THE CHALLENGES IMPLEMENTING ESRF DATA POLICY

ESRF | The European Synchrotron

Andy Götz *on behalf of*

Data Implementation Working Group

THE CHALLENGES IMPLEMENTING ESRF DATA POLICY

THE CHALLENGES IMPLEMENTING ESRF DATA POLICY

1. Data Policy
2. Defining Metadata
3. Data Format
4. Metadata Catalogue
5. E-logbook
6. Finding Data
7. Long Term Archiving
8. Roadmap

Conclusion

WHO SAID THIS ?

Raw
Data
Now!

SIR TIM BERNERS LEE – INVENTOR OF THE WEB

ESRF DATA POLICY – MAIN POINTS

- ESRF is custodian of data and metadata
- ESRF to collect high quality metadata to facilitate reuse of data
- ESRF will keep metadata forever
- ESRF will keep raw (or reduced) data for 10 years
- Data will be registered in a data catalogue (ICAT)
- Data will be published with a Digital Object Indentifier (DOI)
- The experimental team has exclusive access to data during the embargo period (3 years which can be extended on request)
- Data will be made public after the embargo period under licence CC-BY 4.0
- Data Policy will be implemented on all beamlines by 2020

The hard thing to do
and the
right thing to do
are usually
the same thing.

CORE PRODUCT

« DATA is our core product – we must manage it correctly to ensure it is of high quality, traceable, reliable, curated, and reusable »

Andy Götz (Workshop on Active DMP at CERN June 2016)

THE GOAL IS REUSE

« Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. »

Mark D. Wilkinson et al. (March 2016 in Nature www.nature.com/scientificdata)

–

CHALLENGE #1 - DEFINING A DATA POLICY

- **2011 : PaNData Europe FP7 work package 2 delivered a generic data policy**
- **2011 : Proposed the ESRF management to adopt it but failed to convince**
- **2011 to 2015 : rise of Open Science → Data Management Plans → Open Data**
- **2015 : ESRF Data Policy become necessary + feasible, adopted by Council 1/12/2015**
- **2020 : ESRF Data Policy implemented on all beamlines**

« be patient, do not give up, now is the time ! »

« TIMES ARE A CHANGING » … BOB DYLAN

NOBELPREISE

Einsame Helden? Das war gestern

Die Wissenschaft muss sich noch weiter öffnen VON ANDREAS SENTKER

Nur wenige Jurys nehmen es mit der Geheimhaltung so genau wie jene, die über die Nobelpreise befinden. 573-mal sind die Preise von 1901 bis 2015 vergeben worden, an 870 Menschen und 23 Organisationen. Und niemand außer den Juroren wusste zuvor, wer sie erhalten würde. Das macht die alljährliche Bekanntgabe so aufregend.

Die Inszenierung gelingt auch, weil Forschung hier ein Gesicht bekommt. Der Preis wird für – oft längst vergangene – Heldentaten Einzelner verliehen.

Bald jedoch soll die Wissenschaft anders aussehen. Nicht verborgen, nicht einsam, offen soll sie sein und kooperativ: Open Science. Forscher sollen nicht nur die Resultate, sondern auch die Daten ihrer Experimente zugänglich machen (Open Data). Sie sollen Bürger in ihre Arbeit einbeziehen (Participation). Ihre Publikationen sollen jedem zugänglich sein (Open Access).

Warum diese Wende?

Sie hat – zunächst – interne Gründe: Mehr Kooperation verspricht mehr Erkenntnis, Big Data gilt in der Forschung als der nächste große Treiber (siehe Ressort Wissen ab Seite 35). Mehr Transparenz ist zudem zur Qualitätssicherung erforderlich: Sie erschwert Fälschungen, macht Experimente nachvollziehbar und erleichtert so die Selbstkontrolle der Forschung.

Aber auch ein gesellschaftlicher Wandel lässt Forscher umdenken. Den Nobelpreisen zum

CHALLENGE #2 - DEFINING METADATA

- **One of the biggest challenges – why ?**
- ***« Metadata is the data you don't need for your data analysis »*** **Armando Solé**
- → **Hard to find people motivated to work on this**
- **ESRF lucky to have Armando+Alex+Roberto** and **some motivated scientists (Peter+Wout+…)**
- **Nexus** is our **baseline** but we **add local definitions** as we go along **(not using Nexus Applications)**
- **Goal** is to **produce** a **complete set of metadata for experiment logging, debugging and data analysis**

If it's broken - Fix it

If you don't like it - Change it

If you want something - Take it

NeXus

Don't complain about it

Just do something about it.

CHALLENGE #3 – DATA FORMAT

- Ingredients : HDF5,Nexus

- A single master file / experiment
- Master file stores metadata (copy of database)
- Links to data file(s) with multiple datasets
- Easier to analyse, backup + transfer
- → Need to convert data analysis programs
- ESRF developing SILX library with HDF5 support
- → See poster on SILX (A.Solé) today

CHALLENGE #4 – CHOOSING A METADATA CATALOGUE

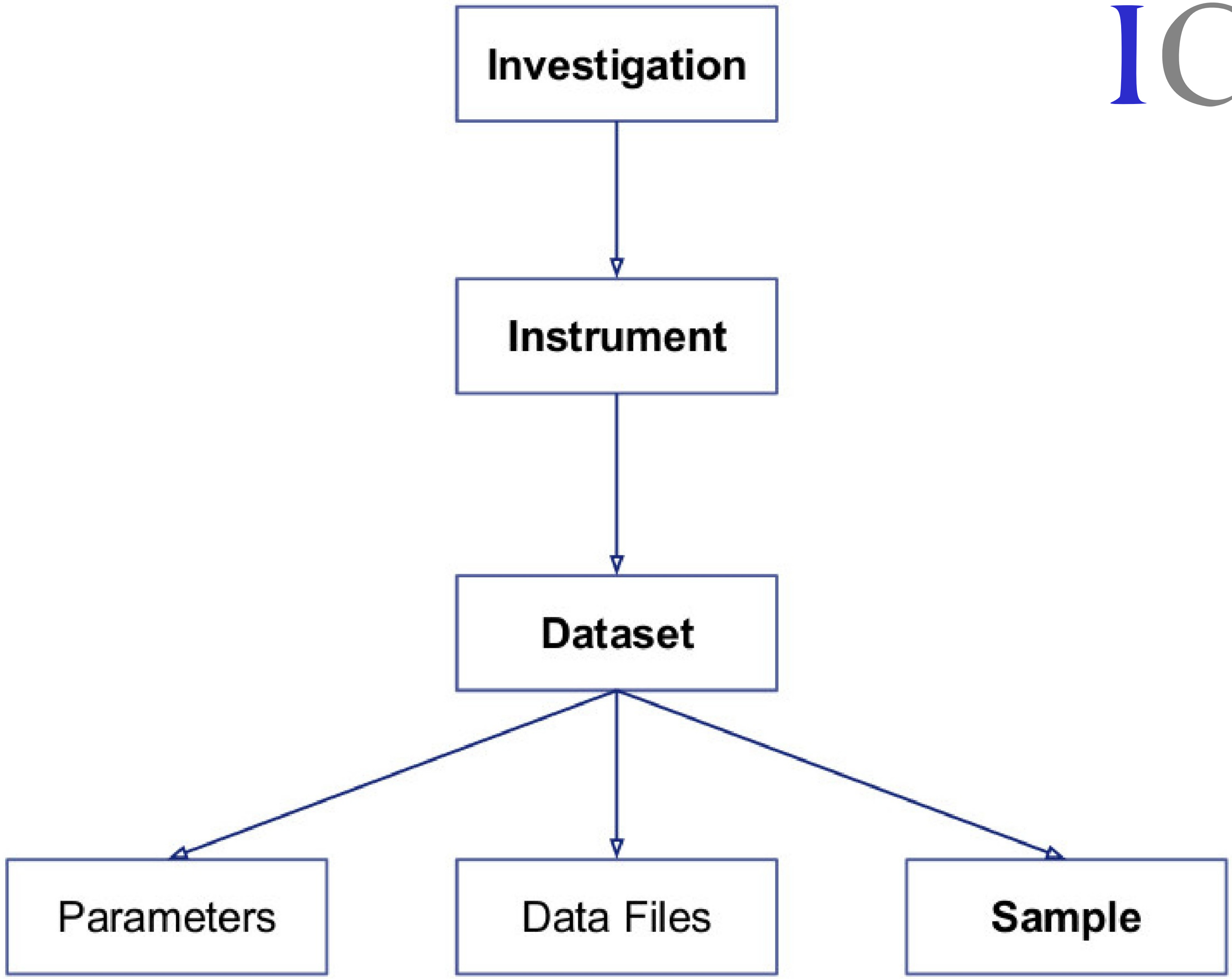
- Many Metadata catalogues
- CRISP project compared 20 metadata catalogues

CHALLENGE #4 – CHOOSING A METADATA CATALOGUE

ICAT

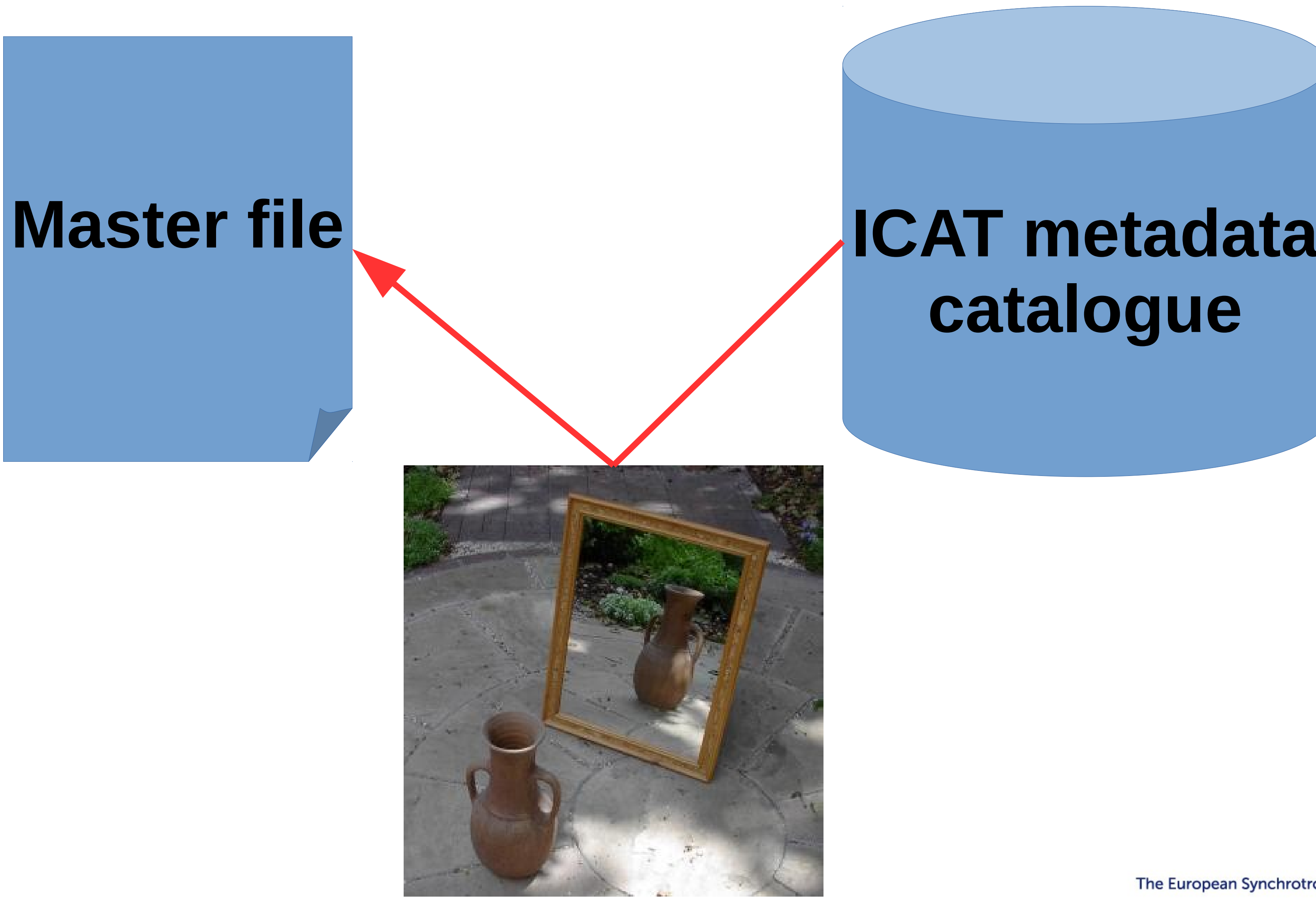
- **ICAT was chosen for its generic data model which captures the scientific experiment**
- **ICAT not invented at here (ESRF)**
- **ISPyB (for MX)** was close second
- **ICAT community** is **collaborating** to **address all steps of data management**
- → **See next two talks (Steve+Frazer)**

ICAT METADATA DATA MODEL



METADATA DEFINITION IN ICAT

- HDF/Nexus Metadata master file is a mirror image of the Metadata stored in ICAT



METADATA DEFINITION IN ICAT

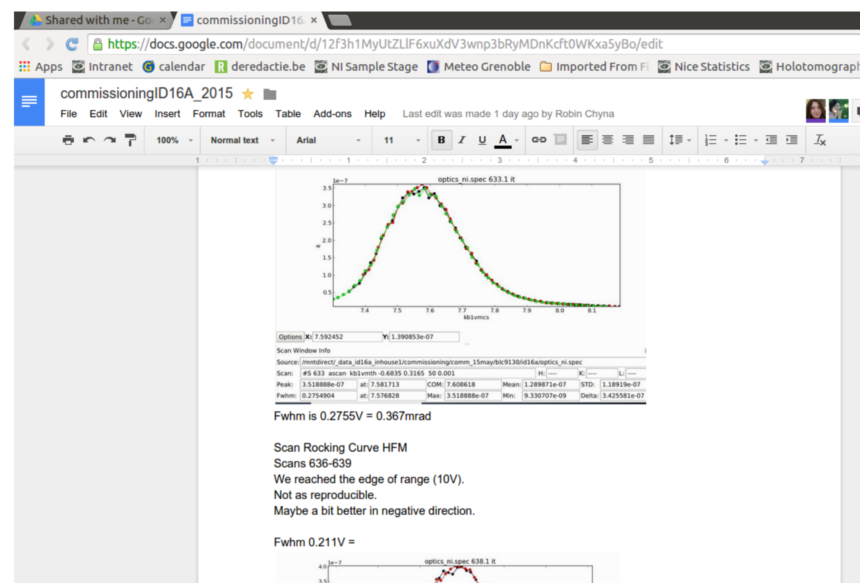
- **Follow the Nexus conventions when defining Dataset Parameters :**

InstrumentPositioners_name InstrumentPositioners_value
InstrumentMonochromator_energy InstrumentMonochromator_wavelength
InstrumentMonochromatorCrystal_usage InstrumentMonochromatorCrystal_d_spacing
InstrumentMonochromatorCrystal_type InstrumentMonochromatorCrystal_reflection
InstrumentSource_mode InstrumentSource_current InstrumentSlitPrimary_name
InstrumentSlitPrimary_vertical_gap InstrumentSlitPrimary_vertical_offset
InstrumentSlitPrimary_horizontal_gap InstrumentSlitPrimary_horizontal_offset
InstrumentSlitPrimary_blade_up InstrumentSlitPrimary_blade_down
InstrumentSlitPrimary_blade_front InstrumentSlitPrimary_blade_back
InstrumentSlitSecondary_name InstrumentSlits_name InstrumentSlits_vertical_gap
InstrumentSlits_vertical_offset InstrumentSlits_horizontal_gap
InstrumentSlits_horizontal_offset InstrumentSlits_blade_up
InstrumentSlits_blade_down InstrumentSlits_blade_front InstrumentSlits_blade_back
InstrumentAttenuatorPositioners_name InstrumentAttenuatorPositioners_value
InstrumentInsertionDevice_gap_name InstrumentInsertionDevice_gap_value
InstrumentInsertionDevice_taper_name InstrumentInsertionDevice_taper_value
InstrumentOpticsPositioners_name InstrumentOpticsPositioners_value
InstrumentDetector01_name InstrumentDetector01Positioners_name
InstrumentDetector01Positioners_value InstrumentDetector02_name
InstrumentDetector02Positioners_name InstrumentDetector02Positioners_value
……

CHALLENGE #5 – E-LOGBOOK

- Get rid of paper logbooks !
- No solution for now – we will work on this in the future.

- Some beamlines using Wikimedia, Google doc ...

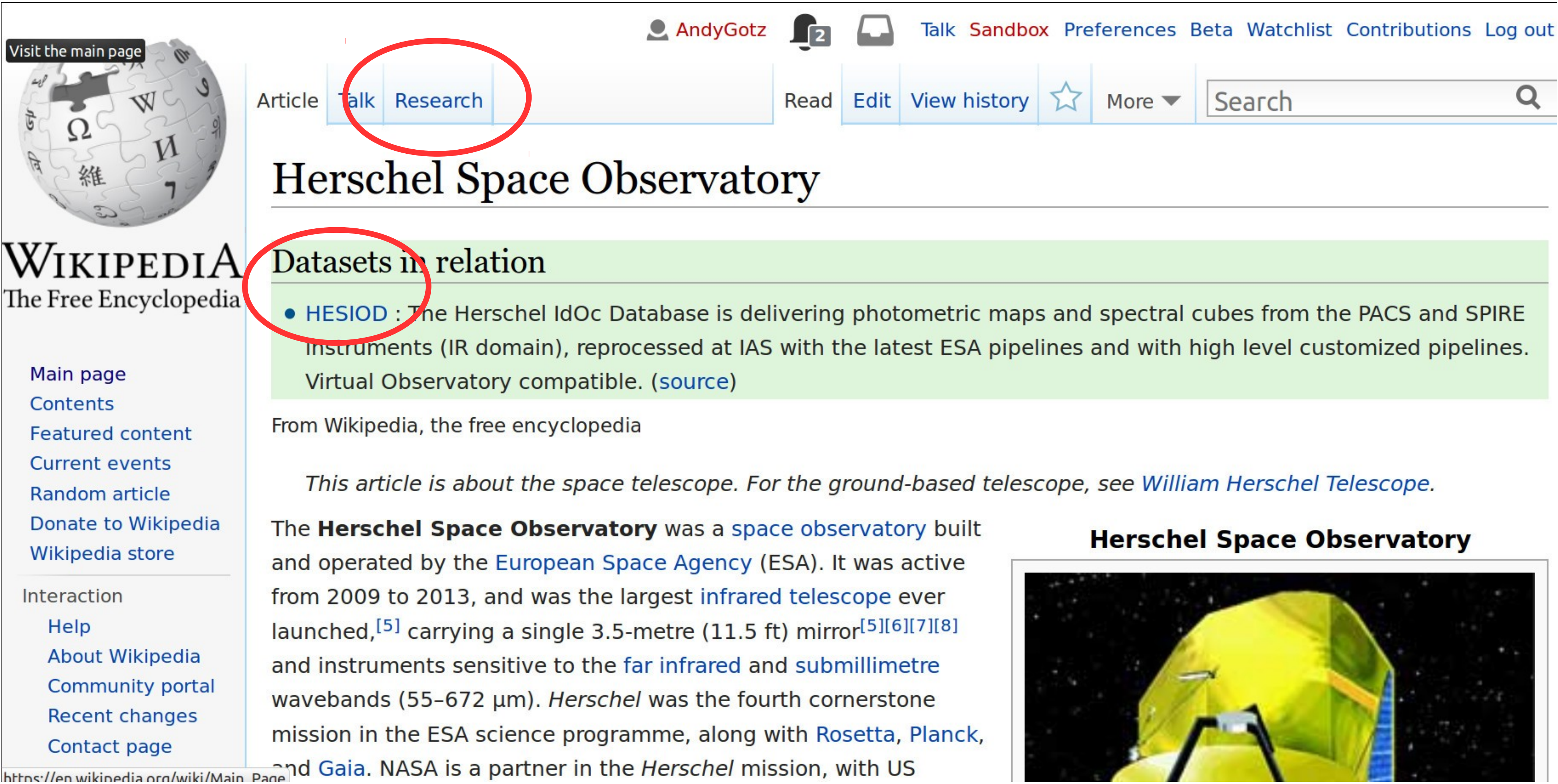


CHALLENGE #6 – FINDING DATA

- **Data must be FAIR i.e. F**indable**, A**ccessible**, I**nter-Operable**, R**e-Usable (http://www.nature.com/articles/sdata201618)

- **Latest version of TopCat (ICAT web UI) offers searching by Investigation, Sample, Dataset**
- **Parameter/metadata searching still needs work**
- **Publish data DOIs, link to public repositories**
- **Idea : implement the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)**
- **Dream (?) : Link data to Wikipedia ...**

FINDING DATASETS VIA WIKIPEDIA



Refer to → https://io.datascience-paris-saclay.fr/map.php

CHALLENGE #7 – LONG TERM ARCHIVING

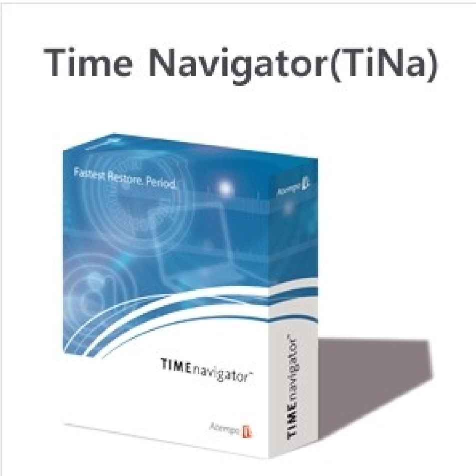
- **What raw data to curate and archive ?**
- **Only data which have been produced at the ESRF by a well defined process**
- **In a well defined format which ESRF provides tools for reading and writing**

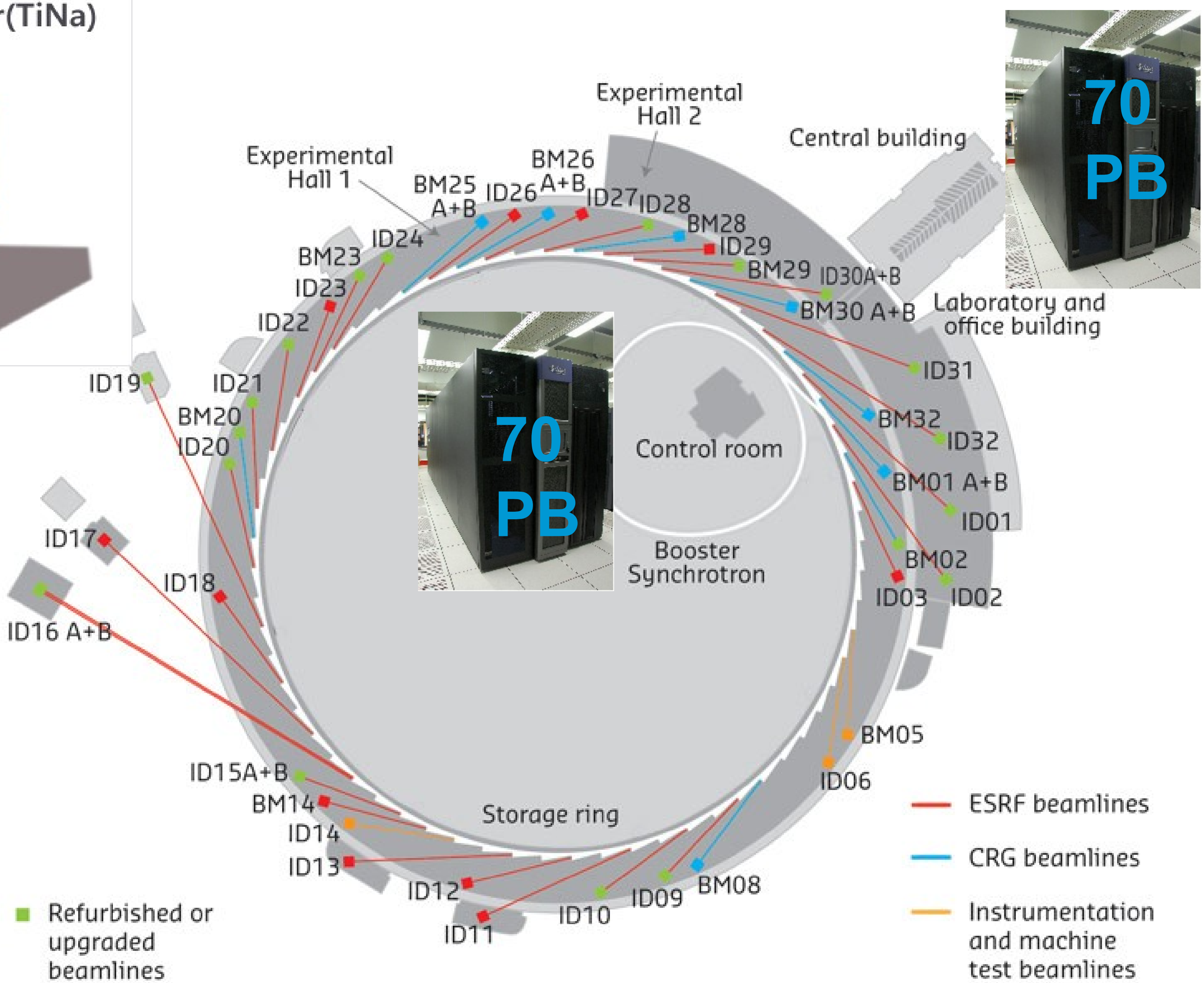
- **What if there is too much raw data ?**
  - Store the reduced data only
  - Limit the amount of data
  - Store the metadata

CHALLENGE #7 – LONG TERM ARCHIVING

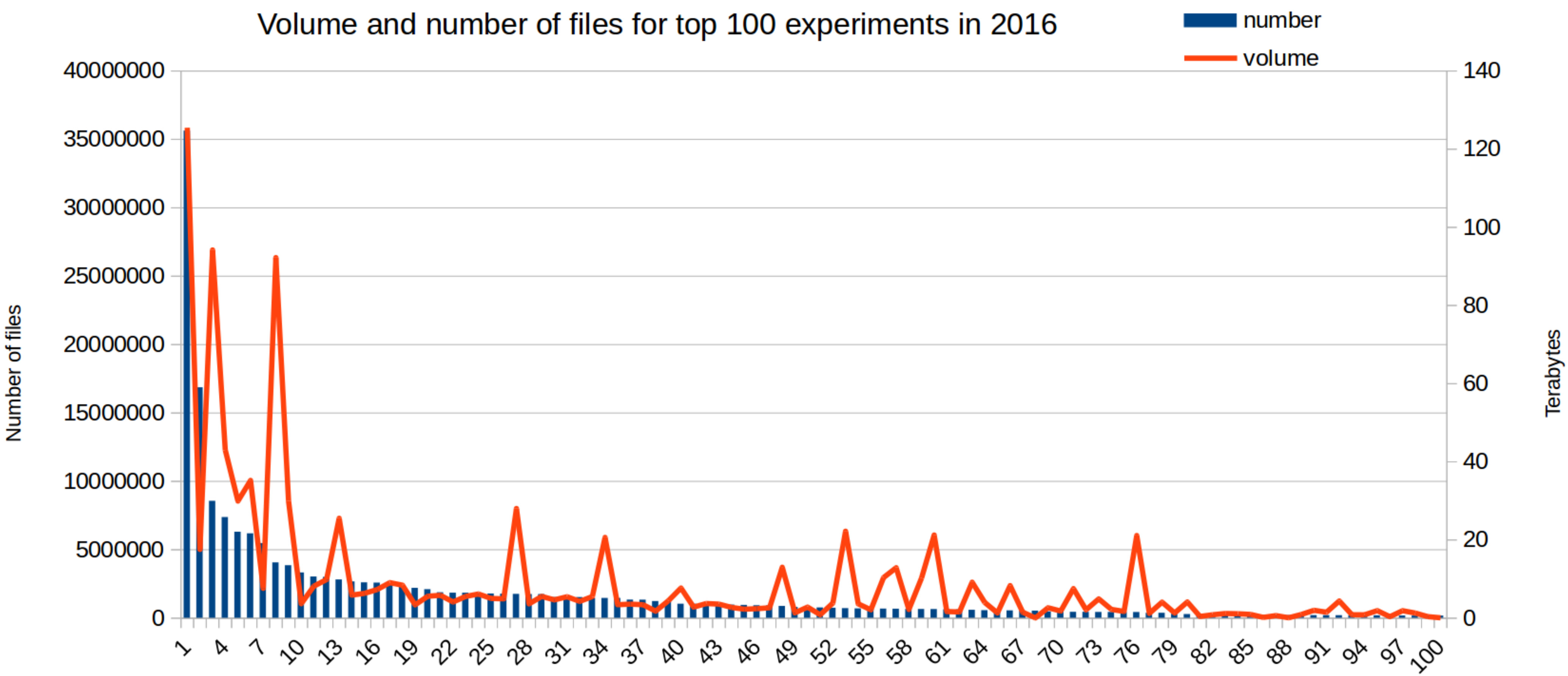
- **How much does it cost ?**
- **Additional costs are :**
  - **Tape drives + tapes = 100 000 euros / year**
  - **Human resources = 2.5 FTEs / yr over 4 years**
    - 1 data manager, 1 metadata, 0.5 for archiving
- **ESRF has 2 tape libraries on site**
- **Upgraded tape drives to T1000 drives and T2 tapes with 8.5 TB capacity and 300 MB/s read/write**
- **Data Policy restricted to data and not compute resources for data analysis**

LONG TERM ARCHIVING (LTA) – 2 COPIES





TOO MANY FILES - CHALLENGE IS TO REDUCE THEM !



→ Limit for LTA is 1000 files / 8 hours / experiment i.e. roughly 20000 per experiment / week ...

CHALLENGE #8 – ROADMAP

- ALL beamlines to implement Data Policy by 2020
- Corresponds to 10 beamlines / yr over 4 years
- So far we are on track for 2016 (10 beamlines) !

CONCLUSION

- A Data Policy is (1) necessary in today's scientific and political landscape and (2) feasible
- Users are the main winners because they get better managed data (Data Management Plans)
- Open Data Policy opens up new possibilities e.g. implementing Data as a Service and Open Data
- Data is our core product, the goal is to ensure its re-use as much as possible !

Kun døde fisk svømmer med strømmen

Only dead fish swim with the current