Current Situation

Raw Data

Data are deleted after 50 days from disk and after one year from tape

No persistent identifiers

No data management plan

Strong difference between in-house research and visitor data

Metadata

Metadata not collected systematically

Experiment report public

Changing landscape

- Scientific data are more and more considered like a publication and/or part of the publication
- Movement to Open Data is growing e.g. OECD, G8, RDA, …
- IUCr dddwg initiative for open data
- Pressure is increasing on publicly funded research institutes to follow
- H2020 participation will be conditioned on a data management plan

Data Policy at other Research Institutes

Neutrons

ILL – PanData-like policy since 3 years

ISIS – PanData-like policy since 3 years

Photons

ELETTRA – PanData-like policy since 1 year

ALBA – PanData-like policy proposed

SLS – Currently under preparation

Other

Alfred Wegener Institut (Helmholtz) – Open Data Policy

Astronomy, Biology, CERN, … – Open Data Policies

The ESRF Data Policy

- **Data needs to be properly managed to allow:**
  - linking to publications (increasingly requested by publishers)
  - re-analysis
  - verification
  - new research
  - preservation of unique data sets

- **The ESRF Data Policy defines the conditions for:**
  - Data ownership
  - Data curation
  - Data archiving
  - Open access to data

One of the two SL8500 tape libraries which will be used for the archival of ESRF beamline data

Following recommendation by SAC, the ESRF Data Policy was approved by Council on

1 December 2015

ESRF data policy – main elements

Raw data and associated metadata

- ✓ ESRF is the custodian of raw data and metadata from all beamlines (including CRGs)
- ✓ ESRF will automatically collect metadata for all experiments
- ✓ ESRF will store metadata in a metadata catalogue (icat)
- ✓ High level metadata will be published as soon as possible, i.e.

  Title, Authors, Beamline, Abstract, Experiment Report
- ✓ Experimental team has sole access to the data during the so-called **embargo period of 3 years**; request to extend embargo period can be made
- ✓ After embargo ESRF will make the data “Open Access” under CC-BY-4 license
- ✓ Users need to create an identifier to get Open Access data
- ✓ Proprietary data belong by default to the PI and are not archived unless explicitly agreed

ESRF data policy

Data Access

- Access to raw data and metadata will be via a searchable on-line catalogue (icat)
- Access to the on-line catalogue of the ESRF will be restricted to registered users of the on-line catalogue. The ESRF will set up an on-line procedure to become a registered user of the catalogue, e.g. with an Umbrella ID
- Access to proposals will only be provided to the experimental team and appropriate facility staff
- PI has the possibility to transfer parts or the totality of her/his rights during the embargo period to another registered person
- PI has the right to create and distribute copies of the raw data
- PI has the possibility to render data public before the end of the embargo period

Implications of the data policy

What do we need to curate data for 5 to 10 years ?

- A metadata catalogue → icat (already installed)
- Good metadata on all beamlines → modify the data acquisition
- Electronic logbook → install a standard electronic logbook on all beamlines
- Hooks in the experiments → modify macros on each beamline
- A catalogue of data to curate → identify what data to register + archive
- Identity management → persistent IDs

- Lots of tape storage → money for tapes and manpower to install
- Automatic way to restore data → manpower to implement workflow

- Current production is **~2 PB / year** in **2015**
- Assuming linear growth to **15 PB / year** in **2025** → **45 PB on tape**

Advantages of the ESRF data policy

Research Teams / BL Scientists / Review Panels / Community

- ✓ Metadata are systematically collected
- ✓ Better and continuously improved metadata
- ✓ Data are managed and archived for long term
- ✓ Metadata can be searched and downloaded easily
- ✓ Compliance with Data Management Plan required by H2020
- ✓ Data can be referenced in publications via PIDs (DOI)

ESRF

- ✓ Better data management and follow-up
- ✓ Data from the ESRF can be traced and verified
- ✓ Better statistics about publications using ESRF data
- ✓ Conformance with France/European/World wide move to Open Access
- ✓ Eventually will lead to ESRF data being used more

Data !

Round Table

- ✓ Open access – Yes or No ? Planned ?
  - ✓ If yes, when ?
  - ✓ If yes, what parameters (embargo period, proposal types, exceptions, access, etc) ?
  - ✓ If yes, implementation, e.g. how users agree, what instruments, electronic logbooks, level of DOI labelling, etc ?
- ✓ Data storage duration ?
- ✓ Data confidentiality level – current and open access if planned ?
- ✓ Problems & advantages ?