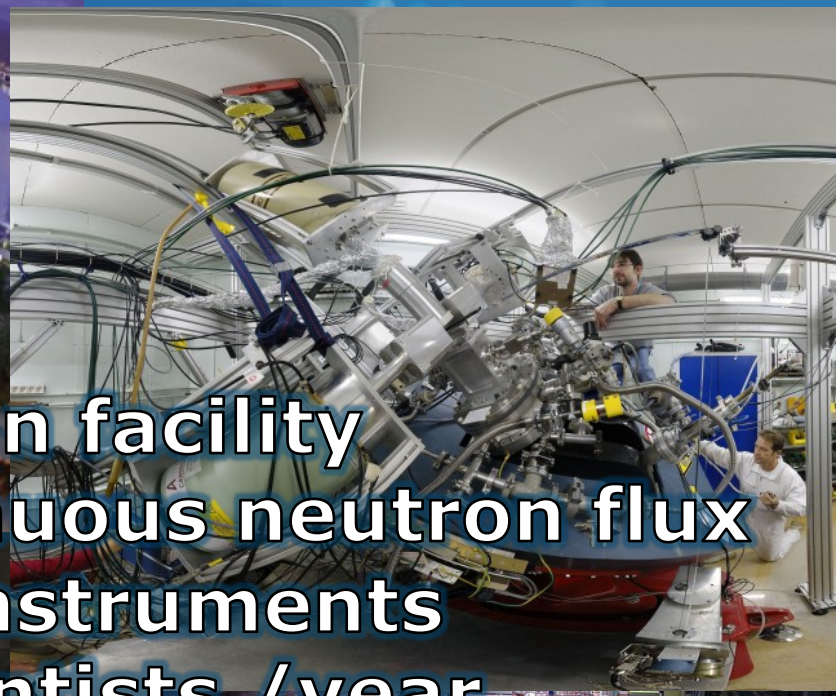
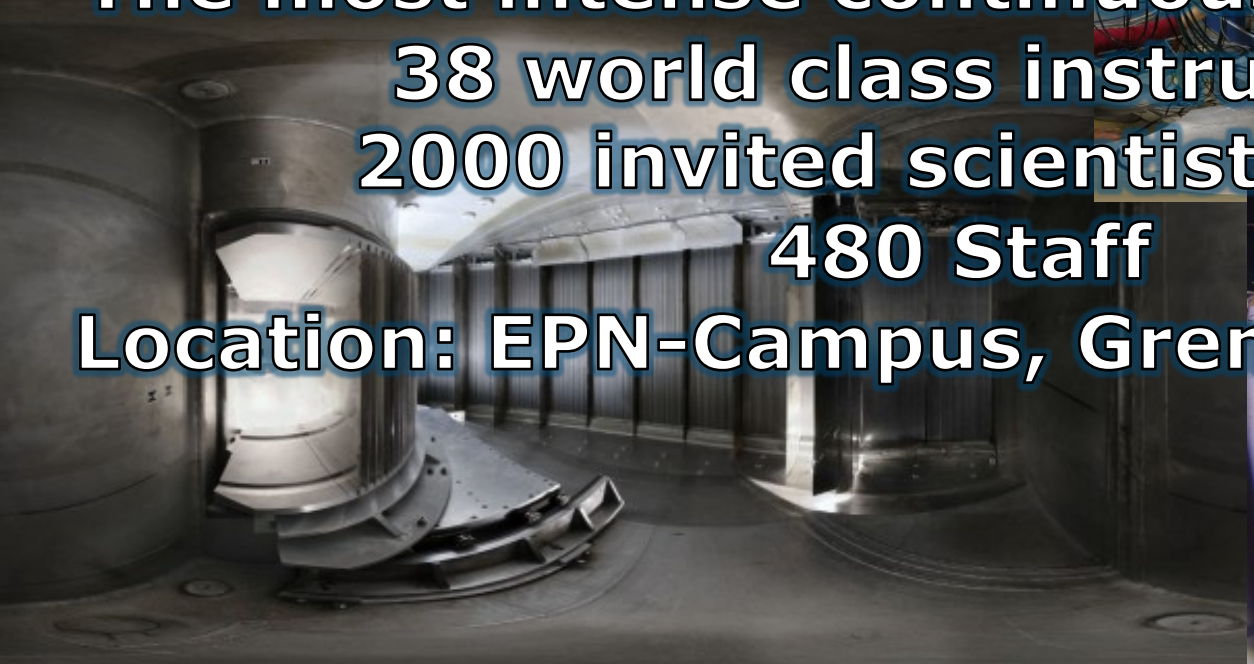
Big data & Open Data @ ILL

Jean-François Perrin (ILL) - 8th of May 2014 - ERF Workshop BigData & OpenData

Who are we ?

NEUTRONS FOR SCIENCE



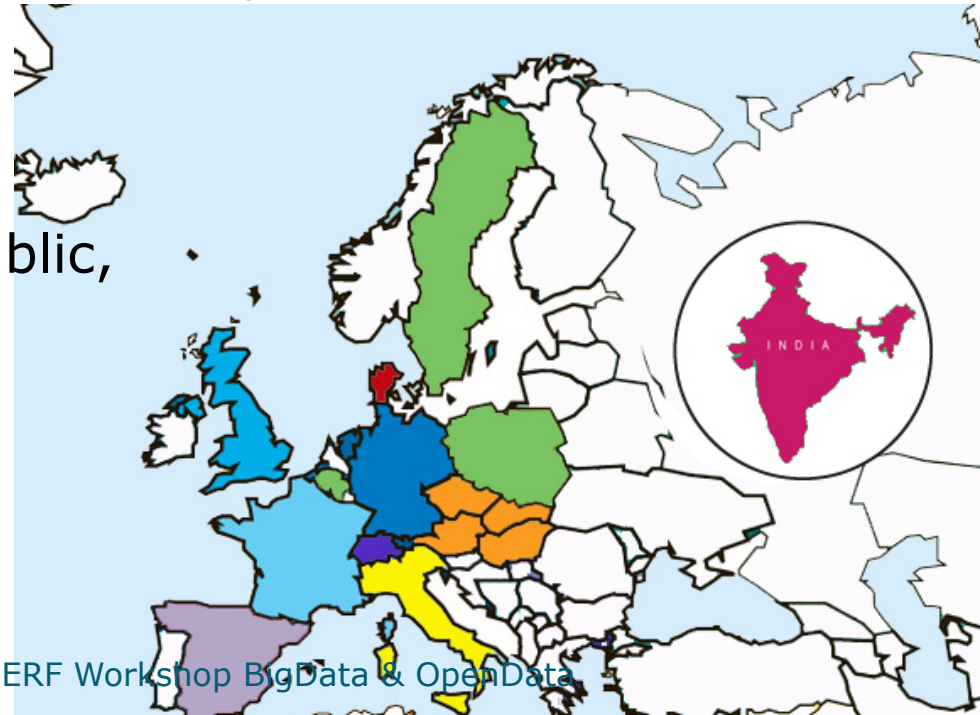


ILL is a neutron facility
The most intense continuous neutron flux
38 world class instruments
2000 invited scientists /year
480 Staff
Location: EPN-Campus, Grenoble (France)

An international scientific collaboration

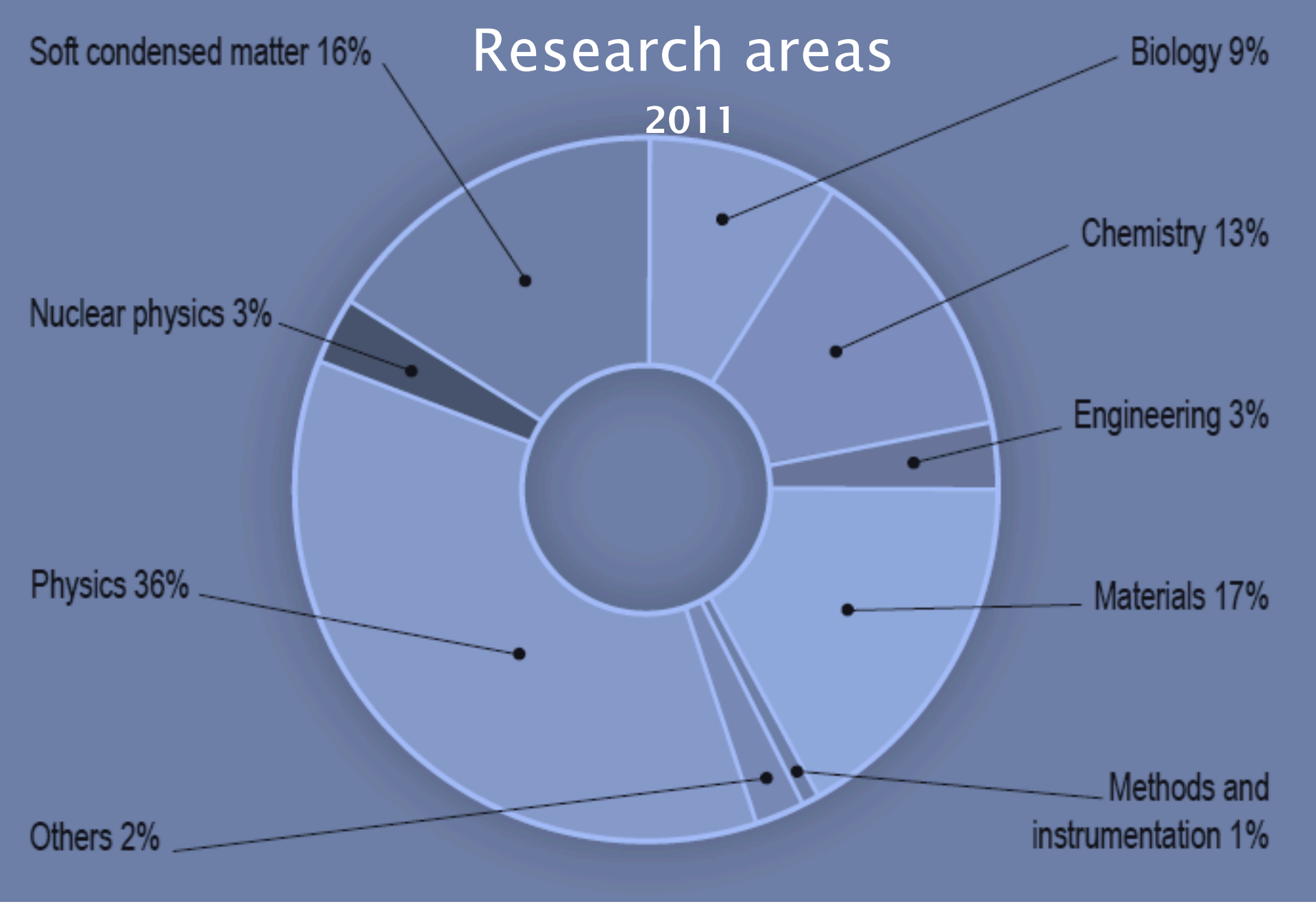
- Founded in 1971 by France, Germany and United Kingdom.
- Scientific partners that have joined in since then:

Spain, Switzerland, Austria, Denmark, Italy, Czech Republic, Sweden, Hungary, Belgium, Slovakia and India.



Research areas

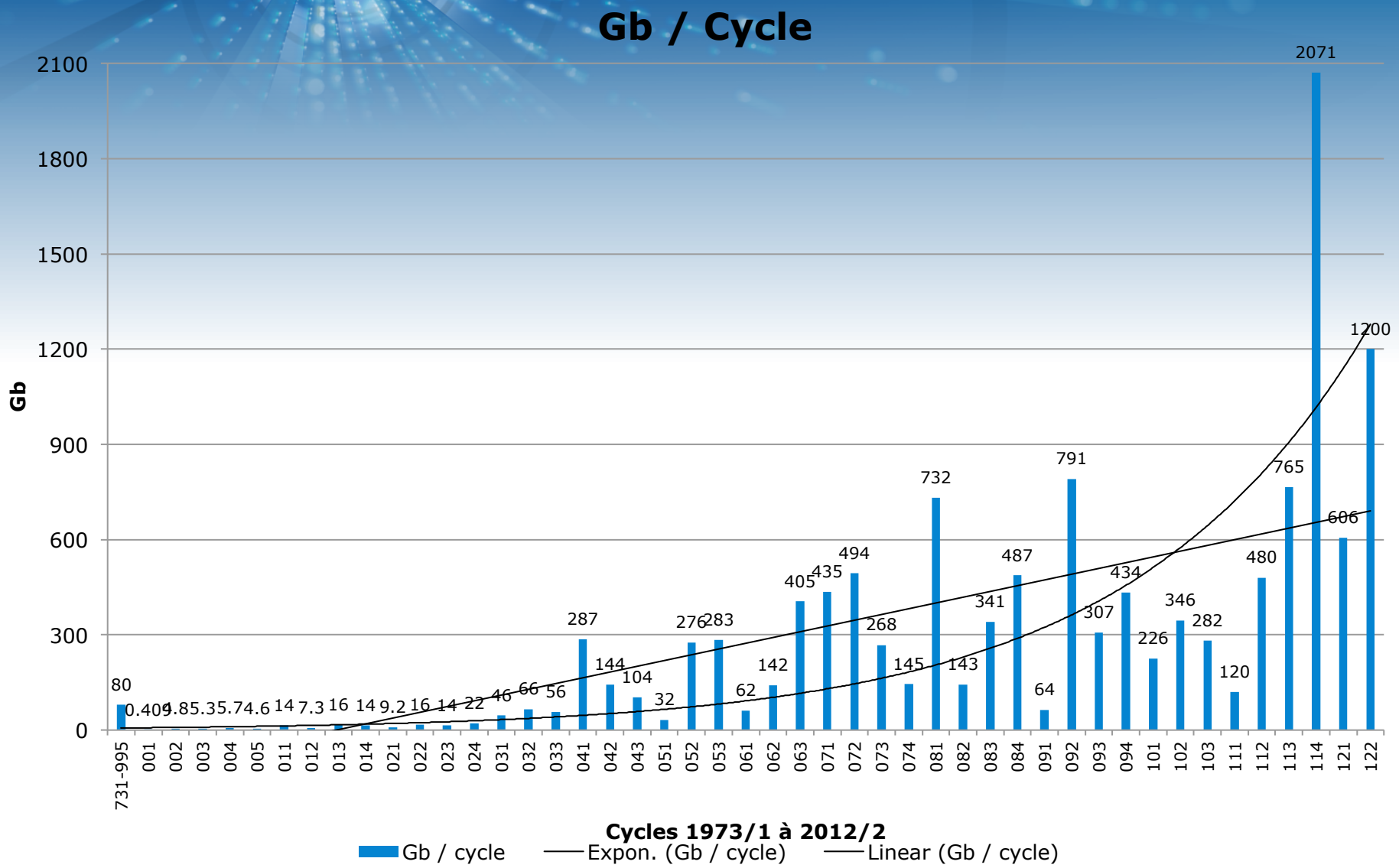
2011



Big DATA

ILL definition: sufficiently large increase in the volume of the experimental data sets to produce a breakthrough in the scientific workflow and question our models.

Experimental raw data 1973-2012



Experimental raw data 1973-2013

Impacts of the “data deluge”

- Storage
  - ILL archive capacity & performance
  - Users’ storage becoming almost impossible
- Moving data
  - Today how to carry 40TB?
  - Why carrying them?
- Analysis
  - Almost impossible in most users’ home labs.

Our vision

- Large raw data sets should stay and be archived at the source (ILL in our case).
- We need to provide remote analysis infrastructure on premises.
- Need to preserve data and the scientific workflow.
- The federation model is more suitable than the centralised one.

Needs for a remote analysis facility

- The aim is to propose to users to access workstation or analysis application already set up for their data remotely using standard web browsers.

- Typical workflow:

  1) The user connects remotely using his web browser and its credentials (preferably FIM like UmbrellaID.org).

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

  3) He is then connected to a computer 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.

  5) Archive the analysis and the process.

Benefits

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

Data management & Open Data

Jean-François Perrin (ILL) - 8th of May 2014 - ERF Workshop BigData & OpenData

Improving data archiving and access to data

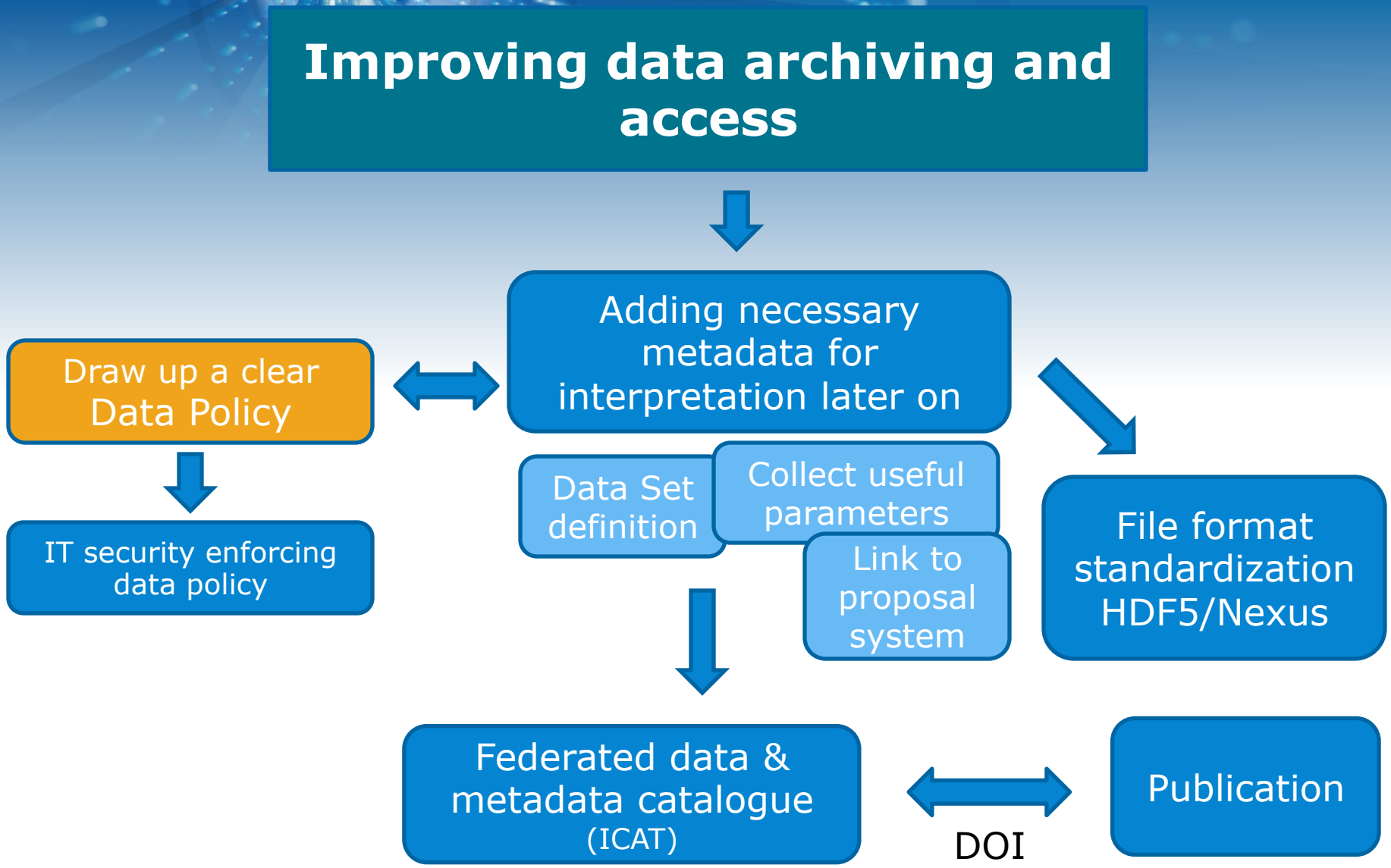
"ILL raw data is available online since 1973, immediately after the experiment.

Is it really valuable ?"

Question raised to our stakeholders in the late 2000's

- Yes for the experimental team
- No for the others (not shareable/bits archive)

The project: global picture



User community

Number of Users shared between facilities																
	BER II	BESSY II	DESY	DLS	ELETTRA	ESRF	ILL	ISIS	LLB	SINQ	SLS	SOLEIL	FRM-II	ANKA	neutron	photon	all
BER II	2761	284	152	54	39	319	556	302	156	130	29	38	207	15	814	657	2761
BESSY II	284	7068	487	110	298	810	165	94	72	46	263	260	89	57	475	1601	7068
DESY	152	487	3563	88	121	735	194	91	55	44	155	130	103	43	402	1259	3563
DLS	54	110	88	3494	72	739	213	336	35	18	145	149	20	12	448	989	3494
ELETTRA	39	298	121	72	2731	455	85	43	23	4	66	316	9	20	162	906	2731
ESRF	319	810	735	739	455	10728	886	406	235	92	600	1069	144	80	1389	3463	10728
ILL	556	165	194	213	85	886	4338	741	343	229	69	176	349	10	1577	1280	4338
ISIS	302	94	91	336	43	406	741	2755	120	119	43	52	155	5	958	740	2755
LLB	156	72	55	35	23	235	343	120	1348	34	12	131	92	3	455	375	1348
SINQ	130	46	44	18	4	92	229	119	34	726	96	9	97	0	348	221	726
SLS	29	263	155	145	66	600	69	43	12	96	2424	182	18	18	177	974	2424
SOLEIL	38	260	130	149	316	1069	176	52	131	9	182	3656	14	26	309	1524	3656
FRM-II	207	89	103	20	9	144	349	155	92	97	18	14	1087	5	522	281	1087
ANKA	15	57	43	12	20	80	10	5	3	0	18	26	5	452	29	157	452

Source http://www.pan-data.eu/CountingUsers

EC Support

These projects have received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no :

- PaNDATA-EU
- PaNDATA-ODI - 283556

  http://www.pan-data.eu
- CRISP - 283745

  http://www.crisp-fp7.eu

Where do we stand (May 2014)

- Open Data Policy published in Nov 2011
- Data Policy implemented in October 2012
- Data Catalogue implemented (ILL + ICAT)
- DOIs available (with DataCite) starting from 2012 experiments.
- NEXUS raw data files on 50% of the Instruments
- Users annotations (experimental e-logbook), available in July 2014
- Federation : ICAT & Umbrella (Authentication)

Data Policy

- We need a clear statement on :
  - Who is the owner
  - If we make them publicly available, how do we protect the Experimental team.
- What was proposed :

1. The facility shall act as a **custodian** for the data.
2. **All raw data will be curated** in a well-defined format with a unique ID.
3. **Metadata** is captured automatically and resides either within the raw data files, and/or in an associated on-line catalogue.
4. **Access to raw data** and the associated metadata obtained from an experiment **is restricted to the experimental team for a maximum period of 3 years**. Thereafter, it will become publicly accessible.
5. The embargo period can be extended on requests.
6. Analysis of openly accessible data must acknowledge the source of the data and cite its unique identifier and any publication linked to the same raw data

http://wiki.pan-data.eu/imagesGHD/0/08/PaN-data-D2-1.pdf

http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

http://www.ill.eu/users/ill-data-policy/

Work is still going on

- Metadata ontology is still a difficult subject
- The implementation of the solutions is still relatively young and could in some case be improved.
- The Id Federation currently based on SAML, need to be extended to support non web technologies.
- Users & Facility (data producers) rewarding (methods & metrics) is still largely insufficient
- …

Thank you for your attention.

Questions ?