# Data Management at ILL.

# Experimental data evolution and its consequences.

Presentation given at the CREMLIN workshop on Big data Management

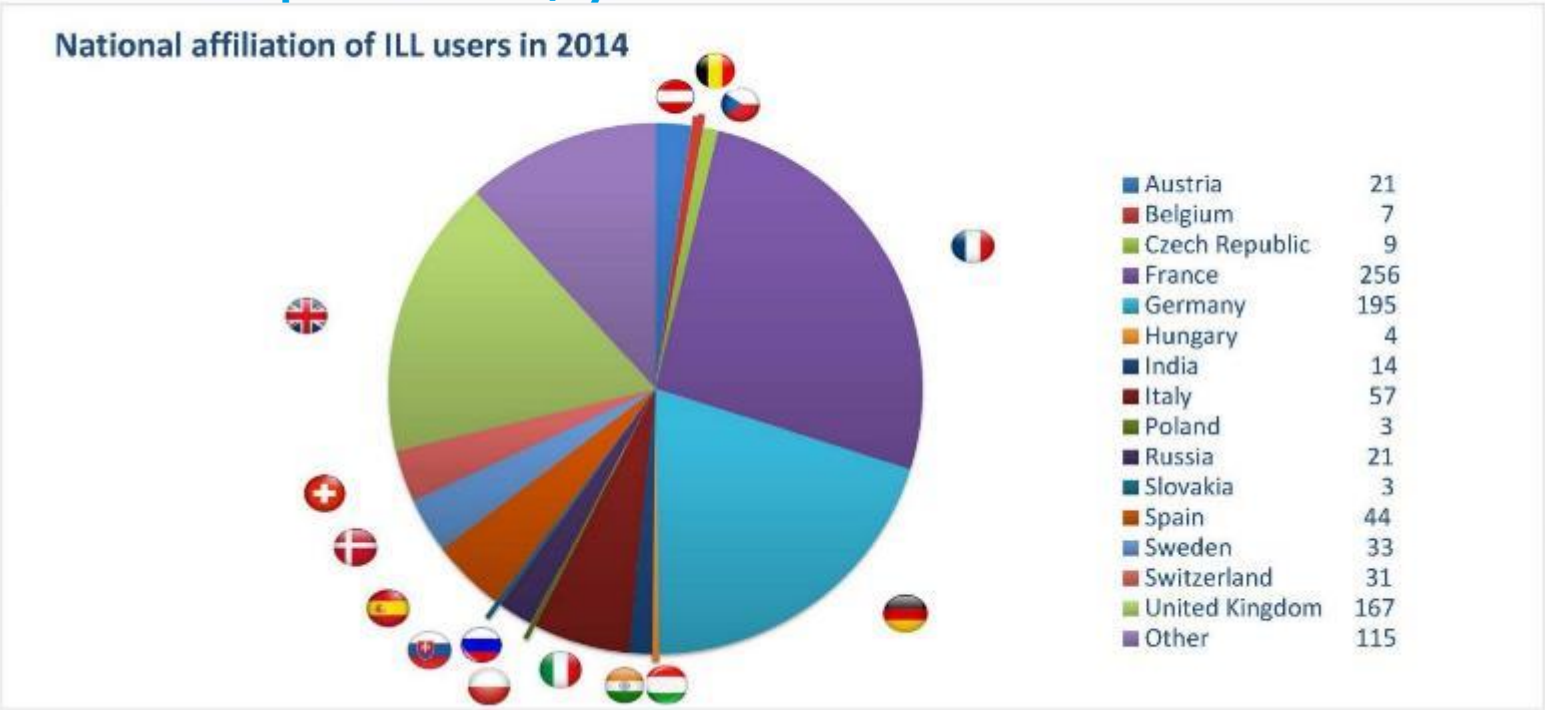
16th of Feb 2017 NRC “Kurchatov Institute” - Jean-François Perrin (ILL IT Services)

# Short introduction to the ILL

# Users

**An active community of 12 000 scientists from 28 countries**

**1400 invited experimenters / year**

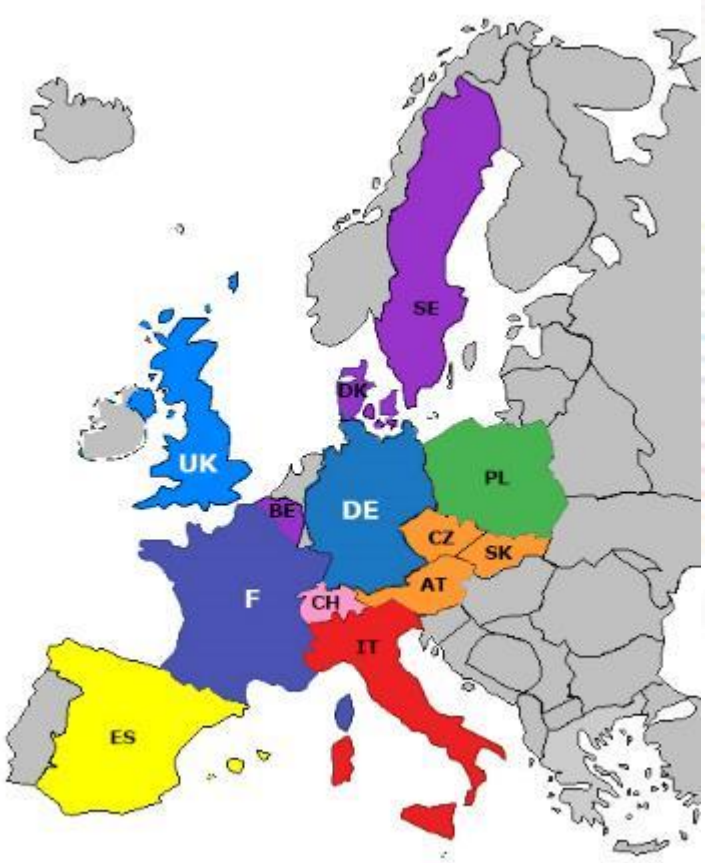



# Users are travelling

Snapshots of unique X-Ray & Neutron scattering users, over 18 months (2013 – 2014).

| | ALBA | BERII | BESSYII | DESY | DLS | ELETTRA | ESRF | MLZ | ILL | ISIS | LCLS | SINQ | SLS | SOLEIL | SNS | neutron | photon | all |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALBA | 1303 | 5 | 43 | 90 | 274 | 128 | 356 | 8 | 83 | 36 | 2 | 23 | 124 | 161 | 7 | 122 | 679 | 1303 |
| BER II | 5 | 237 | 27 | 28 | 20 | 0 | 42 | 66 | 104 | 50 | 0 | 75 | 22 | 7 | 37 | 162 | 86 | 237 |
| BESSY II | 43 | 27 | 838 | 128 | 96 | 95 | 143 | 16 | 31 | 16 | 28 | 29 | 119 | 100 | 11 | 76 | 418 | 838 |
| DESY | 90 | 28 | 128 | 3680 | 396 | 255 | 901 | 110 | 167 | 92 | 151 | 82 | 326 | 246 | 63 | 343 | 1579 | 3680 |
| DLS | 274 | 20 | 96 | 396 | 10445 | 297 | 1606 | 82 | 485 | 763 | 70 | 144 | 559 | 526 | 124 | 1136 | 2598 | 10445 |
| ELETTRA | 128 | 0 | 95 | 255 | 297 | 3422 | 480 | 21 | 99 | 41 | 68 | 14 | 218 | 379 | 12 | 149 | 1171 | 3422 |
| ESRF | 356 | 42 | 143 | 901 | 1606 | 480 | 10786 | 203 | 731 | 356 | 102 | 203 | 899 | 1390 | 155 | 1165 | 4242 | 10786 |
| MLZ | 8 | 66 | 16 | 110 | 82 | 21 | 203 | 1430 | 409 | 167 | 3 | 222 | 52 | 46 | 158 | 601 | 353 | 1430 |
| ILL | 83 | 104 | 31 | 167 | 485 | 99 | 731 | 409 | 4138 | 606 | 3 | 384 | 130 | 239 | 316 | 1252 | 1304 | 4138 |
| ISIS | 36 | 50 | 16 | 92 | 763 | 41 | 356 | 167 | 606 | 3406 | 9 | 236 | 101 | 84 | 267 | 891 | 1052 | 3406 |
| LCLS | 2 | 0 | 28 | 151 | 70 | 68 | 102 | 3 | 3 | 9 | 1123 | 1 | 79 | 44 | 6 | 17 | 329 | 1123 |
| SINQ | 23 | 75 | 29 | 82 | 144 | 14 | 203 | 222 | 384 | 236 | 1 | 1424 | 250 | 65 | 185 | 614 | 501 | 1424 |
| SLS | 124 | 22 | 119 | 326 | 559 | 218 | 899 | 52 | 130 | 101 | 79 | 250 | 3981 | 366 | 64 | 365 | 1637 | 3981 |
| SOLEIL | 161 | 7 | 100 | 246 | 526 | 379 | 1390 | 46 | 239 | 84 | 44 | 65 | 366 | 5134 | 40 | 349 | 2145 | 5134 |
| SNS | 7 | 37 | 11 | 63 | 124 | 12 | 155 | 158 | 316 | 267 | 6 | 185 | 64 | 40 | 3723 | 581 | 327 | 3723 |

# ILL member countries and their financial participation




Germany : 25 %
UK : 25 %
France : 25 %

Spain

Italy

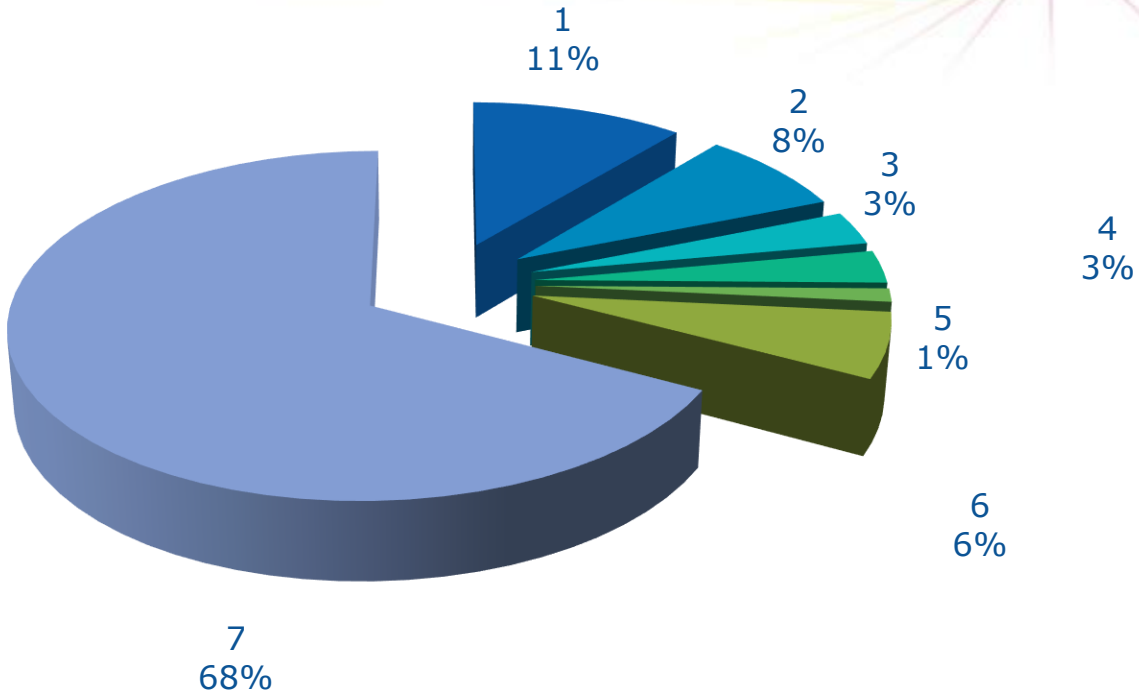
Switzerland

Poland

CENI (Central European Neutron Initiative, Austria, Czech Republic, Slovakia)

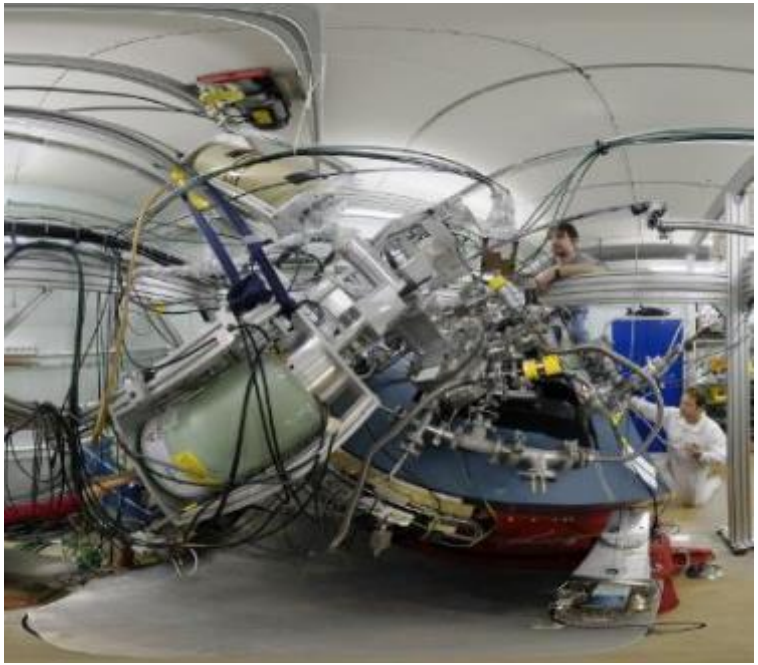
TRANSNI (TRANSnational Neutron Initiative, Belgium, Denmark, Sweden)

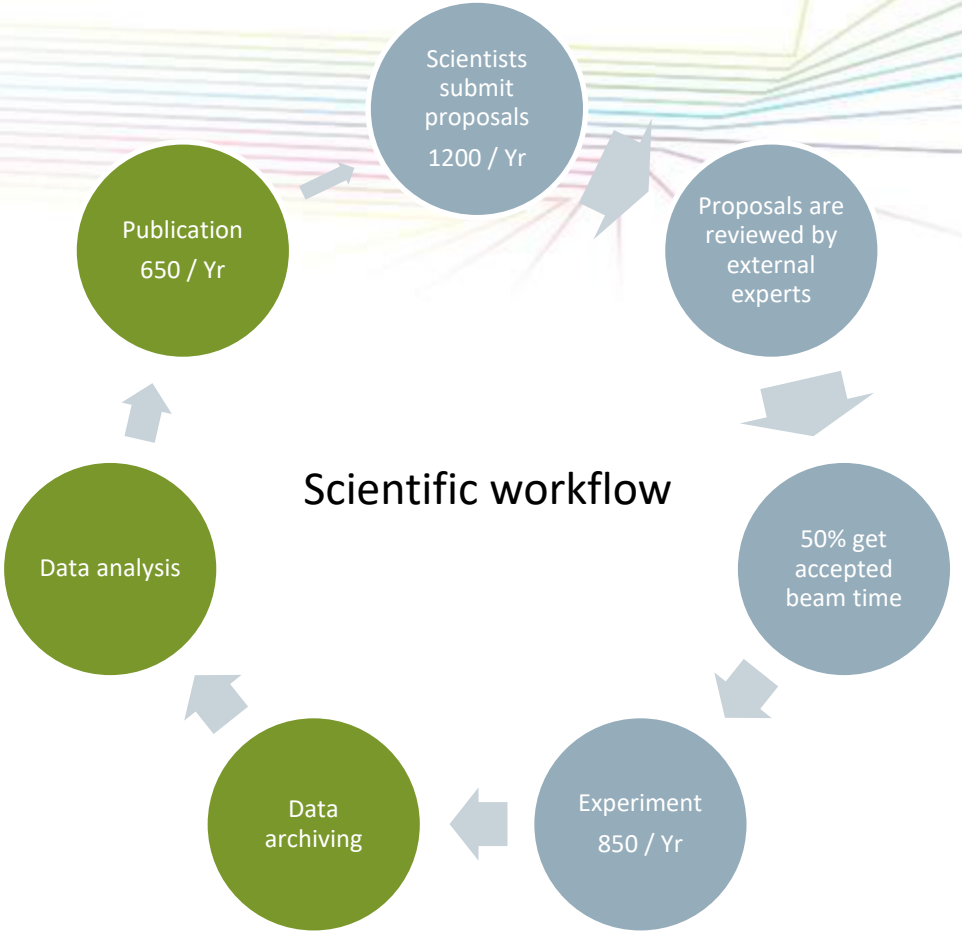
# Societal impact



# Science

## 28 instruments + 10 CRG





# Instuments

Organised in 4 groups

- Spectroscopy:
  - Time-of-flight spectrometers
  - Backscattering spectrometers
  - Spin-echo spectrometers
  - Three-axis spectrometers
- Diffraction:
  - Powder diffractometers
  - Single-crystal diffractometers
- Large scale structures:
  - Diffractometers
  - Reflectometers
- Nuclear and particle physics

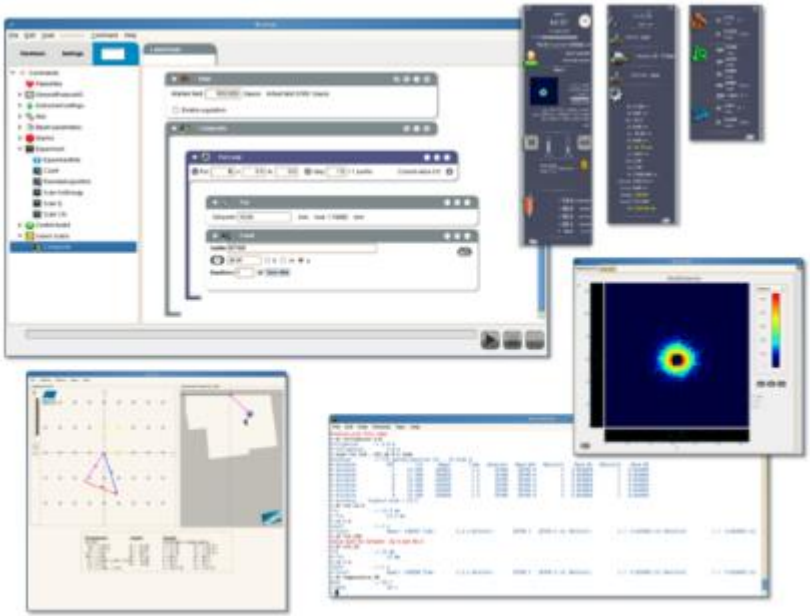
IN5 -©2011, Ecliptique

D11 -©2009 Ecliptique / L. Thion

# Instrument control

Under the responsibility of DPT/SCI

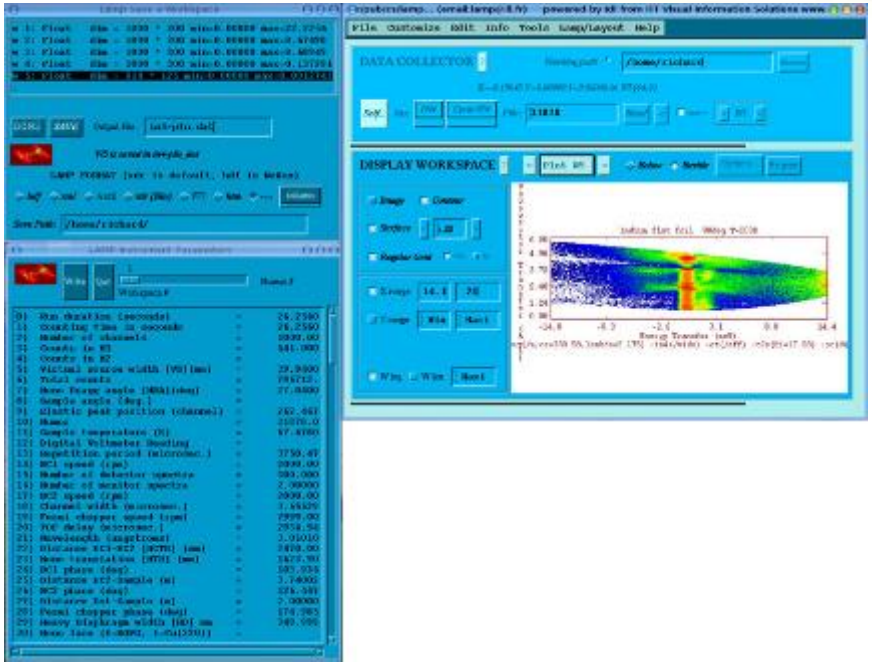
## NOMAD is the ILL's sequencer to control instrument operations.



- Allow to control all operations on the instruments.
- Users can build their specific workflows (GUI/drag & drop or CLI/scripting approach)
- Client/server architecture (Java/C++)
- Open sourced under EUPL
- https://www.ill.eu/instruments-support/instrument-control/software/nomad/

# Data reduction (current situation)

## Under the responsibility of DS/CS

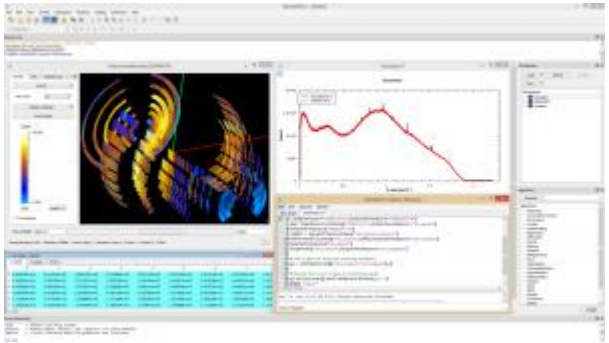



- mostly based on LAMP (http://www.ill.eu/data_treat/lamp)
  - Based on IDL language
  - Reliable and long experience (1996 – 2017)
  - Covers most of the technics
  - Public domain
- Nuclear and Particle Physics have their dedicated software (root, …)
- Some instruments have their dedicated solution.

# Data Reduction (future)

Under the responsibility of DS/CS

- Mantid (http://www.mantidproject.org)
  - Large worldwide collaboration between Neutron RI: ISIS, SNS & HFIR, ESS, ANSTO, PSI, ILL
  - Mantid was developed for spallation sources. ILL is investing 3-4 FTE over 3 yrs to adapt Mantid to the ILL (reactor source) instrument suite. https://www.ill.eu/fr/instruments-support/computing-for-science/data-analysis/mantid-tutorial-ill-resources/
  - For more info, watch Jon presentation.



# Data Analysis

Under the responsibility of DS/CS

- A large set of existing tools
  - http://www.ill.eu/instruments-support/computing-for-science/cs-software/all-software
  - https://www.ill.eu/sites/fullprof/
  - …
- WP10 of SINE2020 project

| Tech. | Software | Lead (+co.) | Improvements |
|---|---|---|---|
| Imaging | MuhRec & KipTool | PSI (+CEA, ISIS, ESS) | Conversion to open source software. More user-friendly interfaces, optimized algorithms for GPU and distributed computing for faster analysis, and new reconstruction algorithms. |
| Reflectometry | BornAgain | FZJ | Addition of GUI, extension to all types of reflectometry, and algorithms optimized for real time analysis. |
| SANS | SASView | ESS (+PSI) | Modularization and new GUI. Addition of API and CLI. Optimization of algorithms for real time analysis and extension with SASFit model fittings. |
| QENS | Mantid | STFC (+ILL, FZJ) | More user-friendly interfaces and extension of model fitting functions. |
| Atomistic modelling | nMoldyn, DFT | ILL (+ESS, ISIS, PSI, UNIPR) | Extension to convert lattice dynamics and Monte Carlo simulations to scattering curves. Development of muon spectroscopy as a complementary tool for neutron scatterers through improved data analysis exploiting atomistic modelling such as density functional theory (DFT). |

**SINE2020 WP10 aims**

- Convergence to a common set of supported tools by Ris
- Straightforward generation of scientific results for non-expert and industry users.
- Data treatment software ready for users at ESS

# Experimental data management

Co-funded by the European Union :
PaNData-Europe Grant Agreement No 261537
PaNData-ODI Grant Agreement No RI-283556

# Open Data/Science for a facility?

- Data are the real/factual production.
- Knowledge (peer-reviewed articles, conference contributions, thematic courses, software …) is the main output.

Openness is a tool for increasing our **impact**.

# What has been done so far?

- 2008 1st discussion on Data Policy (PaNData)
- 2011 “Open” DP published - 3 (max 5) years embargo
- 2012 1st experiment under DP
- 2013 complete set of Data Management Services available for users: search, access, annotate, archive, identify, publish, …
- since then, communication with our users …

# 1) Data Policy

Based on the PaNData framework

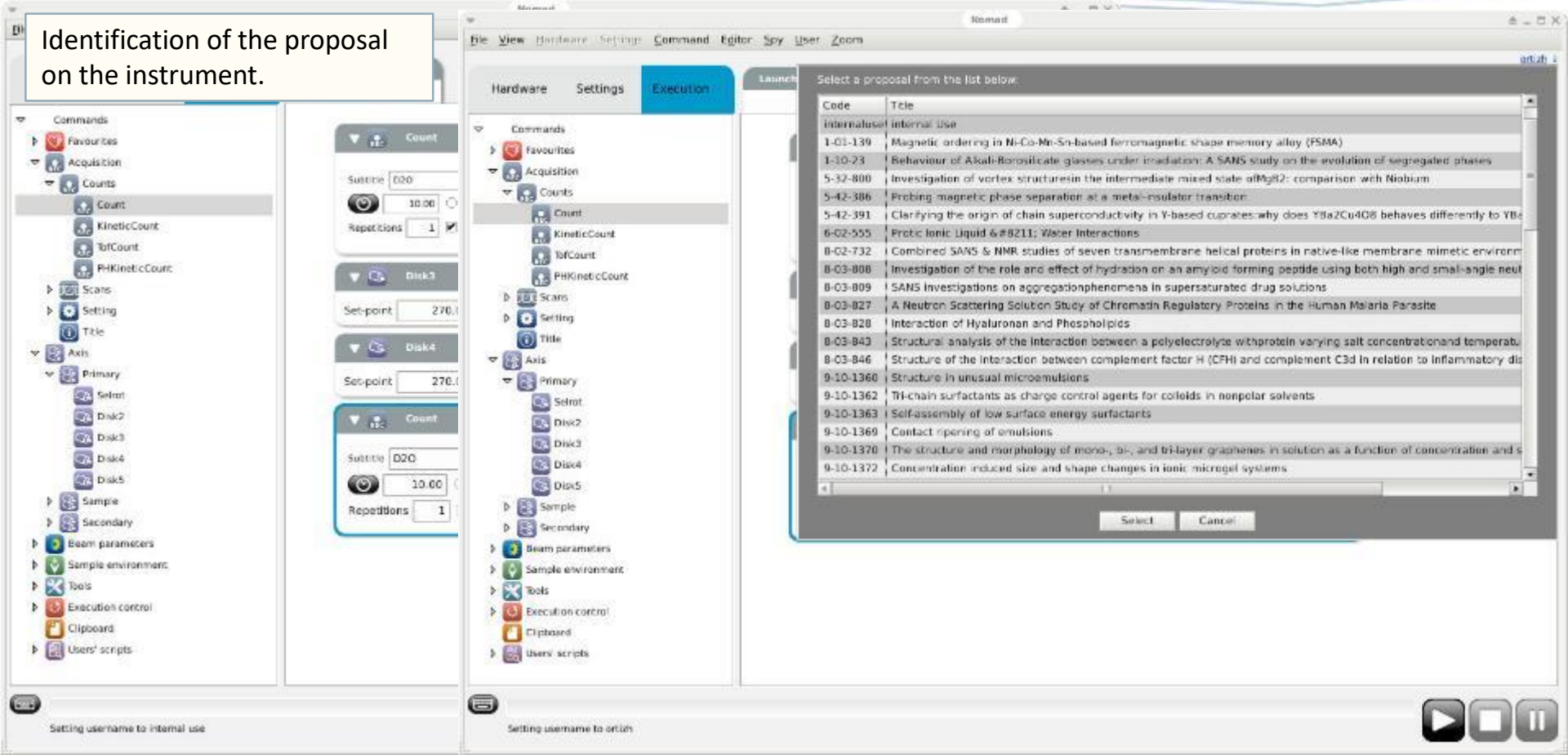
## Open data & how to protect and credit our users?

- The facility shall act as a custodian for the data.
- All raw data will be curated in a well-defined format with a unique ID (DOI).
- Metadata is captured automatically and resides either within the raw data files, and/or in an associated on-line catalogue.
- Users can release or give access to their data at any time, by default access to raw data and the associated metadata is restricted to the experimental team for a maximum period of 3 years. Thereafter, it will become publicly accessible.
- The embargo period can be extended on request to the ILL management.
- Publication based on data must acknowledge the source of the data and cite its unique identifier (CC-BY licence).

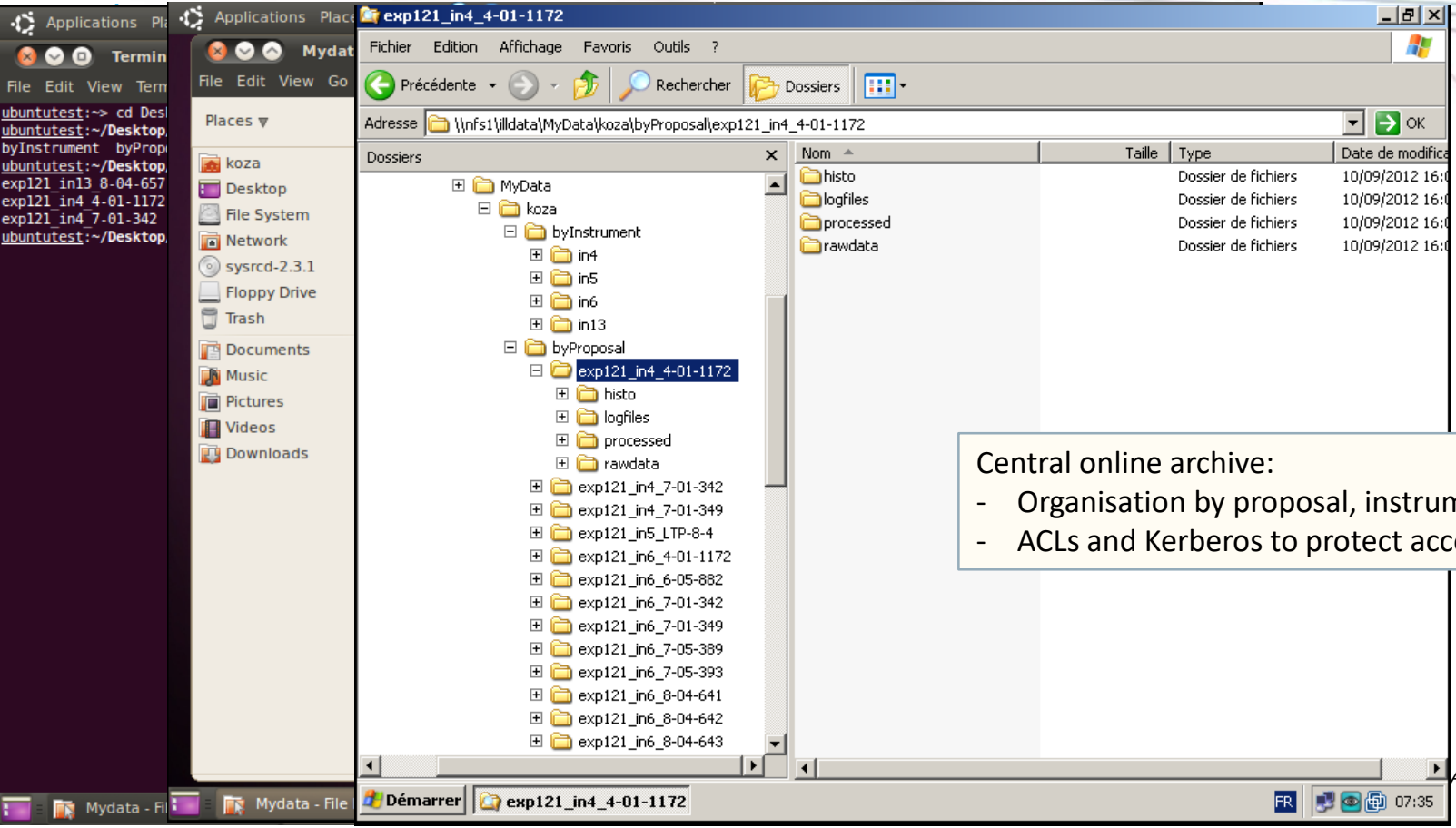
https://www.ill.eu/DataPolicy

# 2) Linking Proposal and data.

Identification of the proposal on the instrument.



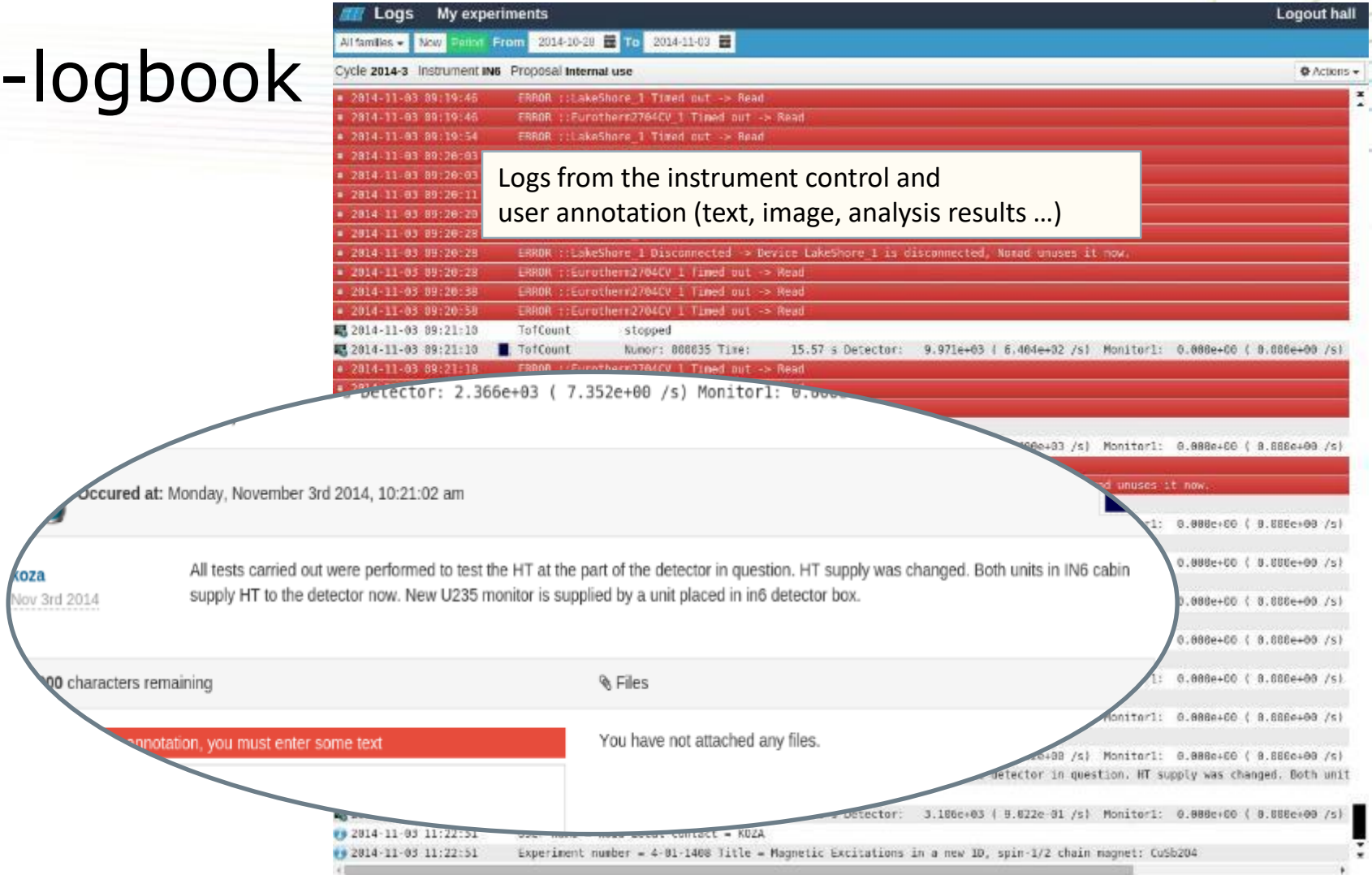
# 3) Archive: ACLs & user experience improvements




Central online archive:

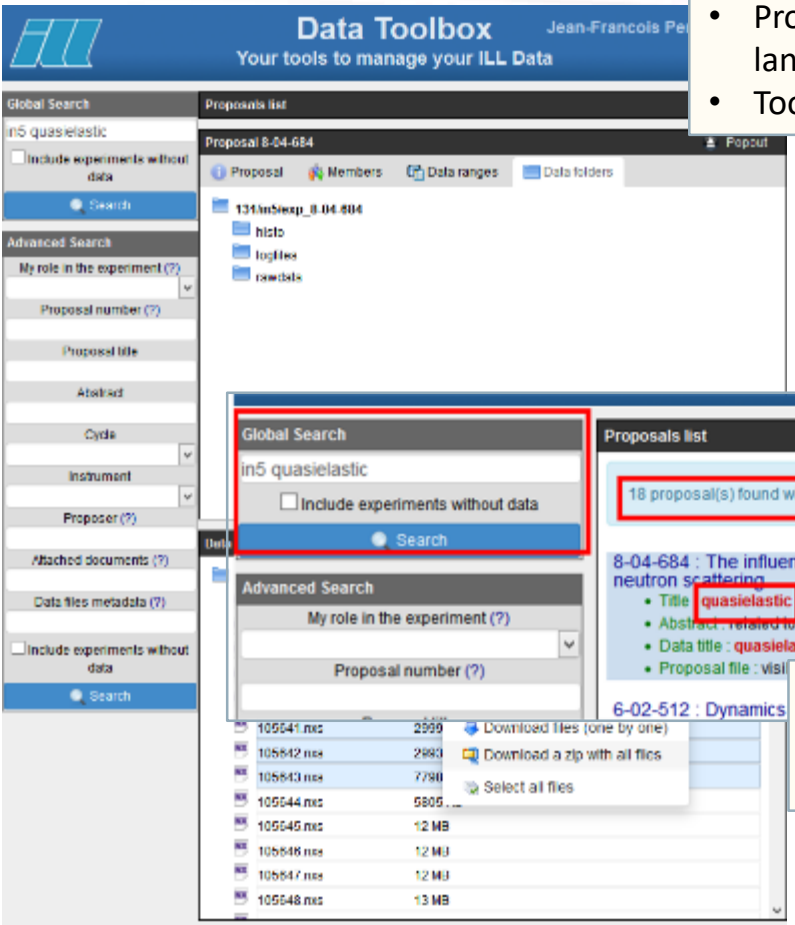
- Organisation by proposal, instrument, dates.
- ACLs and Kerberos to protect accesses.

# 4) E-logbook

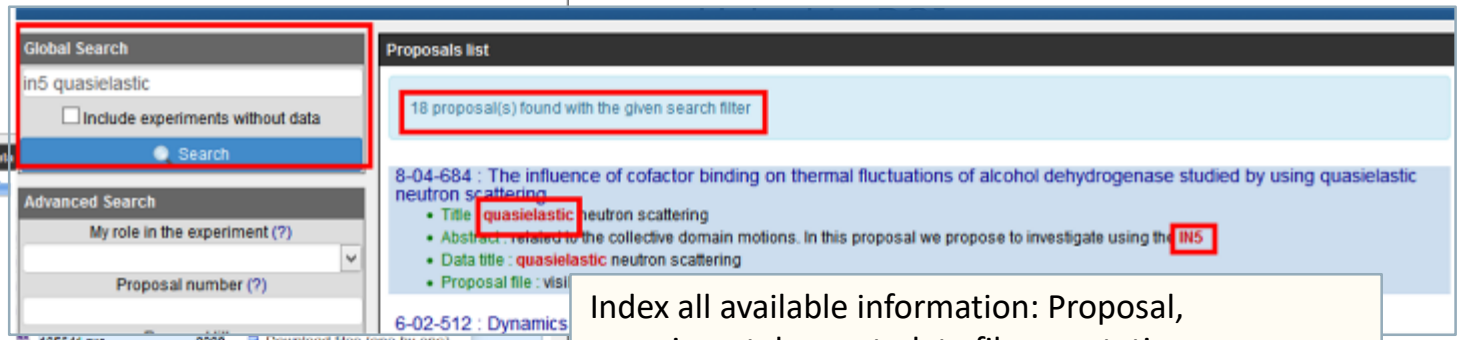

Logs
My experiments
Logout hall
All families
Now
Period
From
2014-10-28
To
2014-11-03
Cycle 2014-3
Instrument IN6
Proposal Internal use
Actions
Logs from the instrument control and
user annotation (text, image, analysis results …)
Occured at: Monday, November 3rd 2014, 10:21:02 am
koza
Nov 3rd 2014
All tests carried out were performed to test the HT at the part of the detector in question. HT supply was changed. Both units in IN6 cabin supply HT to the detector now. New U235 monitor is supplied by a unit placed in in6 detector box.
characters remaining
Files
You have not attached any files.

# 5) Data portal

- Provide access to data, meta-data, logs, DOIs landing page, …
- Tools for managing data authZ



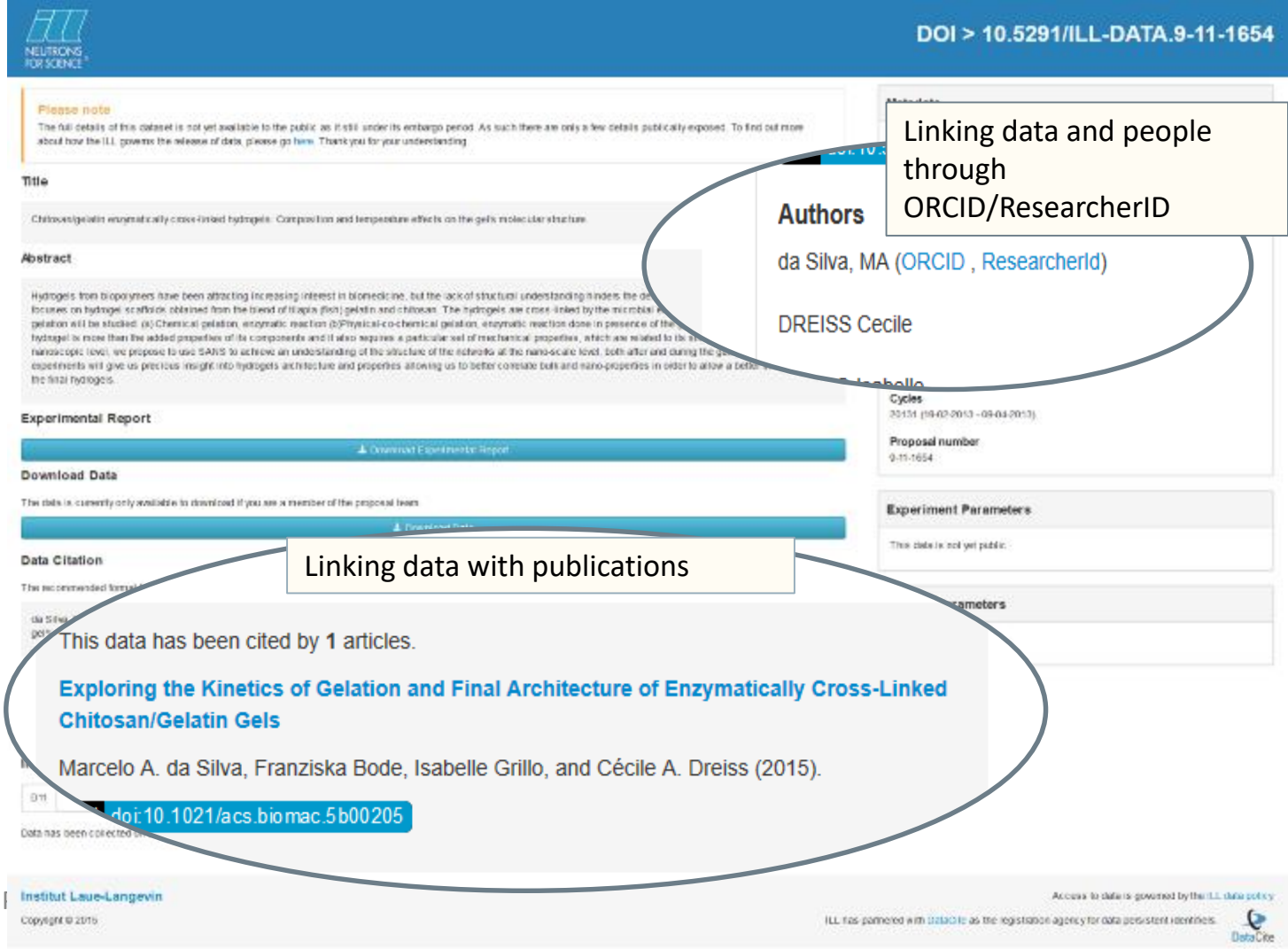

- Tailored to ILL needs
  - User management of data access authorization.
  - Users could decide to publish (open access) their data, before the end of the embargo period.




Index all available information: Proposal, experimental report, data file annotation, publications, …

# 6) DOIs

Collaboration with DataCite
– INIST (French rep)



# Issue #1: Awareness of the scientists

- This is still new for most of the scientists
  "What are DOIs? What are you talking about?"
- We currently feel a bit alone – critical mass. (ESRF, PSI, ESS, have recently joined)
- We need more communication – mentoring – cultural change - education.

Need to fill the gap between what we hear in RDA-like meetings and the daily reality of the scientists.

Still need to convince the scientist that a change is happening regarding experimental data.

# Issue #2: Difficulty to collect the articles exploiting the experimental data

- Technical reason : DOIs in figures instead of references, partial citations …
- No tools yet available to easily collect references
  - CrossRef cited by linking - currently only for article (vs data) publishers ? -, OpenAire.
  - This is a business for the publishers.
- Difficulties to get metrics: how successful are we?
  - We have currently (Dec 2016) collected less than 50 peer reviewed article referencing the data DOI.
  - How many are we missing?

Need to access freely information for building metrics.

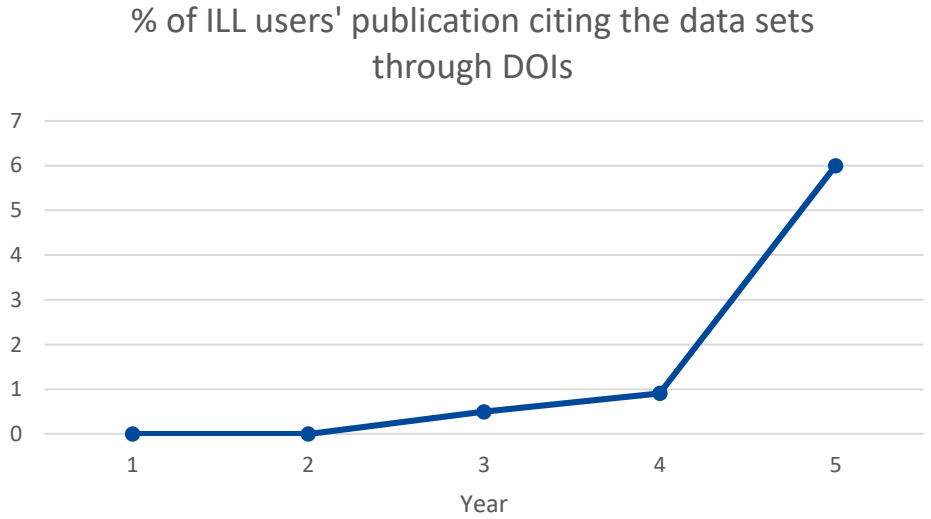
# Issue #3: time

- Time for analyses
- Time for writing articles
- Time for publication
- Time for convincing

This is by nature a long process, but seeing the level of investment needed, we need to convince, we need evidence of success urgently.

# Results as of Dec 2016

- Few data sets have been made public by users before the end of the embargo period.
- The reference to Data sets in scientific articles, through DOIs, is recently improving.
- Real interest of the publishers http://www.elsevier.com/?a=57755
- More user feedback: "Why I don't get a DOI for experiment XYZ?"

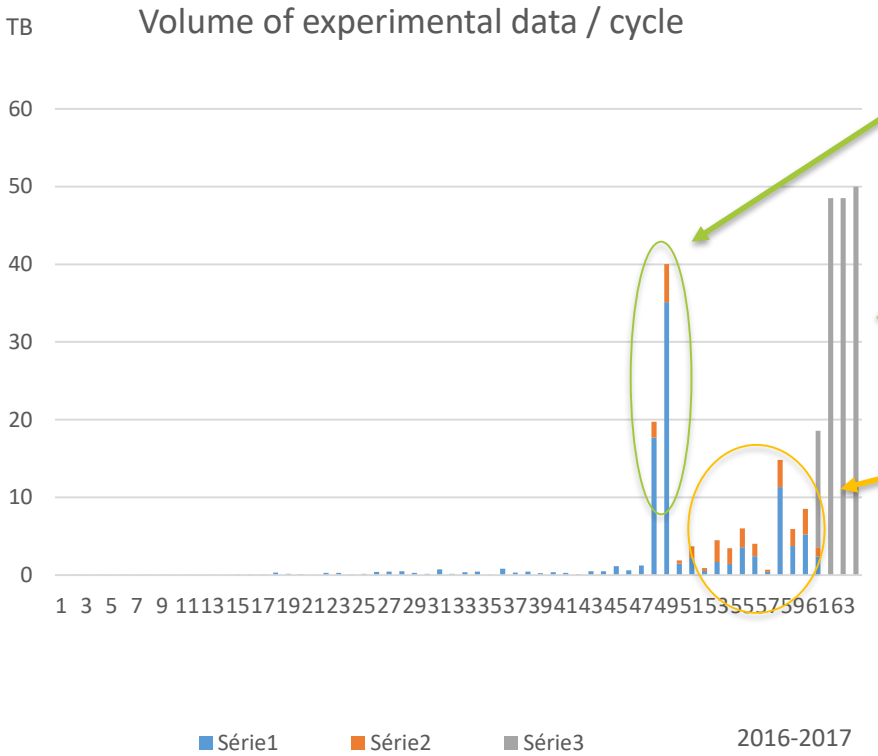


# More generally

- Better tools for our users
  - Following remote experiment: live remote access to data, logs, …
  - Easier access/discovery to data sets.
- Better archiving - from ‘bit level’ responsibility to ‘usability level’.
- New services for users:
  - Ready for Open Data - Some journals (e.g. PLOS) request access to data for every article or for referees.
  - Data Management Plan (more and more mandatory for grant request)

# Data Analysis As a Service.

# Data volume evolution




Evaluation of new detectors leading to permanent instruments starting from Dec 2016.

Moving to list mode (vs Histo)

# Impacts of the data volume evolution

Example of the EXILL campaign

- Storage (2 experiments = 70TB)
  - ILL archive capacity & performance
  - Users' storage becoming almost impossible
- Moving data
  - Today how to carry 40TB to 10 different labs?
  - Why carrying them?
- Analysis
  - Almost impossible in most users' labs with such data sets.
- But
  - 32 direct (h-index 4) peer reviewed articles published
  - 2 Phd-thesis
  - 10+ international conferences
  - …

# Our vision

- Large raw data sets should be archived at the source (ILL in our case).
- Provide remote analysis infrastructure – data and analysis capacity should be collocated.
- Preserve data and the scientific workflow.
- Most of the analytical facilities face the same problem. We share a large part of our User community. We need to work together: **PaNDaaS**

# Data analysis as a Service

- The aim is to proposed to users to access analysis services (**data, software, IT capacity and expertise**) remotely using standard tools (ideally only web browser).

- Typical workflow:

  1) The user connects remotely using his web browser and its credentials (Federated IM)

  2) Then select one of the experiment he has performed in the list.

  3) he is then connected to a service where the necessary analysis applications have been installed and configured for accessing directly the experimental data.

  4) If necessary he could receive help and support from facility expert, during the analysis.

# Benefits

- Provide a user friendly environment (most of or users are not expert neither in data treatment, neither in IT and some have no home IT support).
- Accelerate the analysis process, ease collaboration during analysis.
- Solve the problem of transport of experimental data to home labs.
- Move the work from 'software installation' to 'scientific analysis‘.
- Authorize the preservation of the full workflow.

# Status

- PaNDaaS was not funded
- Coordination meeting between RIs have started to take place (ESRF organisation)
- Work is ongoing mostly with RI budget at the pace allowed by RIs capacity.
- @ILL first users (IN5 instrument) expected for Mid 2017.
- EOSC has the solution?

Contact: data@ill.eu
Portal: https://data.ill.eu
Policy: https://www.ill.eu/DataPolicy
PaNData Collaboration: http://pan-data.eu

NEUTRONS FOR SOCIETY®
71, rue des Martyrs - 38000 Grenoble - **www.ill.eu**