
CERN

International Collaboration for **Data Preservation** and **Long Term Analysis** in High Energy Physics

# DPHEP Update:

## Including: Strategy for & Status of ISO 16363 Certification of CERN

WLCG GDB, October 2017

Jamie.Shiers@cern.ch


Bits Decay: Do Something Today

30th November 2017

# DPHEP March Workshop

1. Provide an update on the changing (or changed) landscape;
   - e.g. FAIR data management (plans), reproducibility, sustainability of data repositories, an update on the status of OAIS and related "standards" and so forth.
2. Status reports of the services / developments in the area of LTDP and their outlook
   - These are now (largely) production services
3. Perform a site / experiment round-table to capture the current situation HEP-wide
   - Are there areas where we can improve collaboration?

# Typical EU H2020 Call Text

- *Research Infrastructures, such as the ones on **the ESFRI roadmap** and others, are characterised by the **very significant data volumes** they generate and handle.*
- *These data are of interest to **thousands** of researchers across scientific disciplines and to other potential users via **Open Access** policies.*

➢ ***Effective data preservation and open access for immediate and future sharing and re-use are a fundamental component of today's research infrastructures.***

# FAIR Data Principles

Expert Group on turning FAIR into reality

**TO BE FINDABLE:**

- F1. (meta)data are assigned a **globally unique and eternally persistent identifier.**
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

**TO BE ACCESSIBLE:**

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

**TO BE INTEROPERABLE:**

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

**TO BE RE-USABLE:**

- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.

# FAIR DMPs & TDRs

- *If we want to be able to **share data**, we need to store them in a **Trustworthy Digital Repository** (TDR).*
  - ***Data created and used by scientists should be managed, curated, and archived in such a way to preserve the initial investment in collecting them.***
  - *Researchers must be certain that data held in archives remain useful and meaningful into the future.*
  - *Funding authorities increasingly require continued access to data produced by the projects they fund, and have made this an important element in **Data Management Plans** (DMPs – H2020 Guidelines).*
  - *Indeed, some funders now stipulate that the data they fund must be deposited in a **trustworthy repository**.*

# How Has FAIR evolved in 2017?

- Increasingly, FAIR has been taken to include not just data + meta-data but also **software**
- What started as "source code" preservation has now evolved to **"running s/w and its environment"**
  - **Much better IMHO**
- But there is still a lot to define / do
  - **How is the data Findable?**
    - **Navigation?** Search? Is there an API? …
  - **How to implement this in a scalable & sustainable way**
    - E.g. how many PID / DOI lookups per unit time, for how long is the service "guaranteed", … **"eternally?"**
  - **How to implement cross project / discipline searches?**
- **I have heard claims that people have been doing this for 20 – 100(!!!) years**
  - **(These people clearly don't need any more project money)**

# OAIS Update: CCSDS/ISO process

- CCSDS-DAI (Data Archive and Ingest) Working Group develops and maintains standards'
  - DAI chair is David Giaretta
- CCSDS is the working arm of ISO TC 20/SC13
  - Standards reviewed and approved in CCSDS go through an ISO review (reviews may be simultaneous)
  - Reference Model for an Open Archival Information System (OAIS). Magenta Book. Issue 2. June 2012 (CCSDS 650.0-M-2) is identical to ISO 14721:2012
- CCSDS and ISO procedures are well defined
  - ISO process at http://www.iso.org/sites/directives/2016/consolidated/index.xhtml
  - CCSDS follows the ISO code of conduct to ensure it addresses consensus, transparency, openness, impartiality.
  - CCSDS process in https://public.ccsds.org/Pubs/A02x1y4c2.pdf

➢ **Require reviews/updates of standards every 5 years**

# Welcome to 5 year review of OAIS and ISO 16363

File a Suggested Change | Search | User Preferences | Documentation

Enter a suggested change # or some search terms | Quick Search | [?]

**Description of the review process:**

The open process used since 1995 to create and revise OAIS has contributed to its success.

In order to ensure the continued usefulness of OAIS any revision must remain backward compatible with regard to major terminology and concepts. Further, for consistency the general level of detail should not be changed nor should the standard be changed from a reference model to an implementation design. Archive implementation standards or implementation profiles or detailed archival process standards or protocols should be addressed, but not in this document. They would become separate standards and would be developed through separate CCSDS projects. A particular interest for the current OAIS update is to reduce ambiguities and to fill in any missing or weak concepts and to add useful terminology.

The CCSDS and ISO process is rather formal; the CCSDS process is documented in https://public.ccsds.org/Pubs/A02x1y4c2.pdf while that of ISO is available at http://www.iso.org/sites/directives/2016/consolidated/index.xhtml. CCSDS follows the ISO code of conduct to ensure it addresses consensus, transparency, openness, impartiality. The integrated set of CCSDS and ISO activities expected to be used for this review is available here and the detailed schedule, which will be updated from time to time, is here.

Although CCSDS is an organisation created and managed by the space agencies it welcomes participation from industry, academia and outside organizations.

The CCSDS-DAI working group is, among other things, in charge of coordinating the Five Year Review of OAIS that includes gathering and adjudicating all comments, creating the consensus revision, and coordinating the review and approval of the revised OAIS standard. The working group is open to all; to participate, one should join the MOIMS_DAI mailing list here. This mailing list is used to co-ordinate discussions and meetings. The website http://review.oais.info is being used to record, discuss and reach consensus on the suggested revisions; to contribute to the site one can create an account at http://review.oais.info/createaccount.cgi. Most work is carried out by email and this is also used to send out agenda and details for the weekly virtual meetings.

CCSDS management has approved the schedule to produce a revised draft of OAIS by the end of 2017. Of course this could be completed earlier but past experience suggests that to reach consensus takes time. The draft is then submitted for review by the management of CCSDS and TC20/SC13 of ISO who will circulate the draft to their national bodies and wide set of contacts around the world and coordinate the responses and return them to CCSDS-DAI. The draft will also be circulated widely by CCSDS-DAI itself. Ideally all the issues will have been addressed during the development of the draft. However further issues may be pointed out and these must be discussed and the issues resolved by reaching consensus. A second round of reviews may then be necessary. The schedule is to publish the new issue of OAIS by CCSDS and ISO by mid-2019.

This is a long and involved process but is the one which has been well proven by CCSDS and ISO and has contributed to the success of OAIS.

# LTDP: How do we measure progress / success?

- **Practice:** through Open Data releases
  - Can the data really be (re-)used by the Designated Community(ies)?
  - What are the support costs?
  - Is this sustainable?
- **Theory:** by applying state of the art "preservation principles"
  - Measured through ISO 16363 (self-) certification and associated policies and strategies
  - Participation in relevant working & interest groups

**One, without the other, is probably not enough. The two together should provide a pretty robust measurement...**

# ISO 16363 Certification

- There are a number of certification procedures but only one developed by a Scientific Community – ISO 16363
  - **All based on OAIS model: ISO 14721**

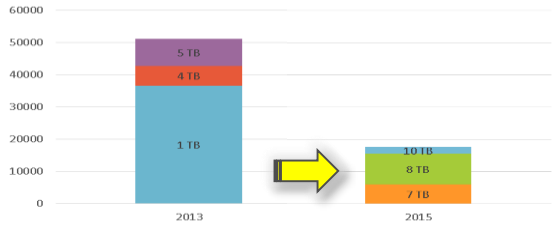
- An initial assessment of the main ISO 16363 criteria related to "bit preservation" was presented at the March 2017 DPHEP w/s

  - ***"Maybe CERN does bit preservation better than anyone else in the world"*** – David Giaretta

➢ Use this as a "template" for other criteria
  - **Bit preservation is only a small (but important) part of ISO 16363**
  - Not a guarantee against loss of a single bit in 200+PB, but a clear statement of what is done – including reporting – plus commitment to improve as technology + experience permit

# Bit Preservation: Steps Include

- Controlled media **lifecycle**
  - **Media kept for 2 max. 2 drive generations**
- Regular media **verification**
  - When tape written, filled, every 2 years…
- **Reducing** tape mounts
  - Reduces media wear-out & increases efficiency
- Data **Redundancy**
  - For "smaller" communities, a 2nd copy can be created: separate library in a different building (e.g. LEP – **3 copies at CERN**!)
- **Protecting** the physical link
  - Between disk caches and tape servers
- Protecting the **environment**
  - Dust sensors! (Don't let users touch tapes)




**Constant improvement: reduction in bit-loss rate: 5 x $10^{-16}$**

# Current Status

- ISO 16363 follows OAIS breakdown:
  3. **Organisational Infrastructure;**
  4. **Digital Object Management;**
  5. **Infrastructure and Security Risk Management.**

- Many of the elements in 3) and 5) covered by existing (and documented) CERN practices
  - **Some "weak" areas – being addressed – include disaster preparedness / recovery (together with EIROForum)**
  - **And we haven't really started to address 4) yet…**

➢ **Next step is "stage 1" external audit to high-light those areas requiring attention**
  - **May just be a question of documentation, e.g. CERN is not going to change its financial practices (MTP etc) as a result of ISO 16363!**

# Scope – Not Just Scientific Data!

- Many (most) of the metrics cover the host organisation
- Scope agreed (WLCG OB) to extend to:
  1. **CERN's Scientific Data (draft OC);**
  2. "Archival material" (OC 3 – CERN's "Digital Memory");
  3. CERN Publications, Reports & Papers (OC 6)
- Some metrics will clearly require specific text for each (relevant) case
  - **But not e.g. site / cyber security, MTP etc.**

# Organisational Infrastructure

| | | |
|---|---|---|
| **3.1** | **Governance & Organisational Viability** | **Mission Statement, Preservation Policy, Implementation plan(s) etc.** **Operational Circular, DPHEP Reports** |
| **3.2** | **Organisational Structure & Staffing** | **Duties, staffing, professional development etc.** |
| **3.3** | **Procedural accountability & preservation policy framework** | **Designated communities, knowledge bases, policies & reviews, change management, transparency & accountability etc.** **Generic descriptions refined by project DMPs** |
| **3.4** | **Financial sustainability** | **Business planning processes, financial practices and procedures etc.** |
| **3.5** | **Contracts, licenses & liabilities** | **For the digital materials preserved…** **See later…** |

# Infrastructure & Security Risk Management

| | | |
|---|---|---|
| 5.1 | Technical Infrastructure Risk Management<br><br>**[ We do all of this, but is it documented? ]** | Technology watches, h/w & s/w changes, detection of bit corruption or loss, reporting, security updates, storage media refreshing, change management, critical processes, handling of multiple data copies etc |
| 5.2 | Security Risk Management<br><br>**[ Do we do all of this, and is it documented? ]** | Security risks (data, systems, personnel, physical plant), disaster preparedness and recovery plans … |

# Digital Object Management

| | | |
|---|---|---|
| 4.1 | Ingest: acquisition of content | |
| 4.2 | Ingest: creation of the AIP | Archival Information Package |
| 4.3 | Preservation planning | |
| 4.4 | AIP Preservation | |
| 4.5 | Information management | “FAIR” etc |
| 4.6 | Access management | |

**The plan is to address these after metrics 3 & 5…**

**Need to agree on scope: only “Open Data”?**

# Selected Metrics for GDB

- **The repository shall have a documented history of the changes to its operations, procedures, software, and hardware**
- **The repository shall define, collect, track, and appropriately provide its information integrity measurements**
- **The repository shall employ technology watches or other technology monitoring notification systems**
- **The repository shall have defined processes for storage media and/or hardware change**
- **The repository shall have implemented controls to adequately address each of the defined security risks**
- **The repository shall have suitable written disaster preparedness and recovery plans, including at least one off-site copy [ of recovery plan and key data ]**

Perhaps a GDB sub-group, including Tier1 representantives, could help elaborate and maintain the responses?

# Implications for WLCG Tier1s

- A number of sites sent people to the June 2015 ISO 16363 training
  - The decision whether to certify, by what method etc lies with the site (and maybe project, e.g. EUDAT)
- The CERN experience may be of value: we could provide advice and / or help review your responses
- **The motivation for Certification may come:**
  - Through WLCG; Funding Agencies, H2020 projects (such as EUDAT, EOSC *) and / or other

# Implications for Experiments

- The draft Operational Circular will need to be reviewed (including through the standard review process defined in OC 1)
  - Existing text based on DPHEP Collaboration Agreement together with (FAIR) DMPs: needs to be stable over period of 1 – 2 decades!

- **The metrics in chapter 4 cannot be addressed without close collaboration with the experiments!**

- **And those in section 3.5 – largely taken from WLCG Computing Model update – would benefit from being checked / updated as appropriate**

- **Target: complete (first) Certification and OC prior to (as part of) 2020 ESPP update**
  - **And have something a bit more concrete about LTDP / data sharing etc in the revised strategy (2013 is below)**

# DPHEP Worldwide Collaboration

- There is general agreement that LTDP in HEP includes: data, documentation, s/w + environment
  - ❑ **Invenio-based services often used for documentation; CVMFS + CernVM for s/w + environment – also in EOSC (Pilot)**
  - Which sites offer “bit preservation as a service”? (effectively required to become a TDR)
  - (CERN) Open Data portal currently limited to LHC experiments, as is Analysis Capture & Preservation

➢ **Clearly room for improvement but how to make it happen? A joint paper at CHEP?**

# 2020 Vision for LTDP in HEP

- ***Long-term – e.g. FCC timescales**: **disruptive change***
  - By 2020, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – easily **findable**, fully **usable** by **designated communities** with clear (Open) access policies and possibilities to annotate further
  - Best practices, tools and services well run-in, fully documented and sustainable; built in common with **other disciplines**, based on standards
  - **DPHEP portal**, through which data / tools accessed
    - ➢ **"HEP FAIRport": Findable, Accessible, Interoperable, Re-usable**

➢ **Agree with Funding Agencies clear targets & metrics**

# Summary

- **LTDP / "Open Data" / FAIR DMPs etc can (should) now be considered "mainstream"**
  - There are many conferences / workshops where issues are discussed and solutions proposed – "visible community" O(CHEP) in size
    - E.g. PV 2018 at RAL! (PV 2020 at CERN?)
  - Our knowledge and experience is regularly sought: e.g. on OECD working groups, e-IRG reports, EOSC "HLEG", RDA etc.
  - See DPHEP Indico pages for pointers to DPHEP and external events
- **Certification should help ensure that LTDP in HEP is both sustainable and sustained**


**Bits Decay: Do Something Today**

30th November 2017


CERN

# What is?

- Preservation
  - Data preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.
- Curation:
  - Digital curation involves maintaining, preserving and **adding value** to digital research data throughout its lifecycle.
- Stewardship:
  - Even more – including decisions on what data to preserve, what is the necessary meta-data (and perhaps also data management during active life of the data).
  - (From cradle to grave, according to EU HLEG report claiming a missing 500,000 data scientists)
  - 5% “total project” tax proposed (and disputed by some)

# Open (Linked) Data

★ Available on the web (whatever format) but with an open license, to be Open Data

★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)

★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★ All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★ All the above, plus: Link your data to other people's data to provide context


PV 2018 CONFERENCE

Science & Technology Facilities Council

RAL Space

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

UK SPACE AGENCY

The PV 2018 Conference welcomes you to its 9th edition, to be held 15th – 17th May 2018 at the Rutherford Appleton Laboratory, Harwell Space Cluster, hosted by the UK Space Agency and jointly organised by STFC and NCEO.