THE CHALLENGES IMPLEMENTING ESRF DATA POLICY

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.

Attribution 4.0 International (CC BY 4.0)

This is a human-readable summary of (and not a substitute for) the **license**.

Disclaimer

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or **technological measures** that legally restrict others from

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

25 Reiseziele im Osten

DIE ZEIT

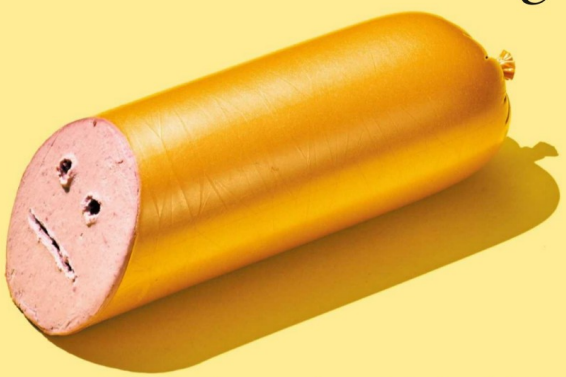
WOCHENZEITUNG FÜR POLITIK WIRTSCHAFT WISSEN UND KULTUR

PREIS DEUTSCHLAND 4,90 €

Diese Woche eine Landkarte in der ZEIT

6. OKTOBER 2016 N° 42

Die Macht der Beleidigten



Politiker, Dozenten, aufgebrachte Bürger – wer heute Aufmerksamkeit erhalten will, muss sich gekränkt zeigen. Und das Verrückte ist: Es funktioniert JENS JESSEN IM FEUILLETON

Merkels Ansage

»Mitgefühl ist nicht mein Motiv«
Kurswechsel? Von wegen! Die Kanzlerin bekräftigt ihre Flüchtlingspolitik
Seite 2/3

Guten Morgen!
Ein Heft über den letzten Rückzugsort – das Badezimmer
ZEITmagazin

Im Totenwald
Drei Jahrzehnte nach der Tat: Auf der Spur eines mutmaßlichen Serienmörders
Recht & Unrecht, Seite 14

PROMINENT IGNORIERT

Umweltfreundlich
bei den spektakulärsten Überfällen der Kriminalgeschichte: Sie hielten die Räuber von Suzuki oder Kawasaki. Fünf Gangster, die am Montag die amerikanische Schauspielerin Kim Kardashian in einem Pariser Appartement überfallen und gefesselt haben, um Schmuck im Wert von neun Millionen Euro zu rauben, entkamen auf Fahrrädern. Die Idee des umweltfreundlichen Radfahrens ist in der Unterwelt angekommen. GRN.

Kl. Fotos (v. o.): D. Butzmann; DigitalVision/Getty Images; M. Piasecki/DG Images/Getty Images

Zeitverlag Gerd Bucerius GmbH & Co.KG, 20079 Hamburg
Telefon 040 / 32 80 - 0; E-Mail: DieZeit@zeit.de, Leserbriefe@zeit.de

ZEIT ONLINE GmbH: www.zeit.de; ZEIT-Stellenmarkt: www.jobs.zeit.de

ABONNENTENSERVICE:
Tel. 040 / 42 23 70 70,
Fax 040 / 42 23 70 90,
E-Mail: abo@zeit.de

PREISE IM AUSLAND:
DK 49,00/FIN 7,50/N 66,00/E 6,10/CAN 6,20/F 6,10/NL 5,30/A 5,00/CH 7,30/I 6,10/GR 6,70/B 5,30/F 6,30/L 5,30/H 2090,00

PSEUDONYM GELÜFTET

Elena Ferrante schuldet uns nichts

Sie wurde verfolgt, ihre Privatsphäre verletzt. Warum nur? VON MARIAM LAU

Als am vergangenen Wochenende fünf internationale Medien das Pseudonym der italienischen Schriftstellerin Elena Ferrante lüfteten, dachten manche Leser erschrocken, der Autorin sei vielleicht etwas zugestoßen. Welchen anderen Grund könnte es geben, die Frau, der die Welt den neapolitanischen Roman *Meine geniale Freundin* verdankt, gegen ihren Willen ans Licht der Öffentlichkeit zu zerren? Sie hat kein Verbrechen begangen. Sie hat keine Panama Papers. Ferrante findet, ihr Romanzyklus *Neapolitanische Saga* solle ein eigenes Leben führen, die Autorin habe dahinter zu verschwinden. Sie sieht ja, dass es im Feuilleton immer öfter umgekehrt gehandhabt wird.

Ein Vierteljahrhundert lang ist es ihr gelungen, ihr Geheimnis zu wahren. Sie hat vor Jahrzehnten einmal angekündigt, dafür notfalls auch zu lügen. Damit rief sie den investigativen Wirtschaftsjournalisten Claudio Gatti auf den Plan, der findet, Ferrante habe damit ihr Recht auf Anonymität verwirkt. Sie habe Journalisten und Kritiker »geradezu herausgefordert« – ein Argument, das aus anderen, unschönen Zusammenhängen im Verhältnis zwischen Männern und Frauen bekannt ist.

Verhältnisse übrigens, die Ferrante in ihrer Terralogie Schicht um Schicht freilegt. Monatelang hat Gatti in ihren Unterlagen gewühlt, in Gehaltszetteln, Grundbüchern; argwöhnisch die Lage ihrer Wohnung »in einer der schönsten Straßen Roms« umschlichen. Nein, das ging zu weit. Eine Übersetzerin wie Anita Raja – die Frau, die er im Verdacht hatte – würde sich so etwas niemals leisten können. Ferrante ist enttarnt.

Gatti ist mit seiner Anmaßung nicht allein. Viele finden, Celebrityx verdanken der Öffentlichkeit ihren Ruhm, also schulden sie ihr auch etwas. Bis weit ins Politische hinein greift die Vorstellung um sich, nur totale Transparenz sei noch mit demokratischen Ver-

KOLUMBIEN

Preis der Wahrheit

Die geplante Amnestie für die Farc-Rebellen empört die Opfer. Doch nur so kann das südamerikanische Land Frieden finden VON ANGELA KÖCKRITZ

Sie haben gemordet, vergewaltigt, gebrandschatzt, entführt. Haben sich an den sagenhaften Einnahmen aus dem Kokain-Business bereichert und Millionen Kolumbianern das Leben zur Hölle gemacht. 52 Jahre lang wütete in Kolumbien der Bürgerkrieg zwischen linken Farc-Rebellen und rechten Paramilitärs. Er kostete 220 000 Menschen das Leben, sechs Millionen Menschen wurden vertrieben.

Und ausgerechnet diesen Farc-Rebellen will man jetzt milde Strafen, Straffreiheit, ja Privilegien gewähren? Wie muss sich das für jene anhören, die ein Kind in diesem Krieg verloren haben? Die Tag für Tag mit dem Schmerz leben müssen, während irgendwo da draußen der Mörder herumläuft, vielleicht sogar Parlamentarier wird?

Die Kolumbianer entschieden am Sonntag: Der Preis ist zu hoch. Präsident Juan Manuel Santos hatte sein Volk zum Referendum über das historische Friedensabkommen mit den Rebellen der Farc-Guerilla gebeten. Die meisten Kolumbianer gingen nicht zur Abstimmung, der Rest stimmte mit einem Vorsprung von gerade mal 54 000 Stimmen: *No!* Der Präsident erklärte daraufhin: Der Waffenstillstand bleibe in Kraft, es werde weiterverhandelt. Das Abkommen hat noch eine Chance.

Die Kolumbianer sind mit ihrem Ringen nicht allein. Den Gegensatz »Frieden oder Gerechtigkeit« mussten schon viele Gesellschaften ausverhandeln. Er hat die Deutschen 1945 genauso beschäftigt wie die Südafrikaner nach dem Ende der Apartheid. Und eines Tages werden sich die Syrer dem zu stellen haben.

Vielen Kolumbianern erschien das Abkommen ungerecht. Die Übergangsjustiz sieht eine Amnestie für einfache Mitglieder der Rebellengruppe vor. Jenen Anführern, die Verbrechen gegen die Menschlichkeit zu verantworten haben – dazu gehören Mord, Entführung, Vergewaltigung und die Rekrutierung von Kindersoldaten –, drohen im Falle eines Geständnisses Farc so viel?, fragen sie. Denn im Gegenzug dafür, dass die 15 000 Rebellen ihre Waffen abgeben und aus dem Narco-Business aussteigen (viele bezweifeln, dass das wirklich allen gelingen werden), soll sich die Farc in eine reguläre politische Partei verwandeln. Und zwar eine, die besondere Privilegien genießt. In den nächsten beiden Wahlen soll sie automatisch eine Mindestanzahl von Sitzen erhalten – und zwar unabhängig vom Wahlergebnis.

Wem ist zu verübeln, das empörend und unmoralisch zu finden? Wem ist zu verdenken, Gerechtigkeit für den Mord an seinem Sohn, seiner Tochter zu verlangen?

Niemandem. Und doch: Frieden und Wandel haben ihren Preis. Vor allem dann, wenn jene, denen etwas abgerungen werden soll, über Waffen und Einkünfte verfügen. Denn kein Rebell und kein Diktator wird sich selbst das eigene Grab schaufeln. Keiner wird dazu bereit sein, sich und die Seinen zu einem Leben im Knast zu verdammen.

Beispiel Spanien. Nach dem Tod des Diktators Francisco Franco im Jahr 1975 erkämpften sich seine Anhänger zum demokratischen Übergang bereit – wenn sie nicht strafrechtlich verfolgt würden.

Beispiel Südafrika. Nach dem Ende der Apartheid setzte Nelson Mandela 1996 eine Wahrheitskommission ein. Den Angeklagten wurde Amnestie zugesagt, wenn sie ihre Verbrechen vor der Kommission gestanden.

Beispiel Myanmar. Bevor Than Shwe, der oberste Führer der Militärjunta, im Jahr 2011 zurücktrat und damit demokratische Reformen ermöglichte, sicherte er sich und seinen Militärs Straffreiheit zu.

Die Liste ließe sich fortsetzen. Die Fälle sind sehr unterschiedlich, haben aber Gemeinsamkeiten. Die Deals waren alles andere als perfekt. In allen Ländern beklagt man, dass die Amnestie eine Aufarbeitung der Vergangenheit behindert habe. Und doch war das, was danach kam, besser als das, was davor war. Un-

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 Trotz: Das Misstrauen ihnen gegenüber wächst. Nicht nur die Politik (Trump versus Clinton), auch die Wissenschaft scheint im postfaktischen Zeitalter angekommen. Verschwörungstheorien und paranormalwissenschaftliche Thesen verbreiten sich. Auf Artikel zum Thema Impfen reagieren viele, sogar gebildete Menschen, mit der Behauptung, Viren seien eine Erfindung der Pharmaindustrie. Open Science ist also eine vertrauensbildende Maßnahme. Die zweifeln an Viren? Hier

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

ESRF

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**

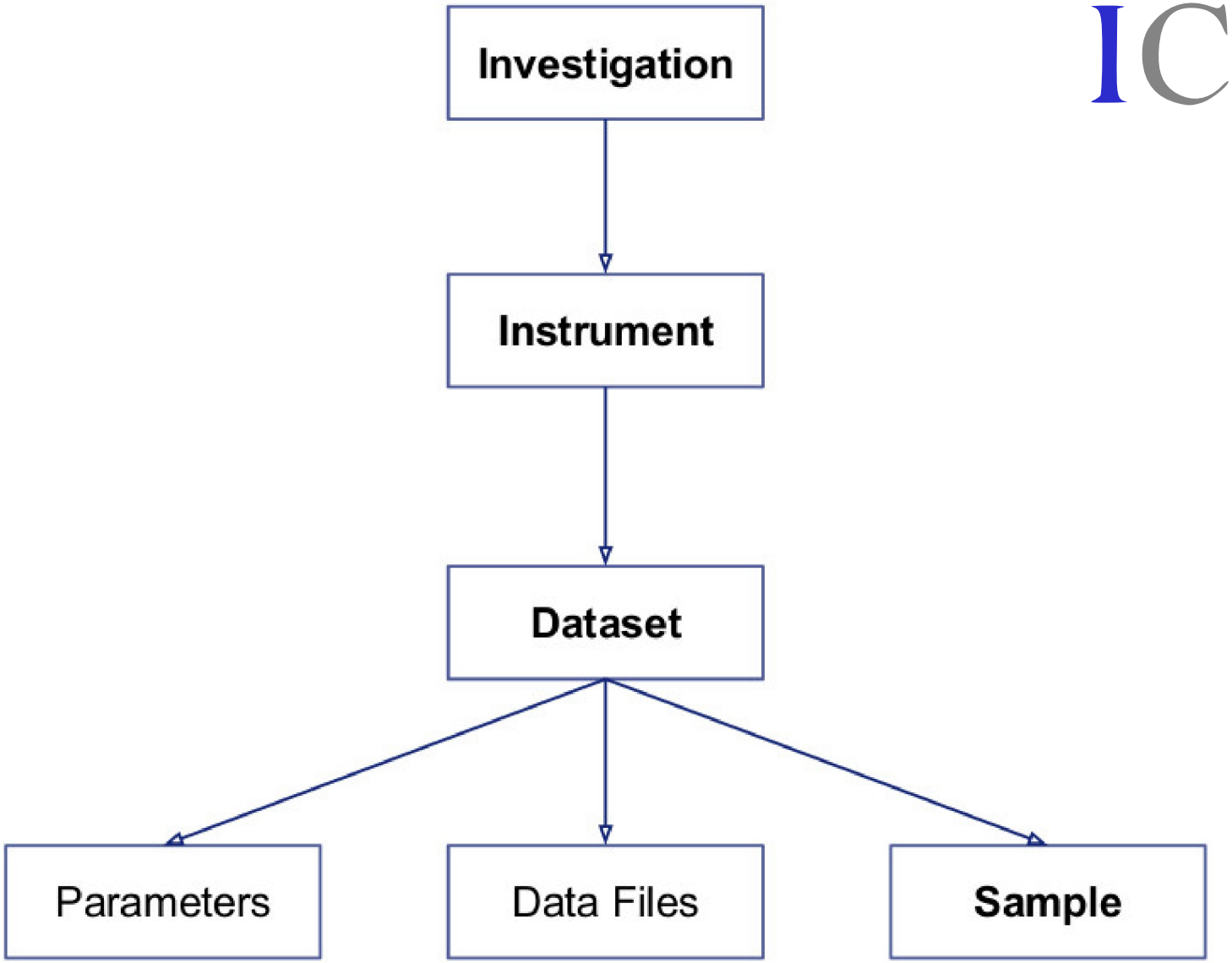
CHALLENGE #4 – CHOOSING A METADATA CATALOGUE

- Many Metadata catalogues
- CRISP project compared
  20 metadata catalogues

CHALLENGE #4 – CHOOSING A METADATA CATALOGUE

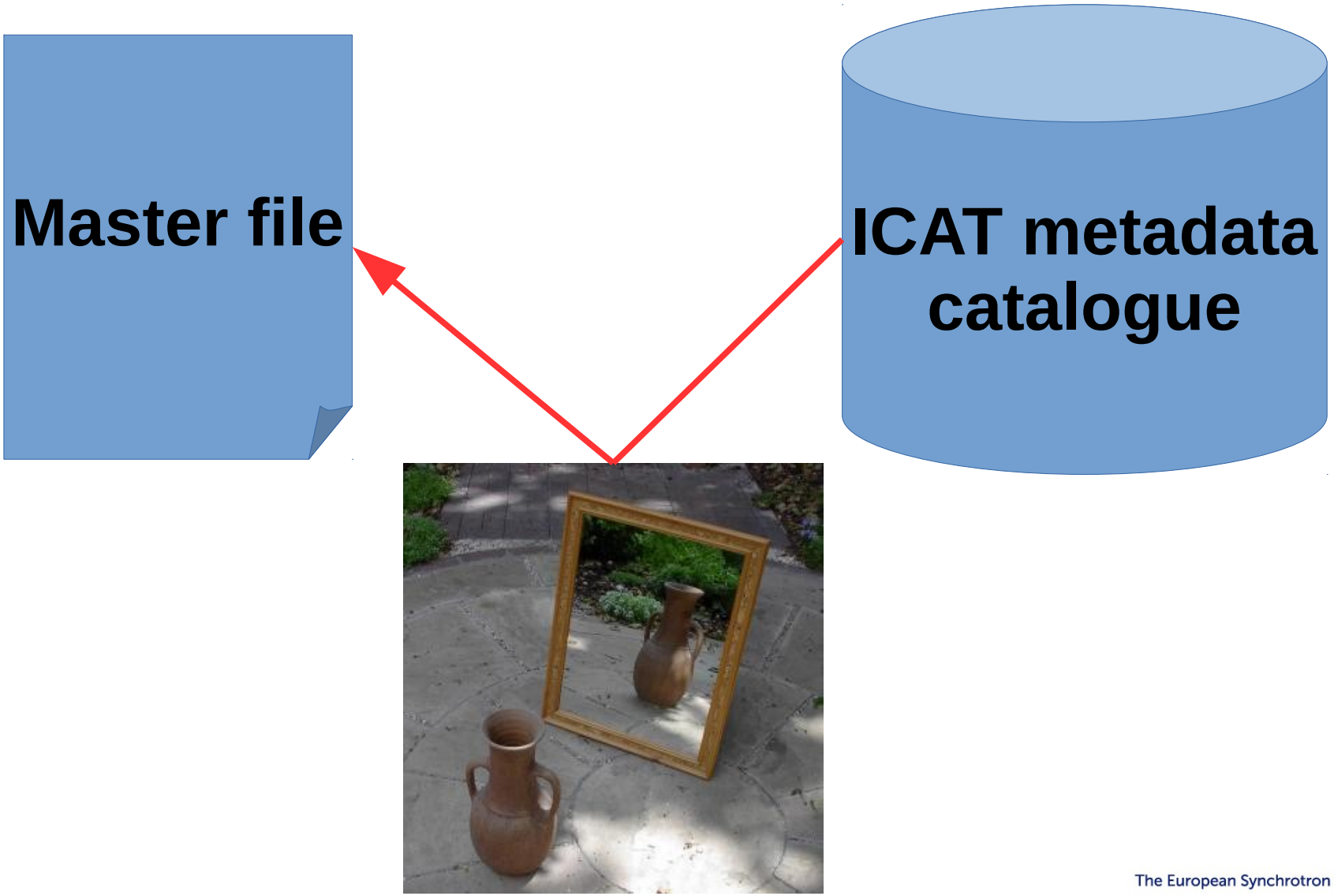
- **ICAT was chosen for its generic data model which captures the scientific experiment**
- **ICAT not invented at here (ESRF)**
- **ISPyB (for MX)** was close second
- **ICAT community** is **collaborating** to **address all steps of data management**
- → **See next two talks (Steve+Frazer)**

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

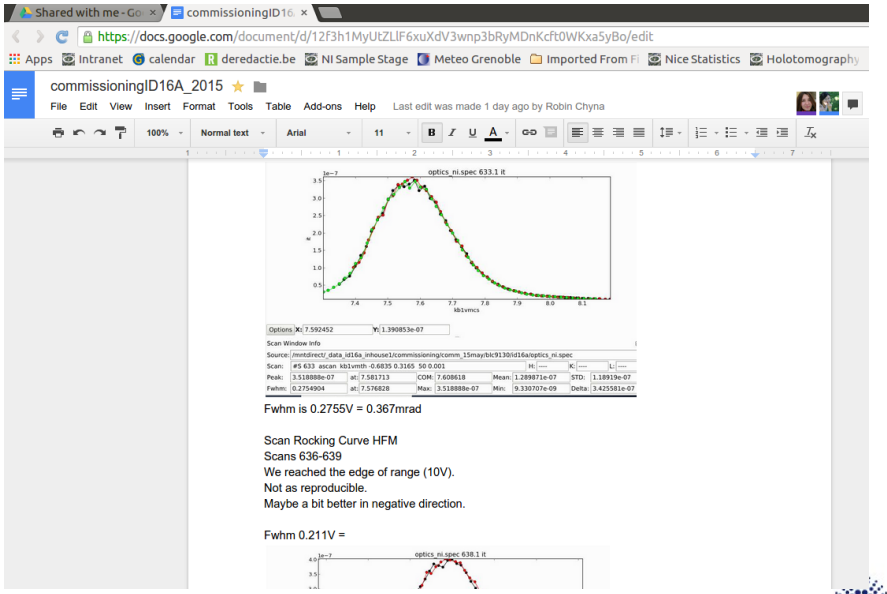
- **Follow the Nexus conventions when defining Dataset Parameters :**

  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

AndyGotz Talk Sandbox Preferences Beta Watchlist Contributions Log out

Visit the main page

WIKIPEDIA
The Free Encyclopedia

Article Talk Research Read Edit View history More Search

Herschel Space Observatory

Datasets in relation

- HESIOD : The Herschel IdOc Database is delivering photometric maps and spectral cubes from the PACS and SPIRE instruments (IR domain), reprocessed at IAS with the latest ESA pipelines and with high level customized pipelines. Virtual Observatory compatible. (source)

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

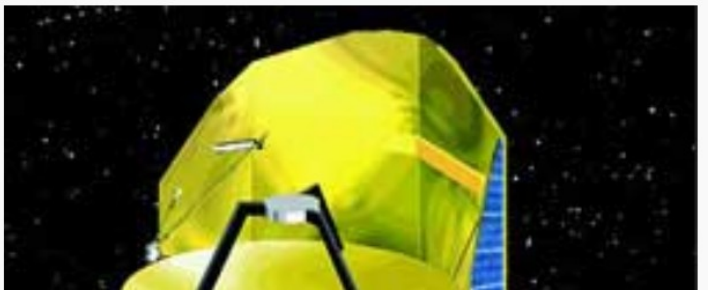
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

From Wikipedia, the free encyclopedia

This article is about the space telescope. For the ground-based telescope, see William Herschel Telescope.

The **Herschel Space Observatory** was a space observatory built and operated by the European Space Agency (ESA). It was active from 2009 to 2013, and was the largest infrared telescope ever launched,[5] carrying a single 3.5-metre (11.5 ft) mirror[5][6][7][8] and instruments sensitive to the far infrared and submillimetre wavebands (55–672 μm). *Herschel* was the fourth cornerstone mission in the ESA science programme, along with Rosetta, Planck, and Gaia. NASA is a partner in the *Herschel* mission, with US

Herschel Space Observatory



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**

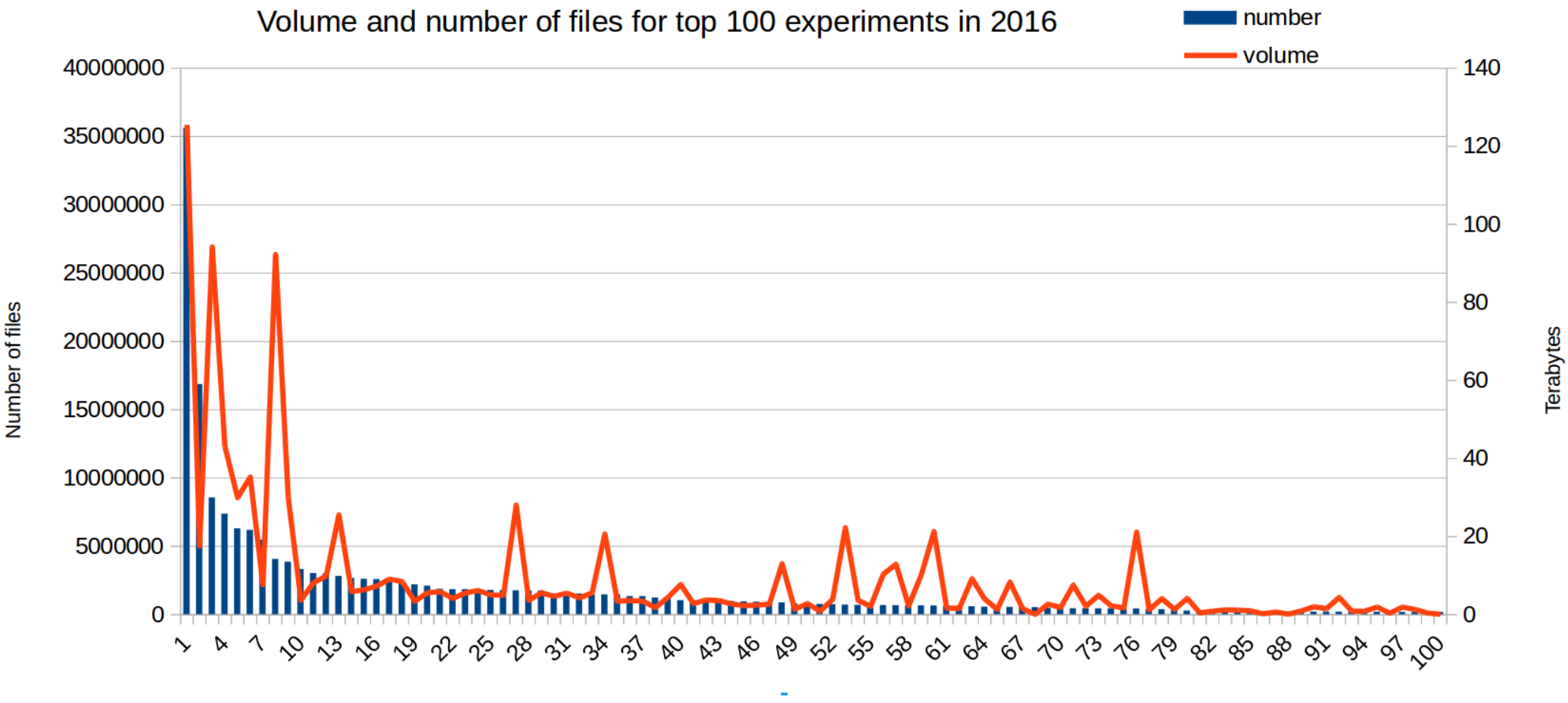
LONG TERM ARCHIVING (LTA) – 2 COPIES

Time Navigator(TiNa)

Experimental Hall 2
Central building
Experimental Hall 1
BM26 A+B
BM25 A+B
ID26
ID27
ID28
BM28
ID29
BM29
ID30A+B
BM30 A+B
Laboratory and office building
ID24
BM23
ID23
ID22
ID21
ID19
BM20
ID20
ID31
BM32
ID32
Control room
BM01 A+B
ID01
ID17
ID18
Booster Synchrotron
BM02
ID03
ID02
ID16 A+B
BM05
ID06
ID15A+B
BM14
ID14
ID13
Storage ring
ID12
ID11
ID10
ID09
BM08

- Refurbished or upgraded beamlines
- ESRF beamlines
- CRG beamlines
- Instrumentation and machine test beamlines

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