



# ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

## ARCHIVING AND PRESERVATION IN THE WIDER WORLD

**Use Case Scoping Session**

DESY, 20 February 2019

[Jamie.Shiers@cern.ch](mailto:Jamie.Shiers@cern.ch)

# OVERVIEW – THE LTDP SCENE

- ☁ Several International Conferences & Workshops
- ☁ Numerous EU (and other) projects
- ☁ National and International Collaborations
- ☁ A well established Reference Model (OAIS)
- ☁ A key component of “modern” data management and stewardship aka “FAIR data management”
- ☁ Covers ~all disciplines from the “long tail” to ESFRI-like projects and beyond: zero bytes to 10s-100s of PB (“zero bytes to exabytes”)

# CONFERENCES & WORKSHOPS

- ☁ **iPRES** – held annually, covering 4 continents – since 2004
  - ☁ This year in Amsterdam: Sep 16 – 20
  - ☁ iPRES 2016 (CH): “CERN Services for LTDP” [ TDR + Invenio-based s/w + CVMFS ]
  - ☁ <https://ipres-conference.org/>
- ☁ **PV** – started by the “space community” – since 2002 (**P**reservation & Adding **V**alue) (PV2020 will be at CERN May 12 – 14)
- ☁ **iDCC** (International Digital Curation Conference) -  
<http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc>
- ☁ **PASIG** – including an active mailing list (Preservation & Archiving SIG):  
<https://preservationandarchivingsig.org/>
- ☁ The Research Data Alliance (RDA): bi-annual meetings with numerous groups on LTDP, DMPs etc.  
<https://rd-alliance.org/node>
- ☁ **WE ARE NOT ALONE!**



# (SOME) PROJECTS

- ☁ InterPARES (Canada): <https://interparestrust.org/>
- ☁ LOCKSS (US): <https://www.lockss.org/>
- ☁ APARSEN (EU): <http://www.alliancepermanentaccess.org/>
  - ☁ Established “hierarchy” of certification procedures
- ☁ 4C (EU): <http://www.4cproject.eu/about-us/>
  - ☁ Costs of Curation – metrics relevant for ARCHIVER
- ☁ RDA (Worldwide, including EU):
  - ☁ DSA + WDS → CoreTrustSeal; Active DMPs; Domain Repositories; Reproducibility; ...
  - ☁ 13<sup>th</sup> plenary just before OMC@CERN, Philadelphia (2-4 April)
- ☁ **At one stage the EU claimed to have invested EUR100M in data preservation projects! [ Presumably more now ]**

# COALITIONS

 Focus on Digital Preservation Coalition (DPC):  
<https://www.dpconline.org/>

 National Coalitions exist, e.g. in D, NL, ...

 The UK also has the Digital Curation Centre... (DMPs)

1. *Could the DPC provide a representative on ARCHIVER's External Advisory Board?*
2. *Could the DPC review the tender technical specifications and provide feedback?*
3. *Could the DPC promote the tender to all their commercial supporters as a way of encouraging them to participate in the open market consultation and submit bids?*
4. *Could the DPC review the total cost of ownership studies to be produced by the suppliers?*
5. *Could the DPC run a training event for ARCHIVER tailored to the supported use-cases?*
6. *Could the DPC review the service specific training material produced by the suppliers?*

 Positive answers to all of these questions – in progress

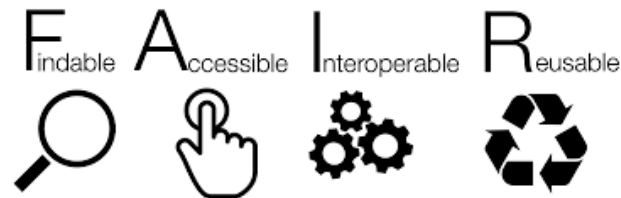


# CONVENTIONAL WISDOM ON LTDP

- ☁ ***A Trustworthy Digital Repository (TDR) is a sine qua non of LTDP*** – Ingrid Dillo, DANS & former RDA SG
- ☁ TDR = certified according to an agreed, OAIS-based, certification procedure
  - ☁ e.g. DSA, DIN, WDS, CoreTrustSeal, ISO 16363
  - ☁ WDS(57), DSA(37), WDS+DSA(1), CTS(48), ISO (2+1?)
- ☁ Attempts to combine (reconcile?) TDRs and FAIR, e.g. DANS “5-star scale”
  - ☁ **Quality (trustworthiness) of data repositories - TDR principles**
  - ☁ **Quality (fitness for use) of datasets - FAIR principles**
- ☁ **FAIR is still evolving, e.g. workshop on FAIR s/w in NL**
- ☁ Final report from EU Expert Group: [Turning fair into reality.](#)

# Assessing the FAIRness of Datasets in Trustworthy Digital Repositories: a 5 star scale

Peter Doorn, Director DANS  
Ingrid Dillo, Deputy Director DANS




## [2nd DPHEP Collaboration Workshop](#)

CERN, Geneva, 13 March 2017



@pkdoorn @dansknaw

# H2020 DMP – General Definition

 As part of making research data Findable, Accessible, Interoperable and Re-usable (FAIR), a DMP should include information on:

1. the handling of research data during and after the end of the project
2. what data will be collected, processed and/or generated
3. which methodology and standards will be applied
4. whether data will be shared/made open access and
5. **how data will be curated and preserved (including after the end of the project).**



# OAIS: Key Concepts & Definitions

- *Open* = developed in an open public forum
- *Archival Information System* = “an organization of people and systems that has accepted the responsibility to
  - **preserve information** and
  - **make it available** for a
  - **Designated Community**”

# OAIS: Mandatory Responsibilities

- Accept content
- Obtain control (including necessary IP rights)
- Define user community
- Ensure that the preserved information is **independently understandable** to the user community
- Follow documented procedures to
  - **Preserve information against reasonable contingencies**
  - Enable dissemination of **authenticated copies**
- Make preserved information available

## **OAIS: Long-Term**

**A period long enough to raise concern about the effect of changing technologies, including support for new media and data formats, and of a changing user community. (OAIS, RLG-OCLC)**

# Other Key Concepts

- **Authenticity**
- **Trust**
- **Sustainability**

# Authenticity

- **Authentic = Genuine = Bona Fide**
- **The trustworthiness of a record as a record: i.e., the quality of a record that is what it purports to be and that it free from tampering or corruption (InterPARES)**
- **Property that a digital object is what it purports to be. (PREMIS)**

## CONFERENCES & WORKSHOPS

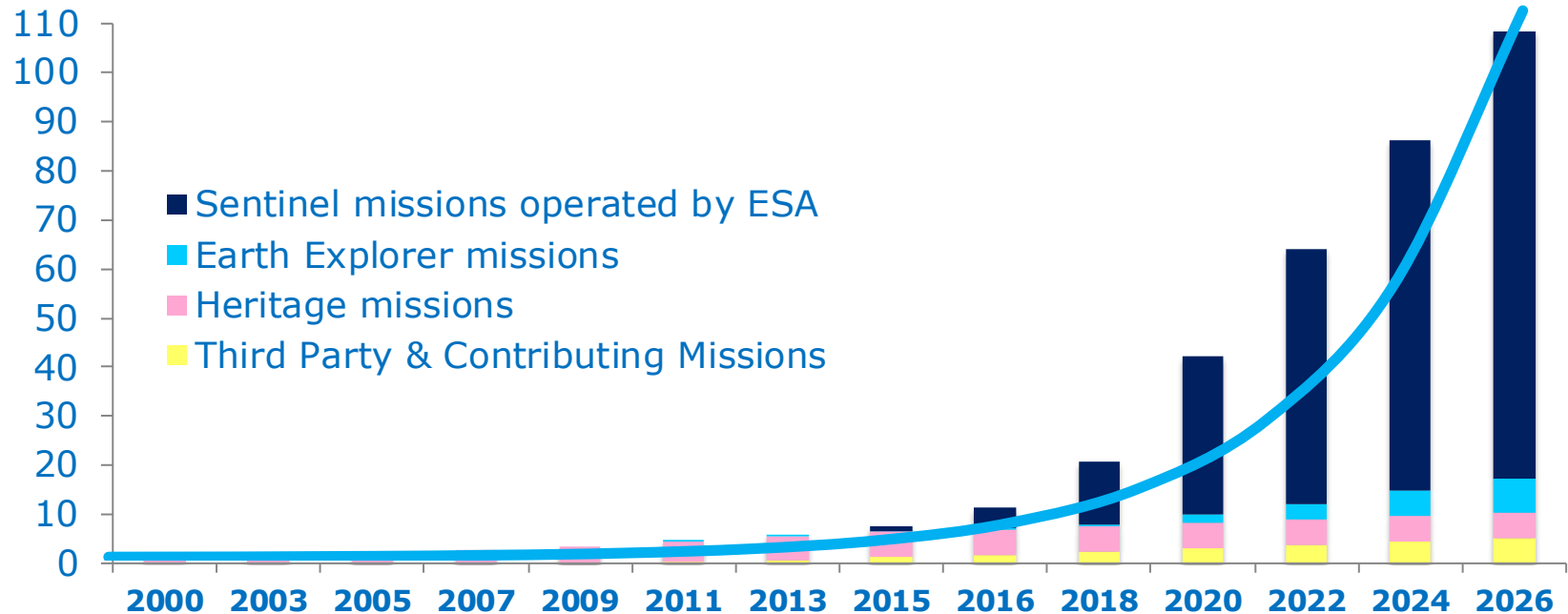
- ☁ iPRES 2019: A draft of a poster exists. Presumably it should be submitted on behalf of the collaboration. Deadline is 18 March 2019
- ☁ PV2020: should be an opportunity to show-case ARCHIVER results(?)
- ☁ I am still looking for help (program committee, LOC) as well as proposals for sessions (good topics + speakers)
- ☁ Plan conference flyer by iPRES 2019
- ☁ Workshop on Sustainable Software Sustainability 19, 24-26th of April, NL: invited talk on LEP-era experience (how badly we did things then and what have we learned).
- ☁ **WE ARE NOT ALONE!**



1. ESA specifies requirements for the archive service and hardware implementation inter alia:
  - a. Archive availability;
  - b. Extended service outage;
  - c. Data Loss events;
  - d. Reporting.
2. Service Level Agreement and Key Performance Indicators.
3. Incentives and penalties.
4. Each Centre provides its own archive infrastructure.
5. Data does not belong to Centres even if the infrastructure does.

## Big Data Revolution

ESA EO Data Archive, in Petabyte



Slide 13



European Space Agency

European Space Agency

# We need to think of the interface(s)



esa

## General Concept

esa

- Data Archiving procured as a Service.
  - No hardware for Data Archiving provided by ESA;
  - Industrial partners free to propose the hardware solution;
  - Interface (Interface Control Document) specified by ESA as interface between PDGS systems generating the data and the LTA - for archive and retrieval;
  - Technical characteristics required:
    - Daily I/O capacity: 30 TB (TeraByte)
    - Maximum single file size: 200GB (GigaByte)
    - Maximum number of files per day: 10000
    - Minimum sustained LTA interface transfer rate: 220 MB/s
  - Service Level Agreement and Key Performance Indicators to measure performance

Slide 6

opernicus



# Authenticity

- **Authenticity (traditional & digital) derives from...**
  - **Source**
  - **Chain of custody**
  - **Processing history**
  - **Fixity**
  - **Trust**

*Maintaining and disseminating authentic information is a primary mission for digital preservation systems.*

# Trust & Trustworthiness

- **As early as 1996, Trust in repositories was sited as a prerequisite for digital preservation**
  - **Task Force on Archiving of Digital Information**
- **Perception of competence, security, long-term commitment is necessary from...**
  - **Producers**
  - **Funders**
  - **Consumers**

# Trust & Trustworthiness

- Trust is granted by a third party to a repository
- *Trustworthiness* is demonstrated by adherence to four principles (nestor, DCC)
  1. Documentation
  2. Transparency
  3. Adequacy
  4. Measurability
- Audits help establish Trustworthiness



# Sustainability

- Long-term preservation, by definition, requires management of information over generations of change in...
  - Technology
  - Users & expectations
  - Staffing
  - Economic conditions
- **Preservation is a journey, not a destination**

# Sustainability

- **Technological Strategies**
  - **Simplicity, component-based, plan for migration, plan for replacement**
- **Organizational Strategies**
  - **Succession planning is a formal process of enabling handoffs across archives**
- **Economic Strategies**
  - **Contain costs, be selective, emphasize value of access**

# Sustainability

- **Make the case for use**
  - **Preservation = Long-Term Access**
  - **“In all cases, access to information tomorrow requires preservation actions taken today” (BRTF)**
- **Not all-or-nothing**
  - **Bit-level preservation + basic access may be more pragmatic than open-ended commitment to format migration / emulation.**
- **Not once-and-for-all**
  - **Rather than a 100 year commitment, think of a 10 year commitment, with an option to renew**

# Review: What Is Digital Preservation?

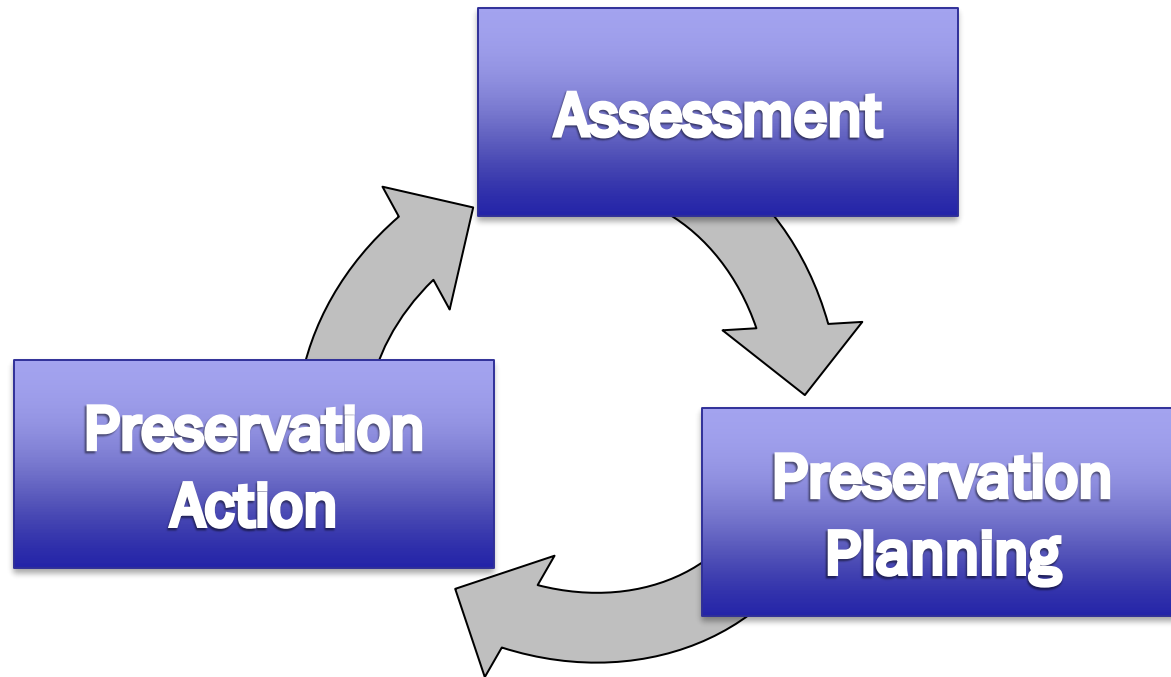
**Digital preservation is the series of strategies and actions taken to promote the availability and usability of authentic digital information over time.**

# Review: Digital Preservation Approaches

- Replication
- Migration
  - Format
  - Media
  - Technology
- Emulation
  - Hardware
  - Software
- Encapsulation
- Redundancy and heterogeneity
  - Technology
  - Location
  - Organization
- Succession Planning

*Digital Preservation (aka Long Term Access)  
is realized through a series of relays over time.*

# Digital Preservation in Action: Plan, Do, Check





# Toolkit for Preservation Planning and Actions

<b>Tool Type</b>	<b>Purpose</b>
<b>Identification</b>	What kind of file is this bit stream?
<b>Characterization</b>	What are its important features?
<b>Bit Audit / Fixity</b>	Have any of its bits changed?
<b>Manipulation</b>	I'd like to modify/update/transform this digital object
<b>Wrapping</b>	I'd like to package this object for storage or transfer
<b>Transfer</b>	I'd like to move this digital object

<http://www.digitalpreservation.gov/tools/>

# Technology Implications

- **Minimize dependencies**
  - Encapsulate your metadata with your objects
- **Minimize correlated errors**
  - Embrace redundancy
  - Embrace diversity
- **Monolithic systems tend to serve poorly**
  - Complex, expensive, inflexible
  - Migration costs can capsize you
- **Keep it simple; have an exit plan for every component**