CERN

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

# Certification: Motivation, Benefits and Status

Joint WLCG – HSF Workshop
March 2018

Jamie.Shiers@cern.ch

# Certification – Introduction

- What is it?
- Who does it benefit? (And how?)
- What does it cost?
- Status and schedule for CERN
- (Examples)

# What is it?

- A set of **metrics**, typically based on the OAIS reference model (ISO 14721) that can be used to judge a repository's end-to-end "data preservation practices"
- Several methodologies exist, which some view as a hierarchy:
  1. *Entry level, e.g. DSA, WDS-DSA*
  2. *Intermediate, e.g. DIN*
  3. ***Full ISO conformance (ISO 16363)***
- ➢ ISO 14721 and 16363 were developed and are maintained by the "space community" (later)

# Who does it benefit?

- **Funding agencies**, who can better judge if the money they are providing will be used according to their requirements
  - e.g**. FAIR DMPs which call for preservation & re-use**
- **Data users** to be able to determine the "trustworthiness" of the data (**user surveys**)
- **Producers** (e.g. LHC experiments) to understand how and what a repository does to preserve their data

➢ **The data of most CERN experiments already lost!**

  - **By number of experiments, not by volume**
    - CERN Greybook: 776 completed experiments, ~20 active
    - "Preserved": LEP(4), LHC(4)

    ➢ **O(10) vs O(1000)**

# What does it cost?

- For CERN, the vast majority of the ISO 16363 metrics are covered by **existing, documented practices**
  - In some cases the documentation needed to be provided or improved
  - **In a small number of cases, e.g. business continuity, practices had to be put in place / improved!**
- It requires knowledge across a wide range of the organisation's activities

- **As a fraction of the cost of the LHC, the LHC experiments and / or compared to the measurable benefits it is simply in the noise**

- **Certification has been presented to WLCG OB & GDB and discussed with both current and previous DRC**

Choose the right metrics for your data and organisation – not the easiest!

# CERN as a “TDR” (ISO 16363)

- We believe **certification** will allow us to ensure that best practices are implemented and followed up on in the long-term: “**written into fabric of organisation**”
- Scope: **Scientific Data** and CERN’s **Digital Memory**
- **Timescale**: complete prior to 2019/2020 ESPP update
- Will also “ensure” adequate resources, staffing, training, succession plans etc.
- CERN can expect to exist until HL/HE LHC (2040/50)
- And beyond? FCC? Depends on physics…

*Current ESPP (approved May 2013):*
*(i) data preservation […] should be maintained and further developed.*

# ISO 16363 certification of CERN

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

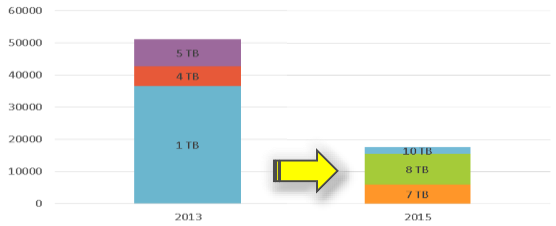
- Many of the elements in 3) and 5) covered by existing (and documented) CERN practices
  - **Some “weak” areas – being addressed – include disaster preparedness / recovery (together with EIROForum)**

➢ **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!**

# 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 $10^{-16}$**

See German’s presentation at March DPHEP workshop

# Organisational Infrastructure

| | | |
|---|---|---|
| 3.1 | Governance & Organisational Viability | Mission Statement, Preservation Policy, Implementation plan(s) etc.<br>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.<br>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… |

# Infrastructure & Security Risk Management

| | | |
|---|---|---|
| 5.1 | Technical Infrastructure Risk Management | 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 | 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 | |

# Current Status

- We have prepared written answers to all ISO 16363 metrics (H2 2017)
- They have been reviewed by some experts
- They have been sent to PTAB – the only accredited body to date in the MS

➢ **We await feedback (June?) before organising a formal on-site audit**
    - Annual "surveillance" audits follow, then re-cert.

- "Plan" is to obtain (and retain) certification in step with ESPP update (see later)

# Summary

- ISO 16363 certification of CERN under way: covers end-to-end: from deposit to re-use.
- Depositors and (re-)users can ask any question, based on >100 agreed metrics
- Some sort of certification expected to become semi-mandatory on Run3 timescales
- Important that large facilities / projects on leading edge in this respect