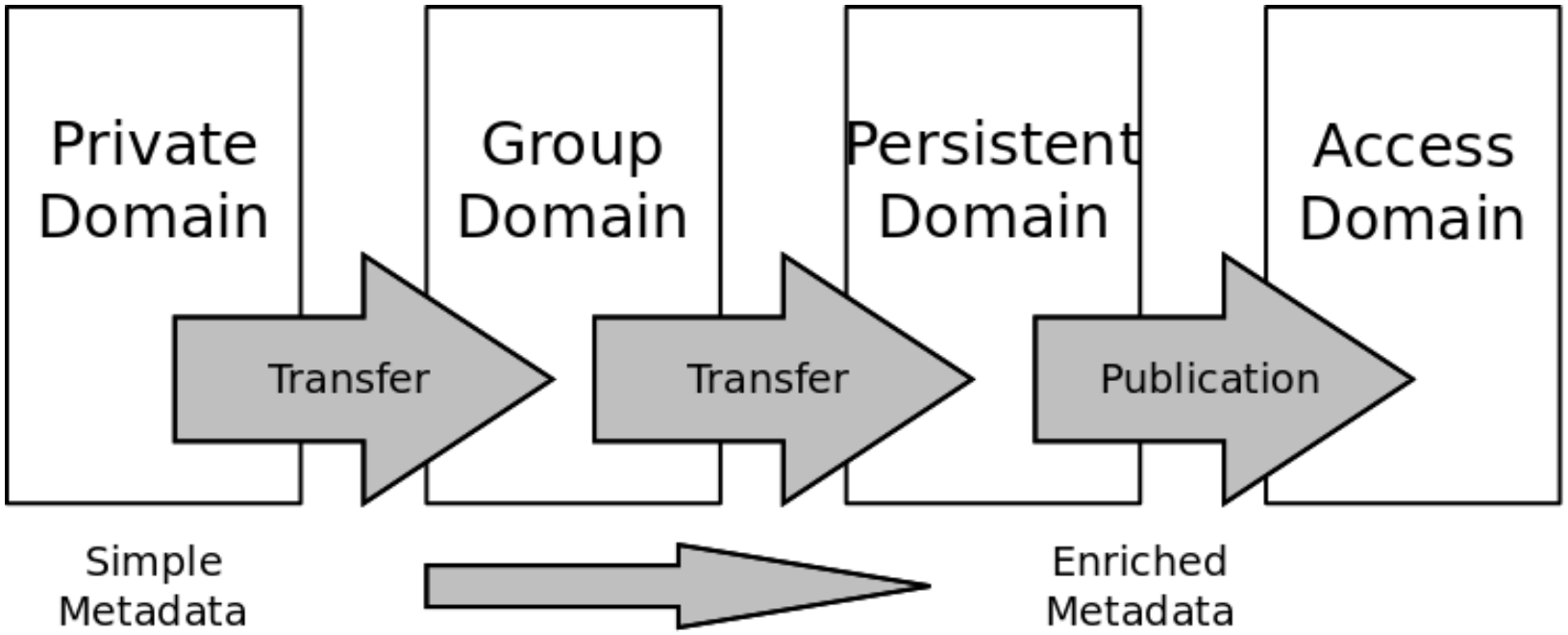
Treloar et al., 2007

Structuring the Continuum



Treloar et al., 2008

Structuring the Continuum



Roles and Responsibilities

Long-term data preservation needs a look at the persons involved and their roles in the processes.

- Scientists
  - Domain experts (+)
  - Not trained in data management (-)
  - No reward for success (-)
  - E-13/14 Salary
- Information specialists
  - Little domain knowledge (-)
  - Trained in data management (+)
  - Reward for success (+)
  - E-8...11 Salary

Risk Structures

- And what if I do nothing?

Common risks:

- Running out of time
- Running out of funds
- More data than expected

Criteria for trustworthiness of digital archives

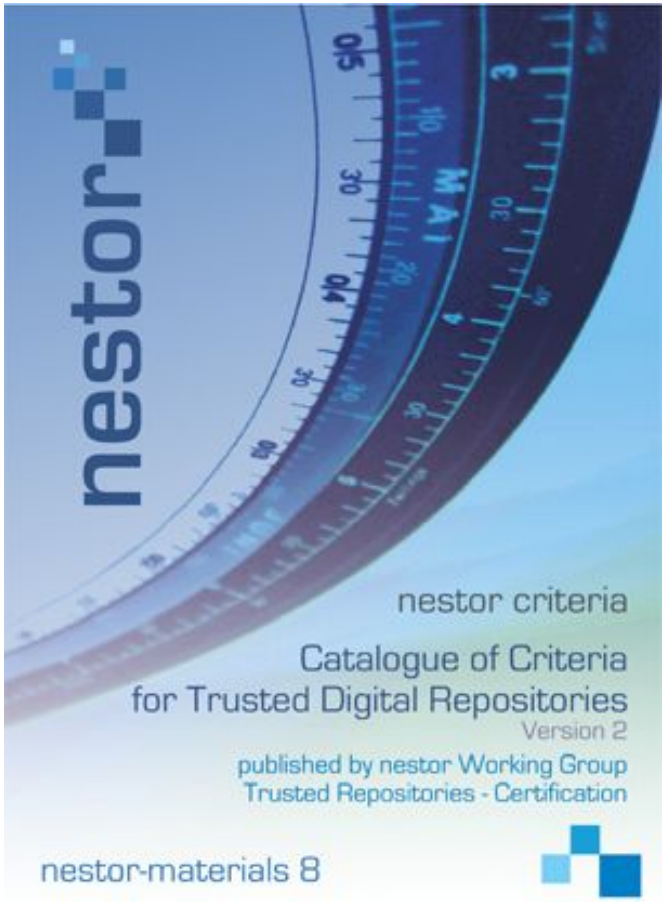
Trusted archives

Audit and Certification

- Whom do we entrust our digital heritage?
- How can we organise long-term data preservation?
- How do we organise a distributed network of digital data archives?

Criteria catalogues for the audit and certification for the trustworthiness of digital archives can be used to set criteria for members of the network and assess their compliance to rules given by the network.

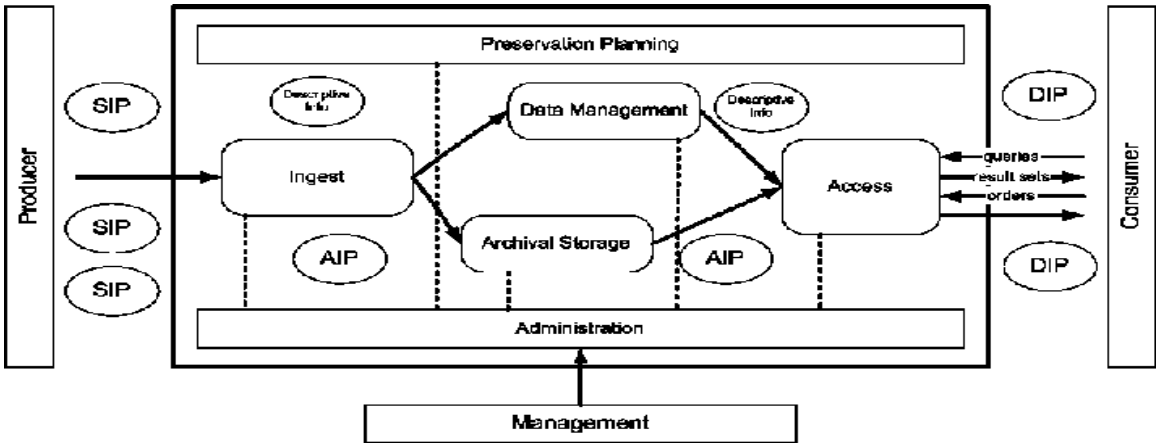
nestor catalogue of criteria for trusted digital repositories



- The nestor working group produced a catalogue of criteria for trusted digital repositories.
- The catalogue of criteria is in the process of being transformed into a DIN standard (DIN 31644) in preparation for international standardisation (ISO 16363).
- The criteria follow the philosophy of ISO 9000.

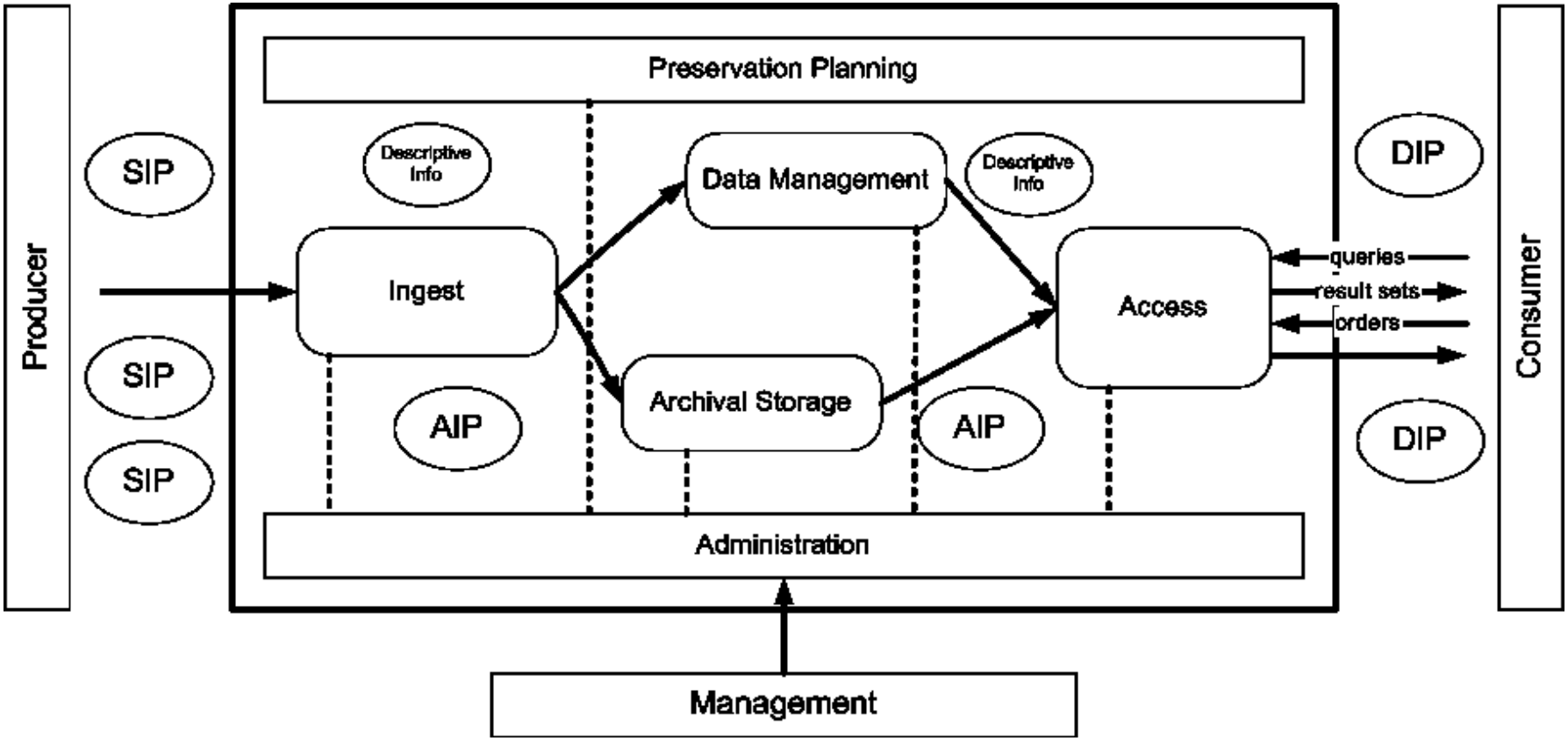
Principles of the criteria catalogue

The catalogue follows the reference model for open archival information systems (OAIS, ISO 14721: 2003)



The criteria of the catalogue have been kept abstract to allow transfer to other fields.

Open Archival Information Systems Reference Model



Structure of the Criteria Catalogue

Organisational Framework	Dealing with Objects	Infrastructure and Security
Documentation	Documentation	Documentation
Transparency	Transparency	Transparency
Adequacy	Adequacy	Adequacy
Measurability	Measurability	Measurability

Running an archive

Cost structures

Life Cycle Costs

$$L_T = Aq + I_T + M_T + Ac_T + S_T + P_T$$

L is the complete lifecycle cost over time 0 to **T**. Other categories are

Aq = Acquisition,

I = Ingest,

M = Metadata,

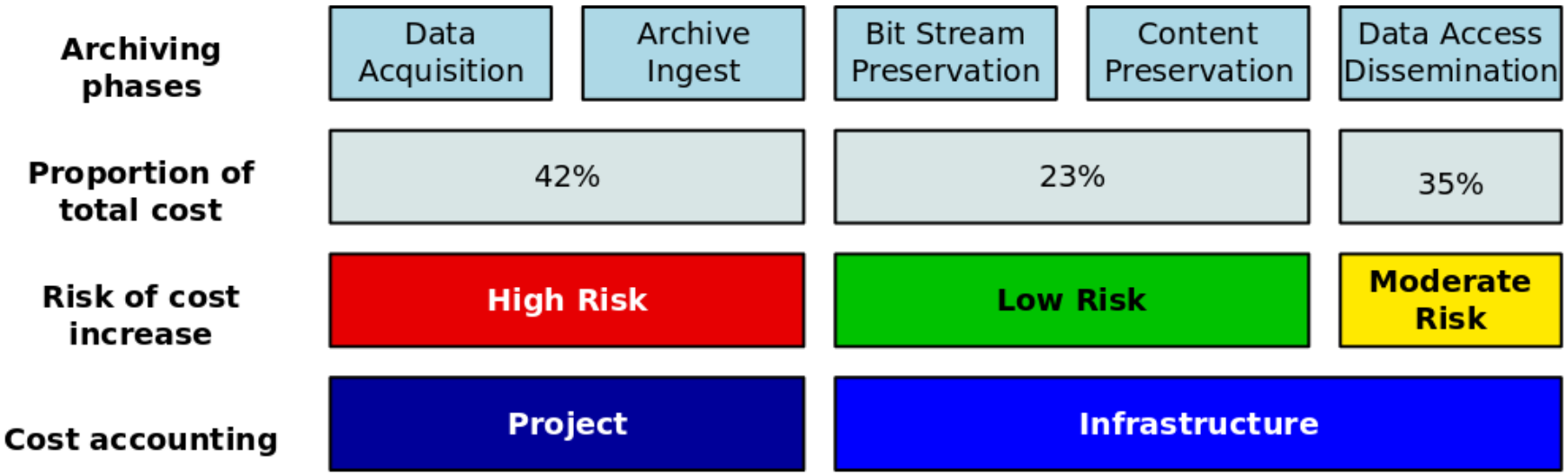
Ac = Access,

S = Storage,

P = Preservation

Source: LIFE2 Report, JISC

Cost Structures



Impact of Fixed Costs

- The costs of long-term data curation/preservation are dominated by fixed costs that do not vary with the size of the collections;
- Staff are the major cost component overall and there is a minimum base-level of staff cover, skills and equipment required for any service;
- Activities characterised by significant fixed costs can reduce the per-unit cost of long-term preservation by leveraging economies of scale.

(KRDS2, pp.32-34, 79-80)

Long-term finance

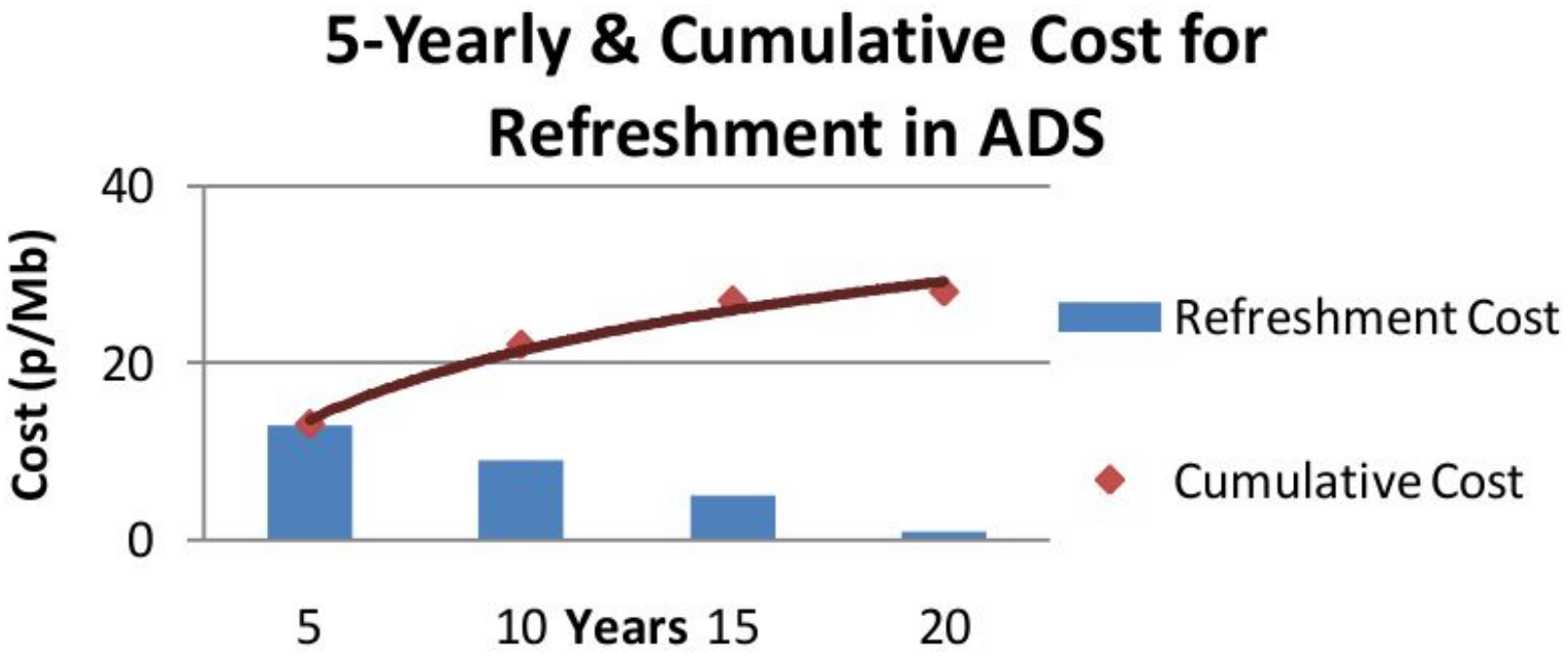
Main options:

- Transfer Model – Cross-financing of preservation actions from current projects (“state pension”)
- Insurance Model – User pays an annual premium to the archive
- Memory Institution – External funding of long term preservation as an infrastructure

- Additional option: Shift cost burden of ingest cost to projects

Most funding agencies will fund data curation and archive ingest in the context of a project.

Long-Term Cost Decline



(KRDS1, pp.4-6)

Service Level Agreements

- The cost of long-term preservation of research data strongly depends on factors determined by services rendered:
  - Archival ingest processes (manual vs. automatic)
  - Archiving scenario (security, number of copies)
  - Access interfaces and speed (SAN, RAID, tape)

Service level agreements between archive and data producers set a framework of technology, organisation and cost.

Long-term preservation of research data

Summary

Summary

- Make use of economies of scale
- Costs of long-term data curation/preservation are dominated by fixed costs
- Data curation is associated with high proportion of staff costs
- Cost of ingest process and apparatus is much higher than cost of maintaining long-term archive
- Service level agreements between data producers and archive define the responsibility for long-term data curation and a business model for long-term funding of archiving activities.

Source: Beagrie, 2010, KRDS Fact Sheet

Beucke, 2010

Thank you!

References

- Ayris, P., R. Davies, R. McLeod, R. Miao, H. Shenton, und P. Wheatley (2008), The LIFE2 final project report, Report, JISC Repositories and Preservation Programme, London, UK. [online] Available from: http://eprints.ucl.ac.uk/11758/
- Beagrie, N., J. Chruszcz, und B. F. Lavoie (2008), Keeping research data safe, JISC Reports, Joint Information Systems Committee (JISC), London, United Kingdom. [online] Available from: http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx
- Beagrie, N., B. F. Lavoie, und M. Woollard (2010), Keeping research data safe 2, Joint Information Systems Committee (JISC), Bristol, UK. [online] Available from: http://www.jisc.ac.uk/publications/reports/2010/keepingresearchd
- Beagrie, N., B. F. Lavoie, J. Chruszcz, und M. Woollard (2010), Keeping Research Data Safe Factsheet, Charles Beagrie Ltd,