



Long-term Preservation of Digital Research Data

23th November 2009, Hamburg

Jens Ludwig

Goettingen State and University Library, R&D

ludwig@sub.uni-goettingen.de

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Remarks in Advance



Long-term preservation does not mean only storing!

Basically: What has to be done to ensure reuseability of digital objects in a different context (temporal, tech., subject-specific, org., ...)?

Overview



- The digital age and the relevance of long-term preservation
- Structure of the challenge and components of the „solutions“
- What is happening in HEP and in WissGrid?

Overview



- The digital age and the relevance of long-term preservation
- Structure of the challenge and components of the „solutions“
- What is happening in HEP and in WissGrid?

Analog Age



Classic memory institutions like

- libraries
- archives
- museums

have collected and preserved information since thousands of years.

Information carriers like stone, clay tablets, paper and microfilm are relatively durable, have low data density and few presuppositions.

Digital Age



In culture, science and administration nearly all information is produced digital:

- Publications
- Records of public administration and companies
- Audio, Video, Images, ...
- Software applications
- Email,
- Internet sites
- ...
- and of course: research data

And digital data is harder to preserve...

Why is preservation important?



- Fundamental liability of politics, administration, companies, ...
- Fundamental for progress and reflection: We have to stand on the shoulders of giants...
- Personal emotional value: pictures, emails, even sms, ...

Why is preservation important?



We are interested in the correspondence of famous people of the past. The famous people of the future have today a facebook site or write emails.

Not everything has to be preserved, but probably everywhere something.

Why is preservation of research data important?



- Good scientific practice and prevention of fraud
- Later, improved analysis of old data
- Not all data is reproducible (climate measurements, astronomy or e.g. for cost reasons)
- „Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly.“

Nature 461, 145 (10 September 2009),
doi:10.1038/461145a

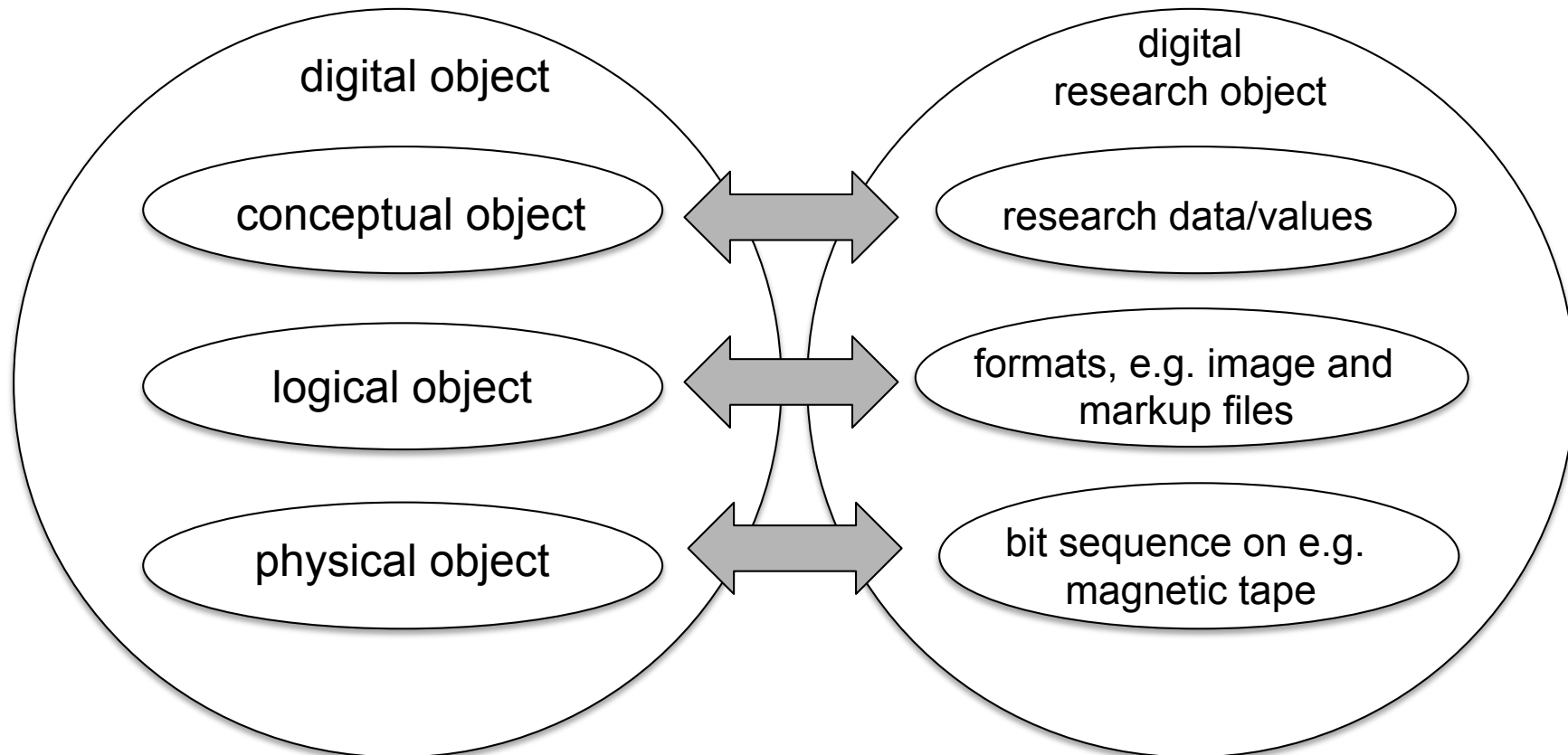


Overview



- The digital age and the relevance of long-term preservation
- Structure of the challenge and components of the „solutions“
- What is happening in HEP and in WissGrid?

Model of Digital Objects



derived from Thibodeau: Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, 2002. <http://www.clir.org/pubs/reports/pub107/thibodeau.html>



Dangers for Digital Objects

- on the physical level
 - decay of media: complete failure, read errors, bit rot
 - obsolescence of hardware: missing drives
- on the logical layer:
 - obsolescence of software: application no longer available or incompatibility
 - obsolescence of formats: wordstar file format
- on the conceptual layer:
 - changes in the background and context knowledge: change of scientific concepts, lacking documentation/ metadata, ...

And “Normal” Dangers

- human error
- technical failures
- disasters
- security gaps, sabotage
- ...

But also

- lack of funding
- lack of qualified staff
- lack of rights (copyright!)
- lack of awareness
- ...





Sometimes it looks even similar to
decay in the paper world:

reire ibe ar e ngen on ponen ia gr ppen n Opera oren i g a en ernen

ür rtin

on or t L ptin in ie e e

ie iri o erna che ar e ng heorie ie cher ponen ia gr ppen ge a e
e , ie irre iben ni ären ar e ngen ie er r ppen a be on er ein ache n
ber ich iche ei e be chreiben. e e erar ige ar e ng einer o chen r ppe
i ono ia , .h. ie i on eine Chara er einer abge ch o enen a enhängen
en n ergr ppe P in ier . Sie ä ich a i in be on ere i i ber chen a
 $L^2(P)$ e a orra e P rea i ieren, er einer ei er öge anoni cher o
or ina en i eine e i i chen \mathbb{R}^m i en ifi ier er en ann, erar , a a ebe g e
che a a \mathbb{R}^m rea i in arian be gich er ir ng on i . ie in egrier e ar e
ng on $L()$, eben a i be eichne , ie er r $L()$ ann rch erne efinier e
Opera oren (), eren erne π a n ionen a $\mathbb{R}^m \times \mathbb{R}^m$ ie ich ing är ein
önnen. a ie eich chön e e a in ie e rei gi r ni po en e r ppen.
ni po en , o gehen er chie ene Para e ri ier ngen on rch anoni che o

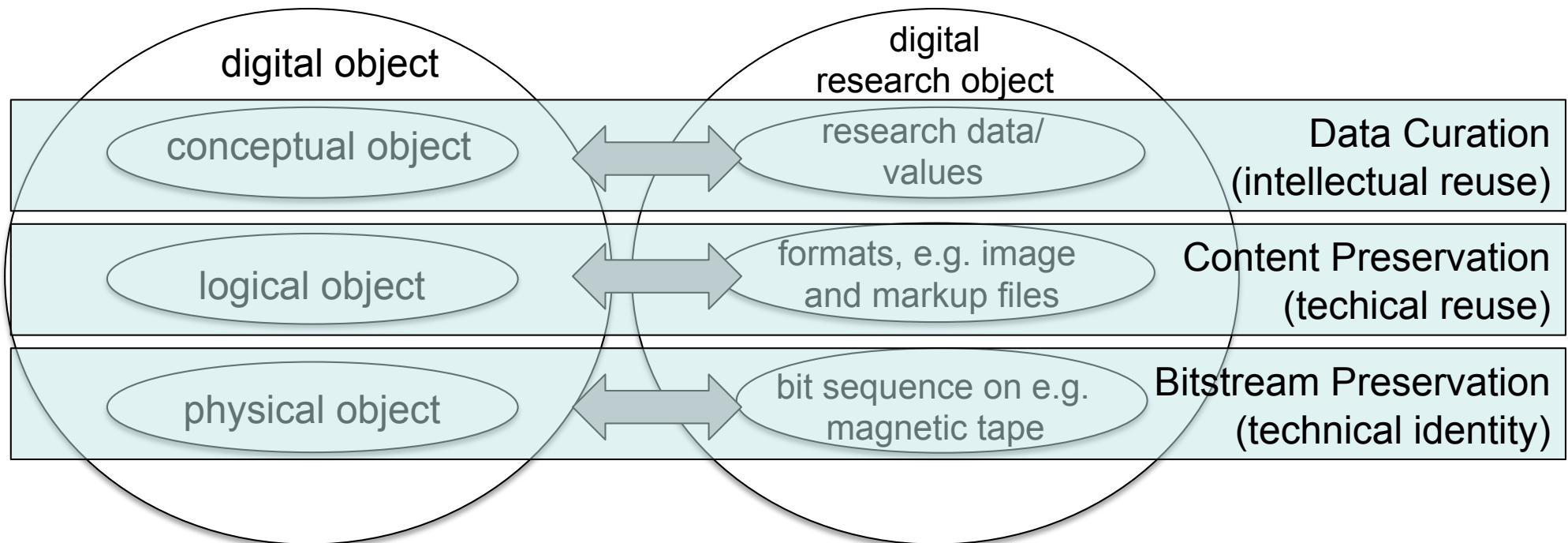
Terminology for Preservation



No precise terminology yet

Overall: long-term preservation (LTP)

In detail:



Tasks (from the perspective of the Object)



- physical-technical object:
 - ensuring the integrity of the bitstream
 - sufficient copies with different locations, institutions and technologies, regular integrity checks
- logical-technical object:
 - ensuring the technical validity and usability
 - persistent identification, technical quality control, preservation actions like format conversion/migration or providing emulators, ...
- intellectual object:
 - ensuring intellectual interpretability and usability
 - planing of data model and metadata, versioning of objects, maintaining access rights, appraisal, collection building, intellectual enrichment/linking, ...

Main Solution Strategies



- Bitstream Preservation:
 - Replication, integrity checks, ...
- Content Preservation:
 - Migration: adapting object to new tech. environment (but can introduce errors, authenticity difficult to judge)
 - Emulation: adapting tech. environment to old object (but you have to use the old and outdated environment)
- Data Curation
 - Provide metadata and context information

Implementation of Strategies



LTP is an ongoing task and can not be solved by technology alone (like in the analog world).

Accelerated obsolescence is equivalent to accelerated progress .

For large amounts of data these strategies can not be implemented

- without technical infrastructure or
- without a political und organisational framework

Examples for International Developments



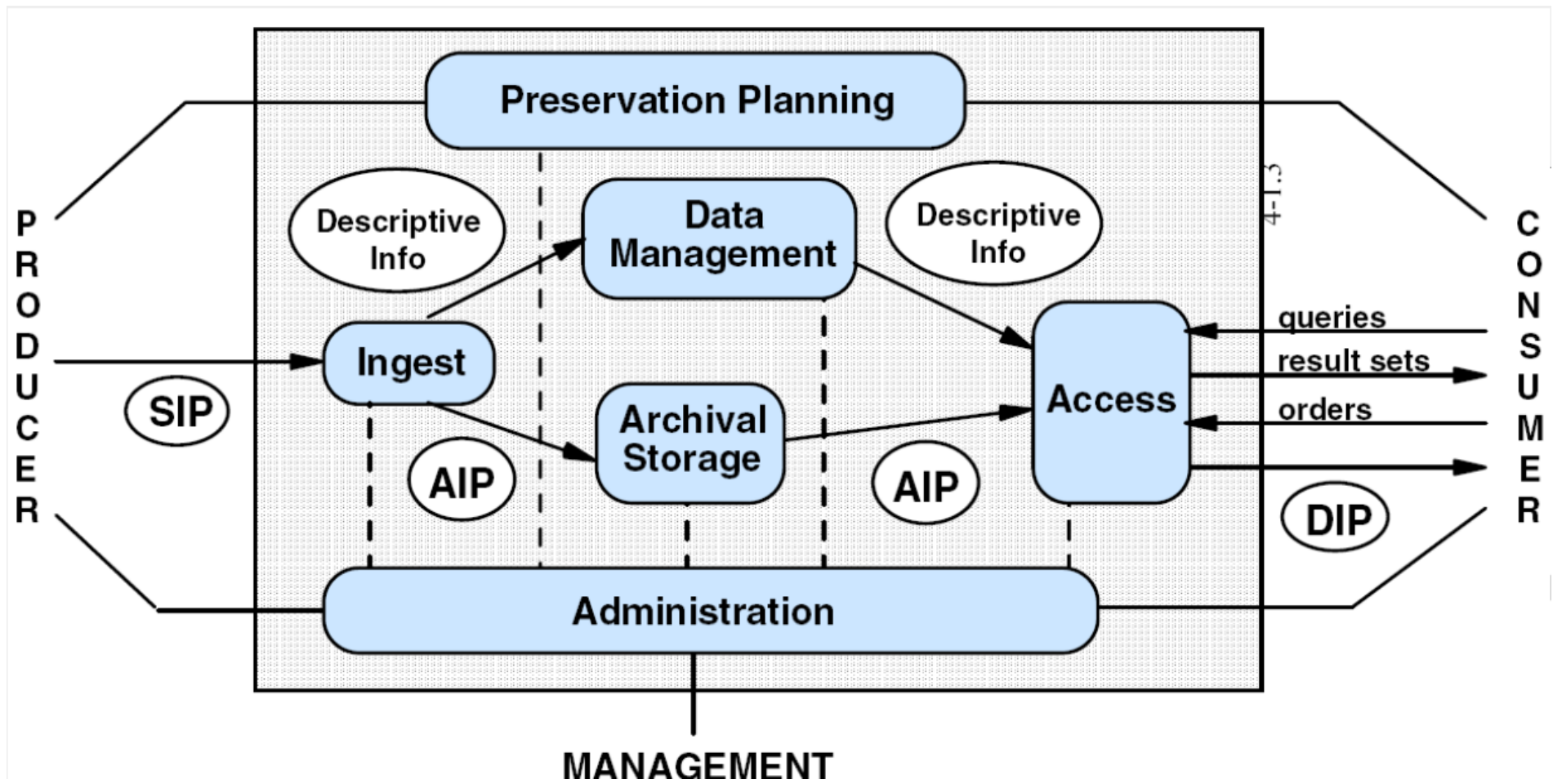
- Standards for long-term preservation
 - archive reference model (OAIS)
 - metadata sets (e.g. PREMIS)
 - criteria catalogues (e.g. TRAC)
- Systems, tools
 - JHOVE (framework for validation of file formats and technical metadata extraction)
 - international file format registries (PRONOM, The National Archives of the UK)
- Organisations
 - Research data centers and infrastructure (DataNet/US, ANDS/AUS, UKRDS/UK)

OAIS Reference Model

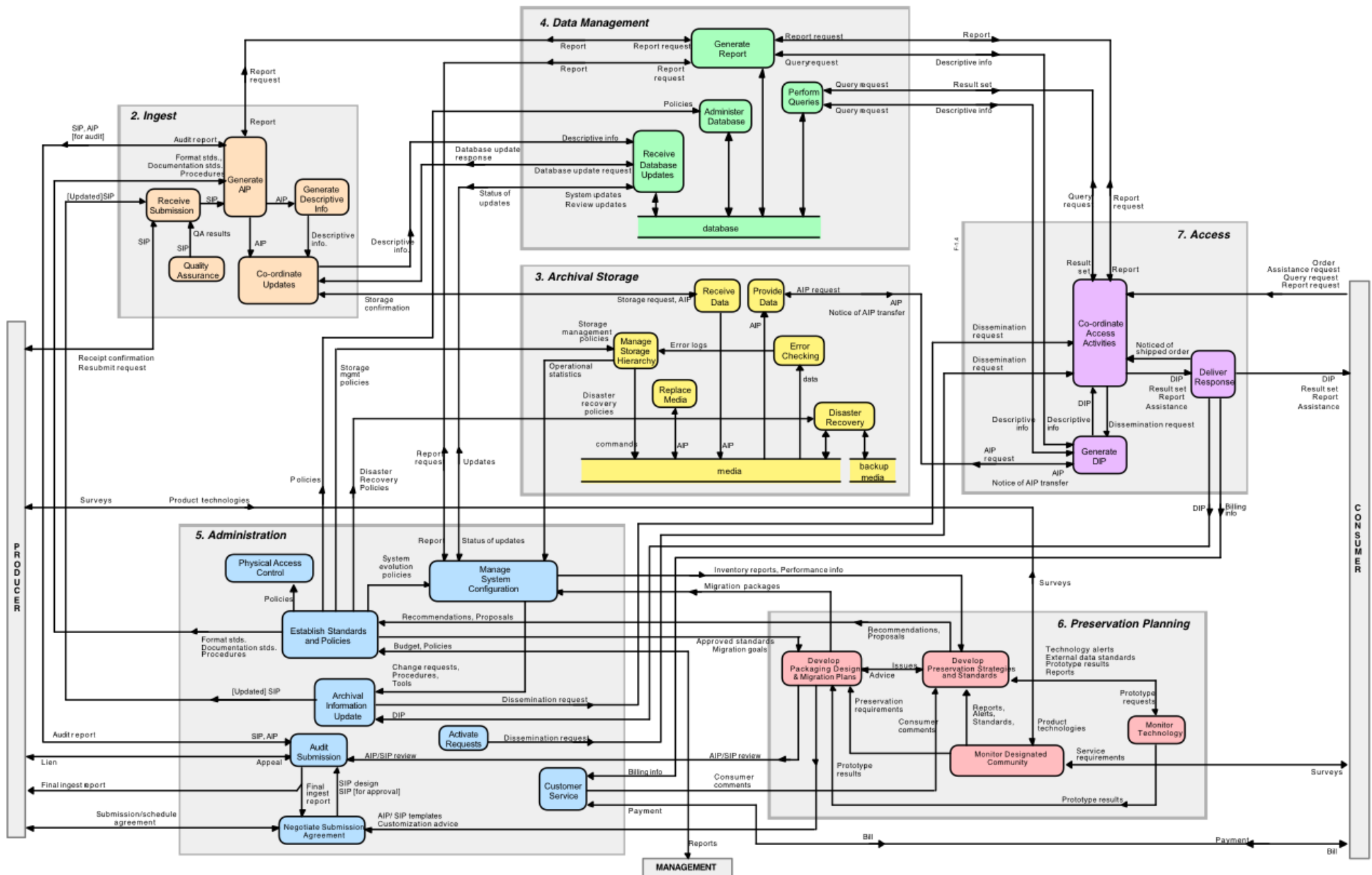


- Originated in the aerospace sector
- Defines functionalities and information types
- No prescriptions for implementation
- main benefit: basic common terminology
- Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System (OAIS). CCSDS 2002.
- <http://public.ccsds.org/publications/archive/650x0b1>

OAIS overview



OAIS in Detail (just to impress you)



Criteria Catalogues



- Problem: success of LTP can be determined only in hindsight
- criteria catalogues define instead what trustworthiness means for an archive
- typical types of criteria
 - organisational frame (e.g. financing)
 - treatment of objects (e.g. integrity checks)
 - infrastructure and security
- e.g. TRAC <http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying>
- e.g. <http://datasealofapproval.org/>

Metadata



Necessary for management and reuse of data

- „Representation Information“
 - Metadata about technical implementation and system requirements
 - Metadata about the semantic of the data, for intellectual reuse
- Provenance metadata
- Descriptive metadata
- Fixity, Context, Reference, ...
- e.g. PREMIS: <http://www.loc.gov/standards/premis/>

Research Data Centers and Infrastructure



- DataNet initiative in the USA
 - 100 Mil. Dollar in five yeas (just started)
 - create five discipline specific data centers
- ANDS in Australia
 - 50 Mil. Australian Dollars in ten years
 - create common infrastructure (for e.g. persistent identification similiar to DOI for research data)
- UKRDS in UK
 - pilot project for creating a shared research data service
- And of course already established ones like the World Data Centers (primarily earth sciences)

Overview



- The digital age and the relevance of long-term preservation
- Structure of the challenge and components of the „solutions“
- What is happening in HEP and in WissGrid?

HEP overview



- Rolf-Dieter Heuer (CERN and former DESY):
Need for open data formats and explication of
implicit knowledge for reuse

<http://www.computerweekly.com/Articles/2008/08/06/231762/in-search-of-the-big-bang.htm>

- CERN is member of PARSE.Insight
 - EU project which develops a roadmap and
recommendations for a european preservation
infrastructure
- ICFA Study Group on Data Preservation and
Long Term Analysis in HEP

PARSE.Insight



Holznera, Igo-Kemenes, Melea: „First results from the PARSE.Insight project: HEP survey on data preservation, re-use and (open) access“

<http://arxiv.org/pdf/0906.0485>

- about 1200 responses
- 69% see preservation as very important or crucial
- only 16% think that their institution has the necessary resources
- about 40% suspect that relevant data has already been lost (not in that article)

ICFA Group on Data Pres. and Long Term Analysis in HEP



- <http://dphep.org/>
- Started with a series of workshops this year (January/DESY, May/SLAC, Dec./CERN, ...)
- The different experiments report their experiences with Data Preservation
 - e.g. migration of data formats
 - e.g. keeping software alive
- Working groups dealing with different aspects

Reuse Story from the 1st Workshop



- Siegfried Bethke (Max-Planck-Institut for physics), „Experience from re-analysis of PETRA (and LEP) Data“, <http://indico.cern.ch/getFile.py/access?contribId=11&sessionId=3&resId=0&materialId=slides&confId=42722>
- The JADE Experiment at the PETRA e^+e^- storage ring @ DESY, operation time: 1978 – 1986
- It was possible to get new results from old data with new insights although ...

Reuse Story from the 1st Workshop



- A lot of know-how was stored in private accounts and lost
- For reuse it was necessary to:
 - copy data from old tapes to new media
 - convert some old data formats to new ones
 - rewrite and adapt JADE software to new platforms
 - track missing software and data all over the planet
- *„Jan Olsson, when cleaning up his office in ~1997, found an old ASCII-printout of the JADE luminosity file. Unfortunately, it was printed on green recycling paper - not suitable for scanning and OCR-ing. A secretary at Aachen re-typed it within 4 weeks. A checksum routine found (and recovered) only 4 typos.“*

LTP in HEP from an Outsider Perspective



- It should be easier for HEP than for other disciplines
 - because it is Big Science
 - limited data sources
 - very organised
 - a „lot“ of money
 - high technology affinity
- But:
 - Infrastructure and IT is probably highly proprietary and uncommon
 - high responsibility and need to justify: don't let that JADE thing happen again?

So, what is this Logo about? ->



- WissGrid is a new project of the five scientific grid projects (HEPGrid, C3-Grid, MediGrid, AstroGrid-D, TextGrid) in D-Grid
- Basic aims: sustainability and propagation of Grid for academic users in Germany
 - operational model, grid user representation and independent financing
 - consulting of communities interested in grid tech.
 - and: provide basic tools and know-how for preservation of research data with grid technology

LTP of Research Data in WissGrid



Reasons for LTP with grid technology

- Grid provides abstraction from underlying technology
- Grid resources beneficial for storage of huge amounts of data and e.g. task like file format conversion
- LTP is a task for the whole life-cycle from production to archiving to reuse. And the grid is part of this life-cycle.

What will the LTP-WP of WissGrid do?



- pre-configured systems
 - research data repository
 - provenance registry
- frameworks for
 - metadata extraction from files
 - validation of file formats
 - conversion of file formats
- guidelines for LTP of research data

What will the LTP-WP of WissGrid do?



WissGrid =

Community =

D-Grid/Infrastructure provider =

