Big Data in Science, Best Practice

Steve Aplin Scientific Computing DESY
GridKa School 2013
Karlsruhe, 30th August 2013

HELMHOLTZ ASSOCIATION

LSDMA

pandata

DESY

“May All Your Problems be Technical”

Jim Gray

“A small amount of data may be precious, but a glut of data is a curse”

Anon

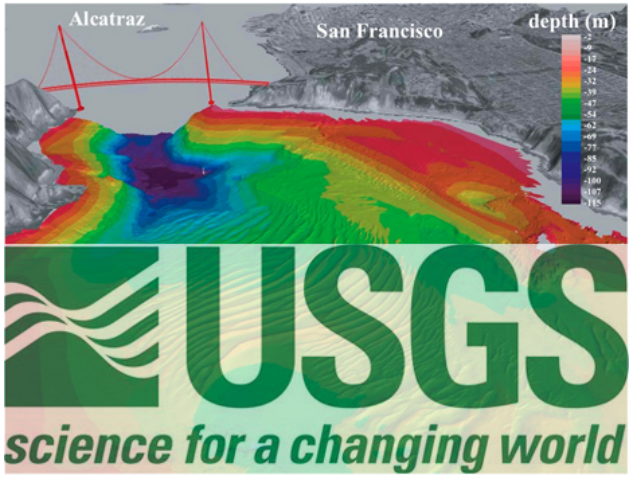
Big Data

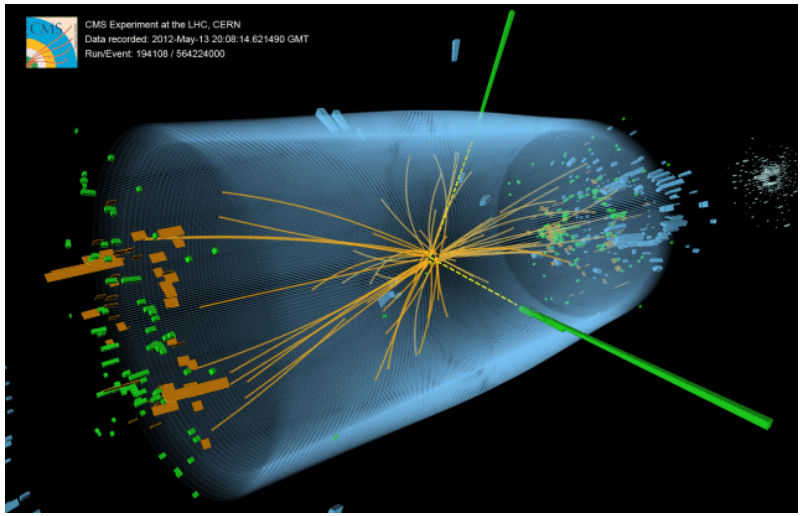
Open Access

The Challenges Posed by Big Data in Science

Technical Challenges

Scientific Challenges





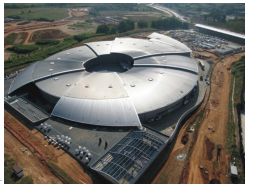
DESY

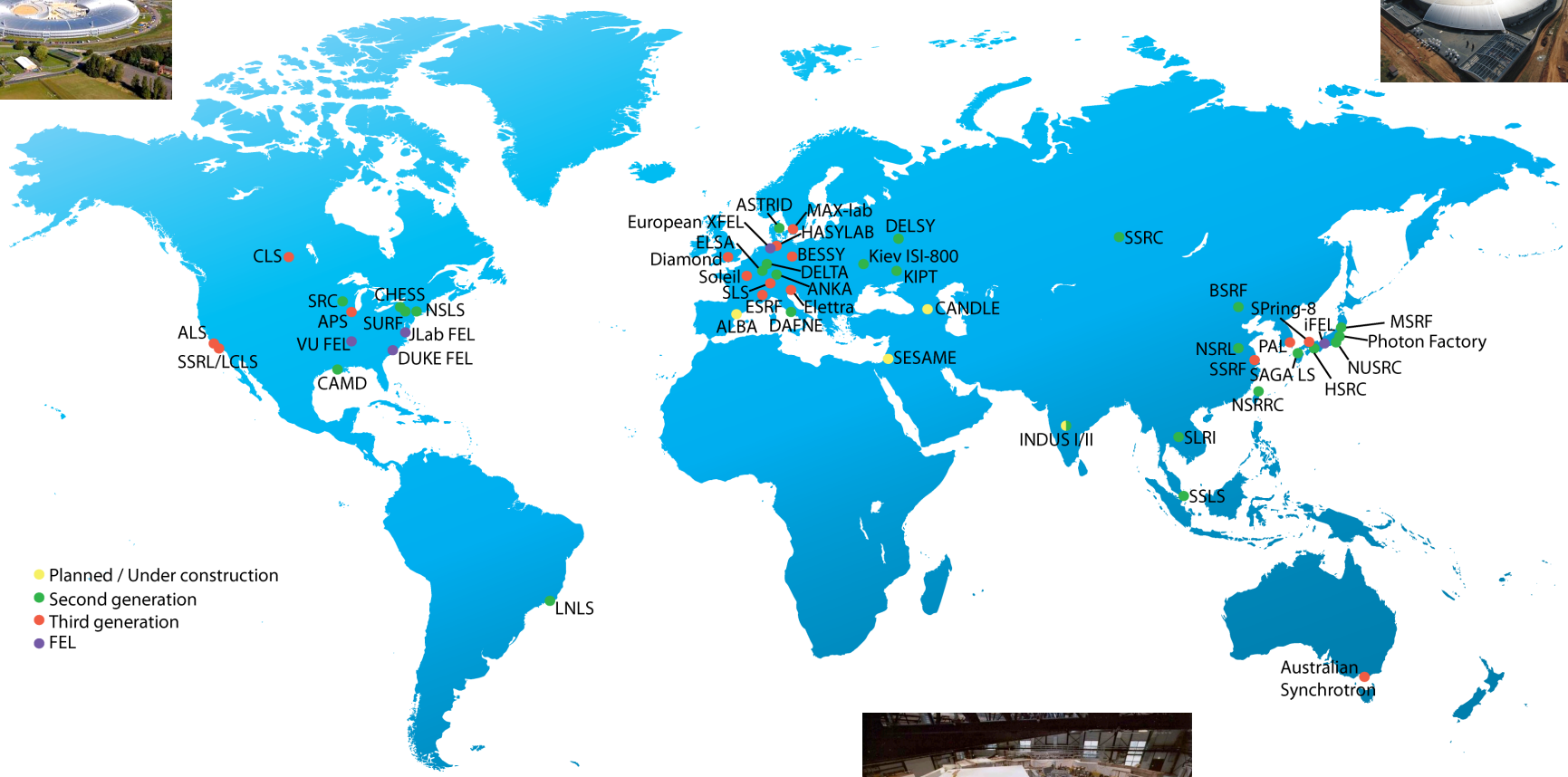
What Does Best Practice Entail Regarding Research Data

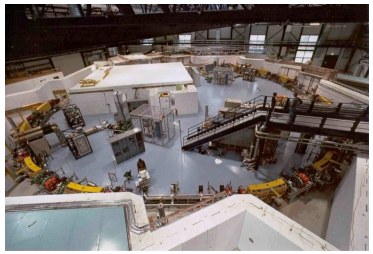
- Data Management Planning
- Documenting Your Data
- Formatting Your Data
- Storing Your Data
- Sharing Your Data
- Ethics and Consent

BEST PRACTICE FOR RESEARCHERS UK Data Archive 2011

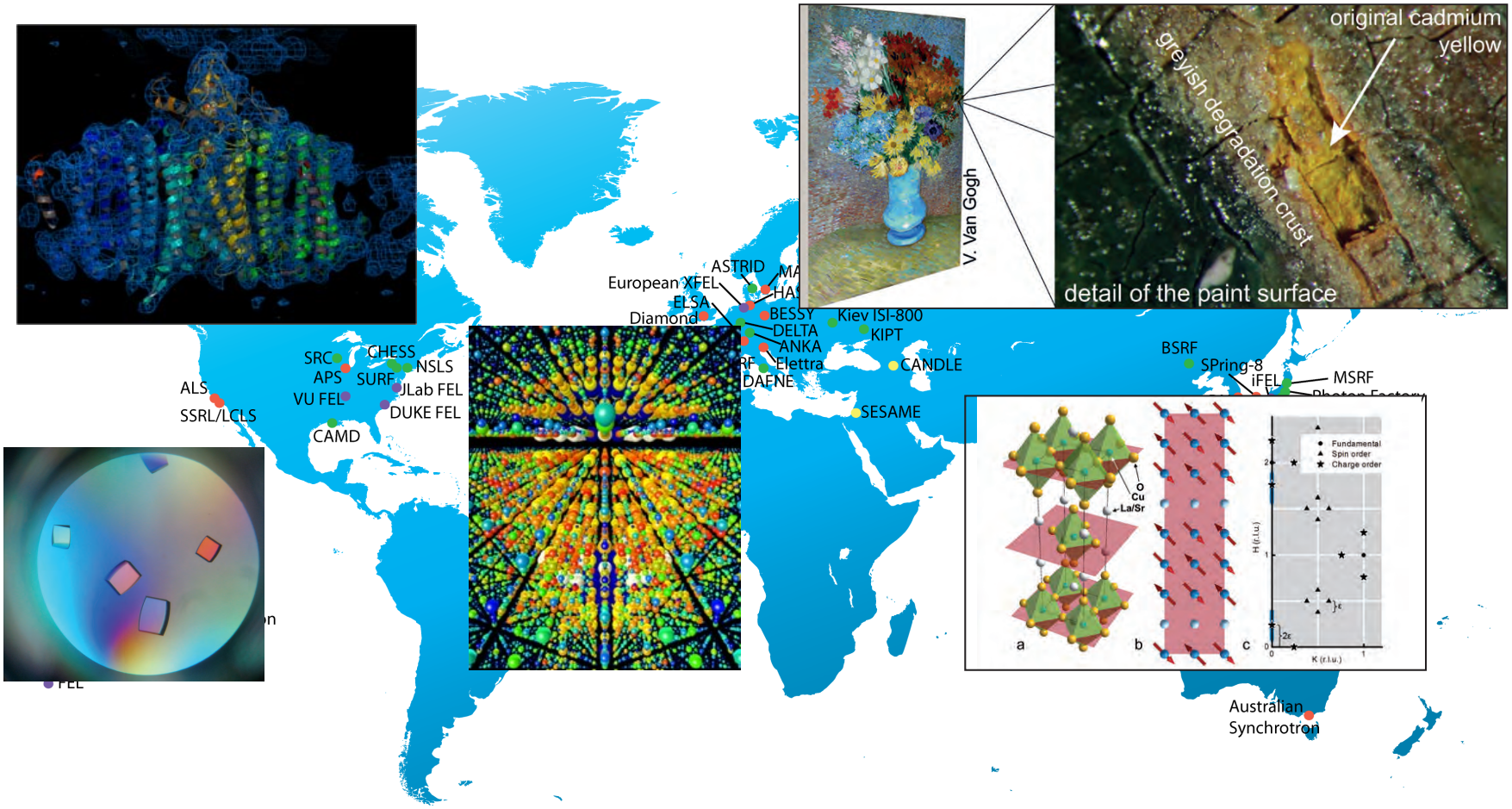
Synchrotrons Shedding New Light onto Sciences





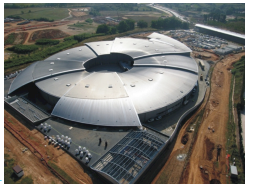


Synchrotrons Shedding New Light onto Sciences



Very Diverse Scientific Usage:
Physics, Chemistry, Biology, even Cultural Heritage

Synchrotrons Shedding New Light onto Sciences



Game Changers

- Dectris Pilatus 6M
- 2463 x 2527 pixels
- 7 MB Images
- 25 frames per sec.
- 175 MB/s

- High Duty Cycles means that 10 TB / day is quite possible

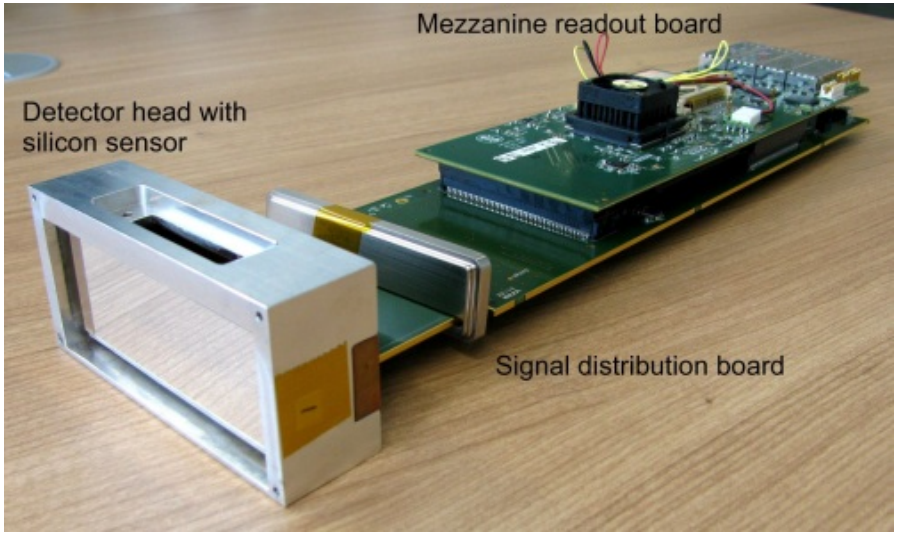
"The Pilatus detector has completely transformed the way X-ray photons are being detected today at synchrotron radiation sources, such as Diamond. This is something we could only have dreamt of in the early days of synchrotron sciences."

Prof. Dr. Gerhard Materlik (CBE), CEO of Diamond Light Source, June 18th, 2012

DESY

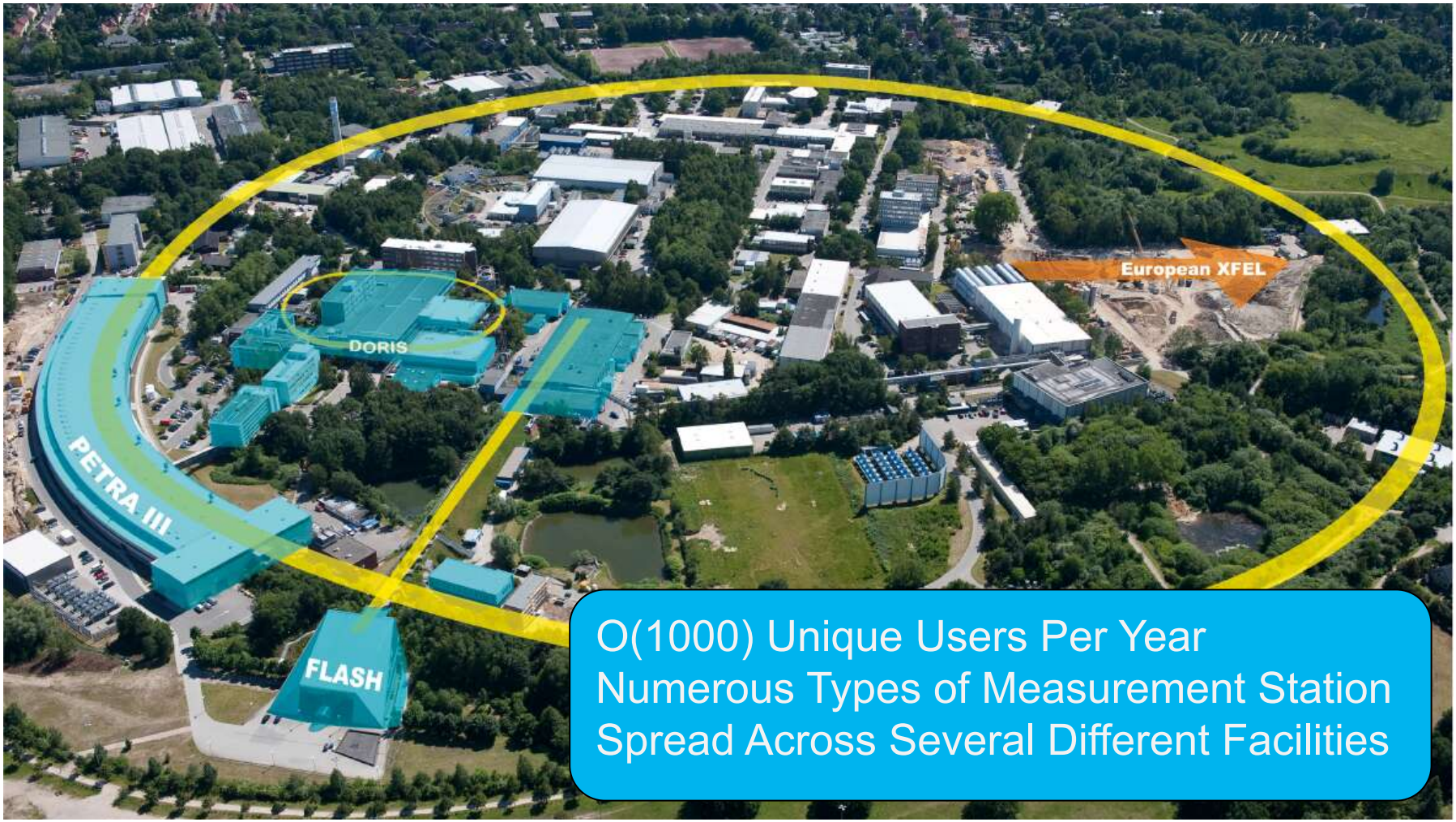
Next Generation

- LAMBDA (Large Area Medipix Based Detector Array)
- Developed at DESY using Medipix developed at CERN
- A single LAMBDA module has 1536 x 512 pixels, and multiple modules can be tiled to cover a large area.
- Operates in a “continuous read-write” mode with negligible dead time between images.
- Expected to allow continuous 1000 fps readout using 10 Gigabit Ethernet links.



- ~ GB/s per module

DESY Light Sources



O(1000) Unique Users Per Year
Numerous Types of Measurement Station
Spread Across Several Different Facilities

The Problem

- Data Sets Become Too Large to Take Home.
- Data Rates Begin to Require Dedicated Central IT Infrastructure, Way Beyond Previous Requirements.
- Wide Variety of Scientific Users Means Significant Number of Data Formats and Analysis Software.
- Large Number of Users With Little Affiliation to the Lab.

So What Does Best Practice Dictate

- Data Sets Become Too Large to Take Home.

  Define Data Management Policies

- Data Rates Begin to Require Dedicated Central IT Infrastructure, Way Beyond Previous Requirements.

  Realistic Performance Estimates, and Plan Ahead

- Wide Variety of Scientific Users Means Significant Number of Data Formats and Analysis Software.

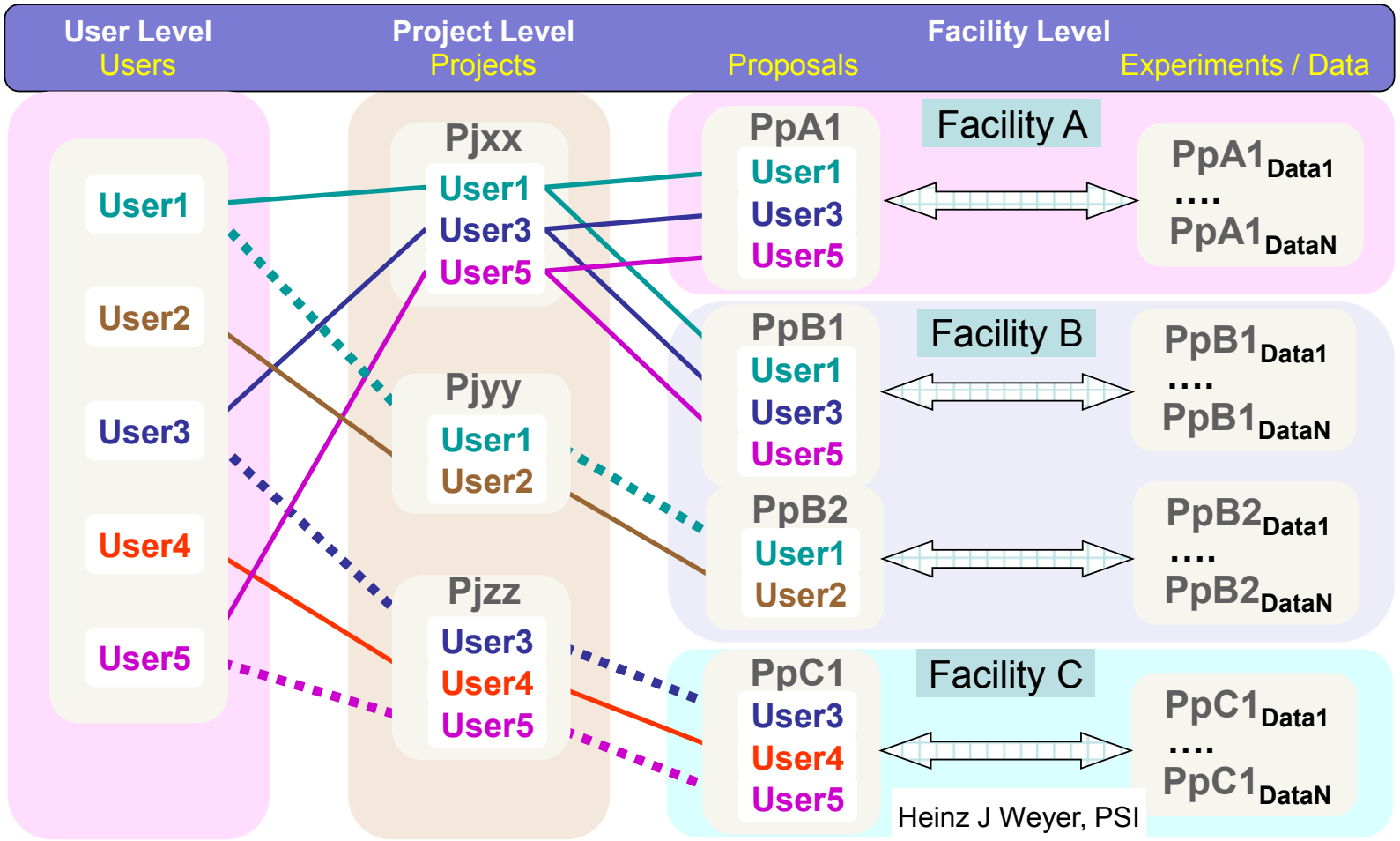
  Insist, Insist, Insist …

- Large Number of Users With Little Affiliation to the Lab.

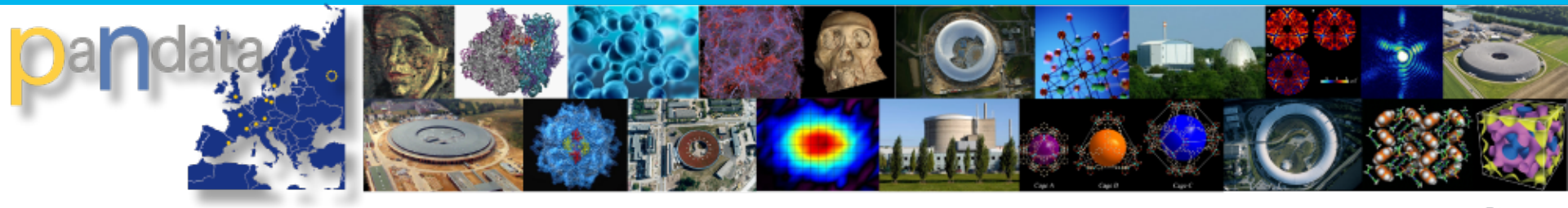
  Pray

Research Facilities as Data Custodians

- With data-sets increasing to sizes which become infeasible to take home, Research Facilities get left holding the baby.

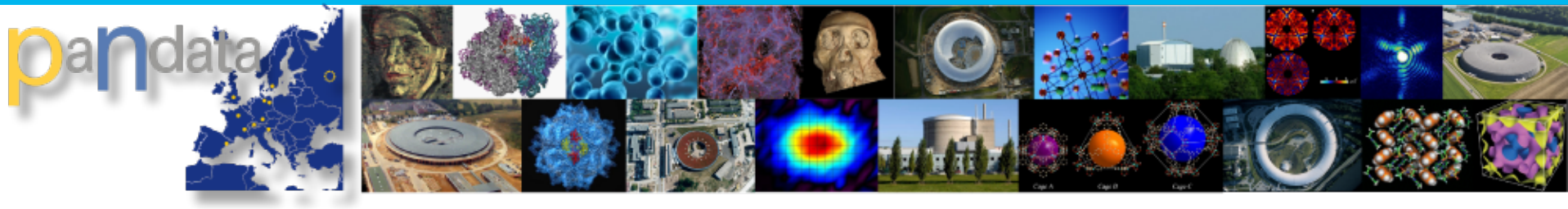


Strength in Numbers



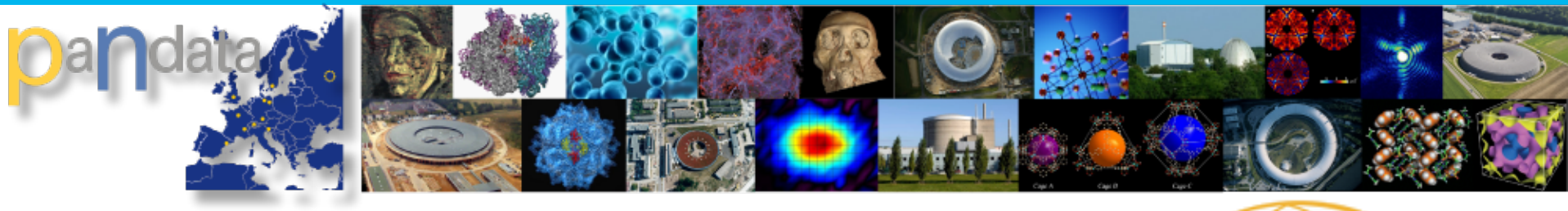
- Photon and Neutron Data Infrastructure
- Established in 2007 with 4 facilities, now standing at 13
- Combined Number of Unique Users more than 35000 in 2011
- Combines Scientific and IT staff from the collaborating facilities
- European Framework 7 Project

Aims of the Project



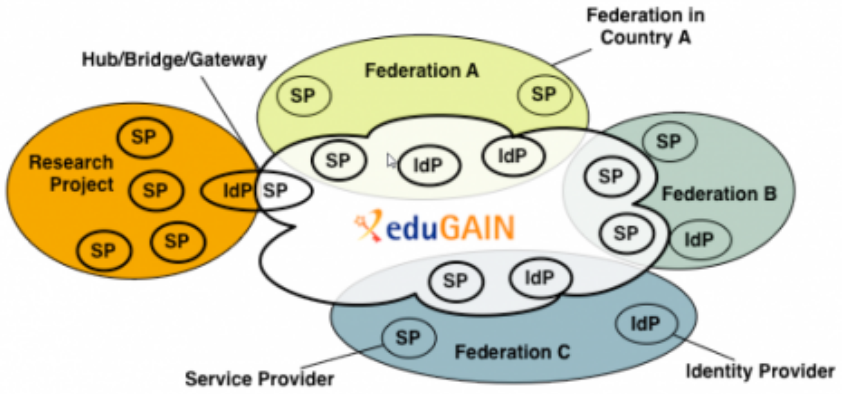
- Harmonize authentication and authorization
- Standardize data formats and annotation of data
- Allow transparent and secure remote access to data
- Establish sustainable and compatible distributed data catalogues
- Allow long term preservation of data
- Provide compatible data analysis software
- Promote data policies in laboratories

Harmonize Authentication and Authorization

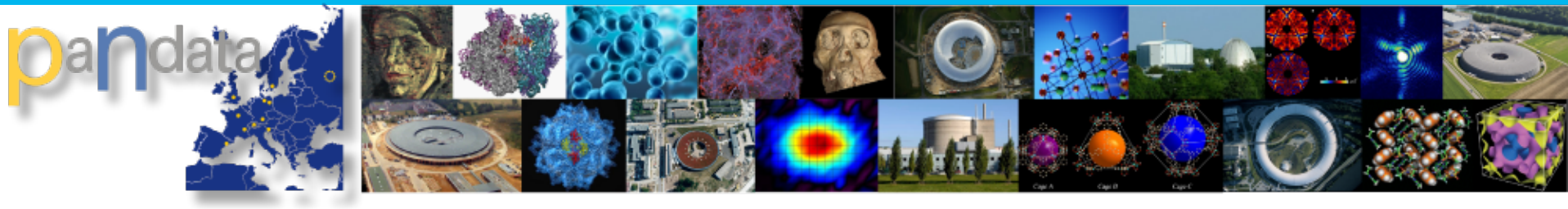


umbrella

- Umbrella Identity Management System (IdM)
- Built on top of the existing well established User Offices
- Utilises Internet2's Shibbolleth Federated AAI
- Based on a single Umbrella IdP with the facilities acting as Federated Services, i.e. they maintain Authorisation role
- Recently GEANT GN3plus committed to allowing Research Projects to use its European wide AAI infrastructure eduGain



Standardize data formats and annotation of data



- NeXus/HDF5 Chosen as the common format. The HDF Group
- NeXus is developed as an international standard by scientists representing major scientific facilities across the world.
- NeXus builds on top of HDF5 so any NeXus file is a fully valid HDF5 file, which can be read by a large number of applications without any further modification.
- HDF5 is a widely adopted, standardized data format and has been proposed by the European Commission as an ISO standard for all binary data.

DESY

Data Access and Catalogues

pandata europe

Metadata Catalogue

http://www.icatproject.org/

ICAT

Metadata Repository

Proposal

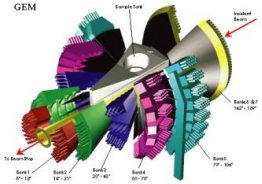
Scientist submits application for beamtime

Approval

Facility committee approves application

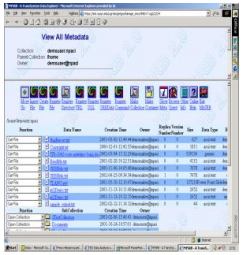
Scheduling

Facility registers, trains, and schedules scientist's visit



Experiment

Scientists visits, facility run's experiment

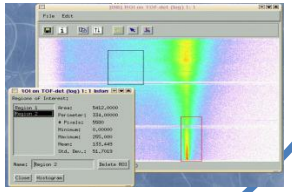


Data cleansing

Raw data filtered and cleansed

Data analysis

Tools for processing made available



Record Publication

Subsequent publication registered with facility

DESY

Sustainable Infrastructure

Large Scale Data Management and Analysis

- Provide tools, services and processes required for the uniform access to computing and storage resources, regardless of the scientific discipline.
- Federated Identity Management
- Federated Data Access
- Meta Data Catalogues & Repositories
- Archive Service
- Monitoring
- Data Intensive Computing
- **4 Helmholtz research centers, 6 Univ. and DKRZ**

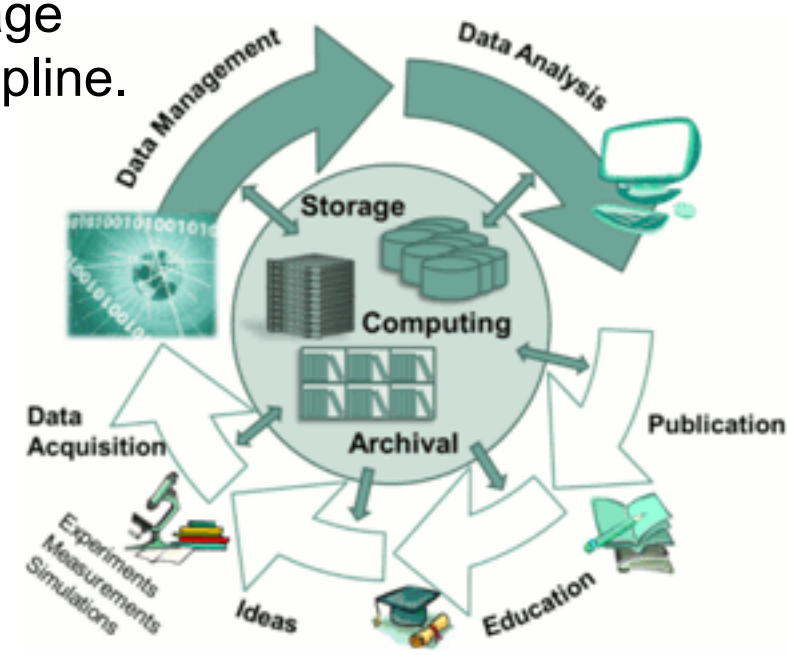
Sustainable Infrastructure

LSDMA

Large Scale Data Management and Analysis

HELMHOLTZ GEMEINSCHAFT

- Provide tools, services and processes required for the uniform access to computing and storage resources, regardless of the scientific discipline.
- Federated Identity Management
- Federated Data Access
- Meta Data Catalogues & Repositories
- Archive Service
- Monitoring
- Data Intensive Computing



- **Started in 2012, will run to 2016. After which it is projected to be integrated into sustainable POF of Helmholtz Association**

DESY

OPEN ACCESS TO DATA – Correcting The Data-Gap

- "Technologies capable of acquiring and storing vast and complex datasets challenge the principle that science is a self-correcting enterprise. How can a theory be challenged or corrected if the data that underlies it is neither accessible nor assessable?"
- "A great deal of data has become detached from the published conclusions that depend upon it, such that the two vital complementary components of the scientific endeavour - the idea and the evidence - are too frequently separated."
- "The rewards for attracting resources for research and making important scientific discoveries are considerable. These can create temptations for some scientists, whether in the public or private sector, to indulge in poor practices that range from blatant fraud, where data are invented, to selective reporting of findings in order to promote a particular hypothesis."

Science as an Open Enterprise The Royal Society 2012

OPEN ACCESS TO DATA – Correcting The Data-Gap

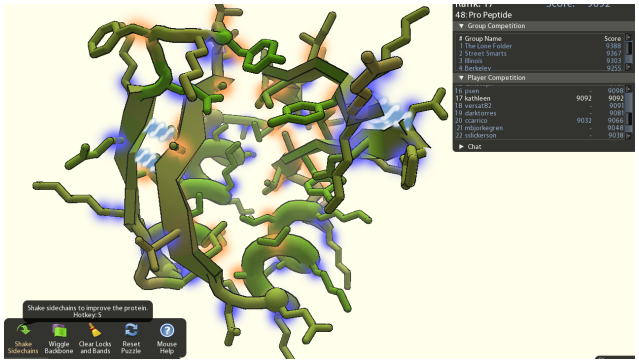
Office of Science and Technology Policy

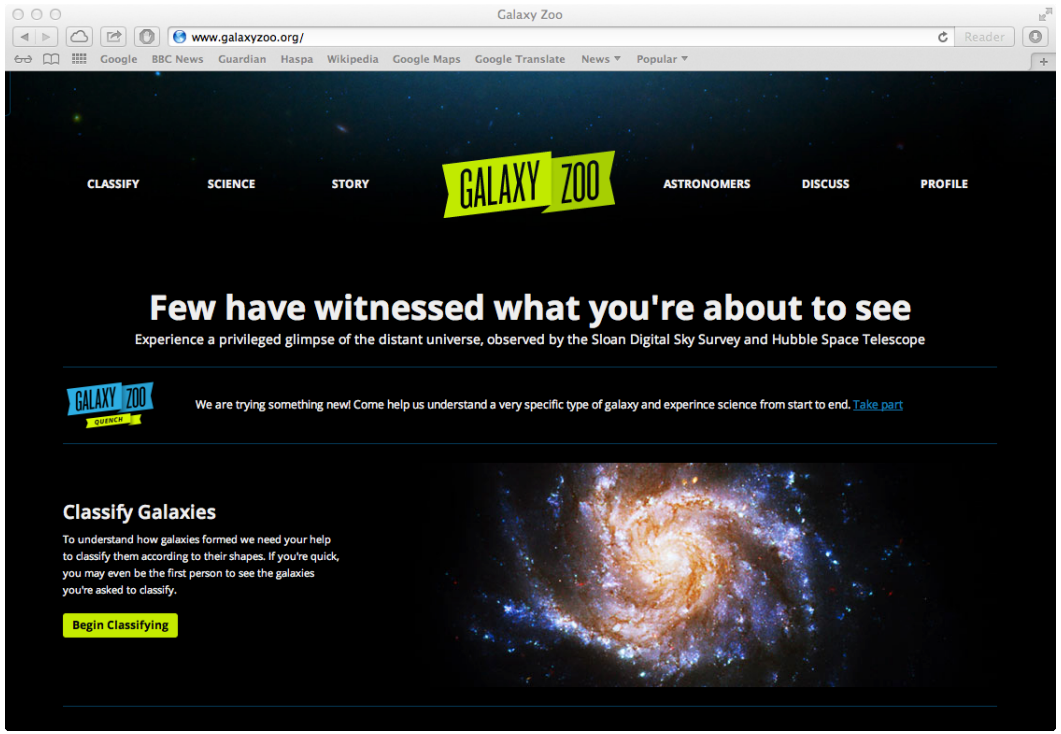
OSTP Public Access Policy Forum

"The Obama Administration is committed to the proposition that citizens deserve easy access to the results of scientific research their tax dollars have paid for."

Public funders of research increasingly follow guidance from the Organisation for Economic Co-operation and Development (OECD) that publicly funded research data should as far as possible be openly available to the scientific community.

OPEN ACCESS TO DATA – A Reality





OPEN ACCESS TO DATA – Data Preservation

- **Data from space explorer Viking were partly useless in 1985 after only 9 years of storage. Data from 30 years of space exploration were stored on 1.2 million magnetic tapes by NASA in 1976 but in the mid-nineties many were useless[1].**

- **"Last month, researchers working out of an abandoned McDonald's restaurant on the grounds of NASA Ames Research Center recovered data collected by NASA's Nimbus II satellite on 23 September 1966 .. . The LOIRP[2] team obtained $750,000 from NASA and private enterprise and enlisted the assistance of a retired Ampex engineer. They cleaned, rebuilt, and reassembled one drive, then designed and built equipment to convert the analog signals into an exact 16-bit digital copy."[3]**

© NASA, http://www.nasa.gov/topics/moonmars/features/LOIRP/loirp-gallery-index.html#

[1]Hilmar Schmundt: Im Dschungel der Formate. In: Der Spiegel 26/2000. URL: http://www.spiegel.de/spiegel/print/d-16748341.html

[2]Lunar Orbiter Image Recovery Project

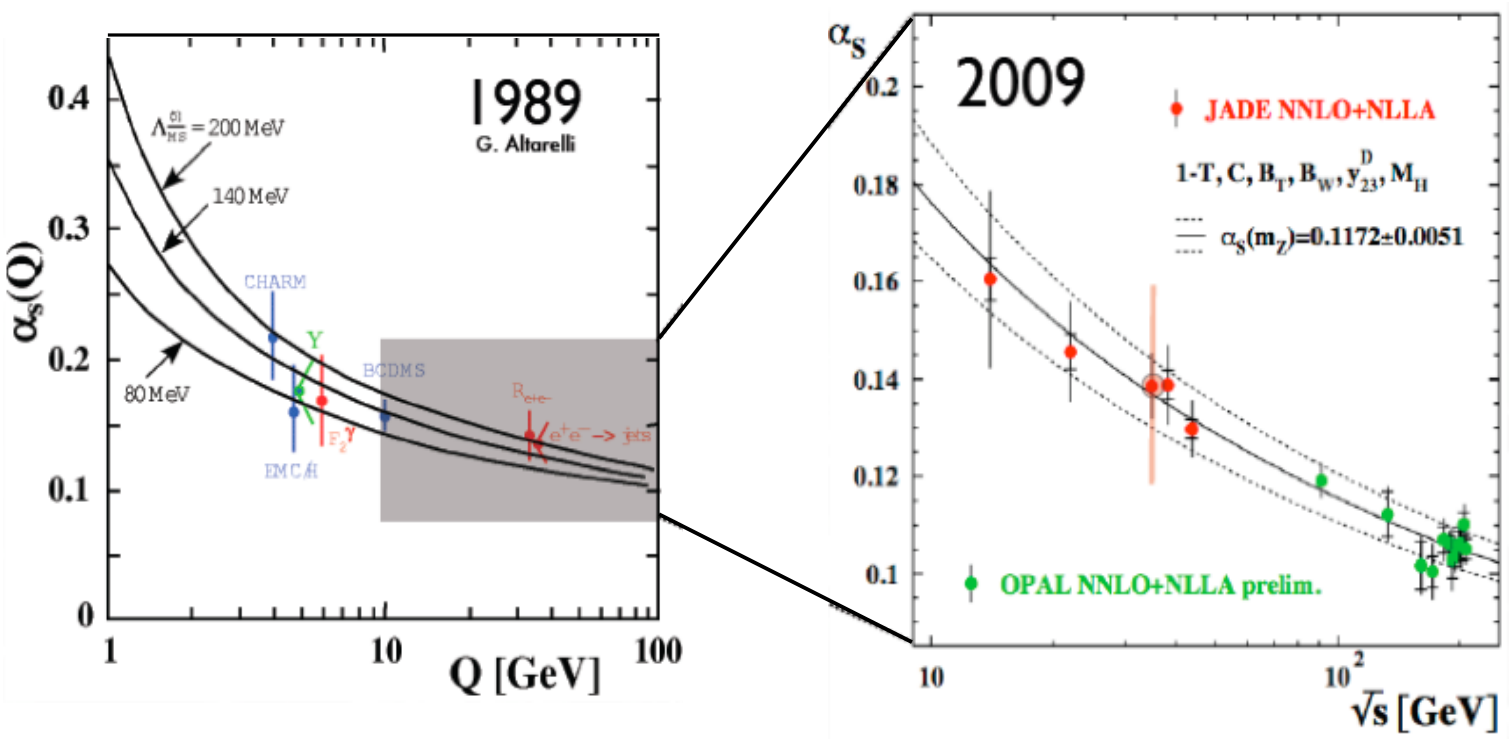
[3]SCIENCE Magazine, 12 March 2010, Vol. 327, p. 1322

Problems in Research Data Archiving
Iver Lauermann HZB

DESY

OPEN ACCESS TO DATA – Data Preservation

Precision measurements of α_S and tests of asymptotic freedom



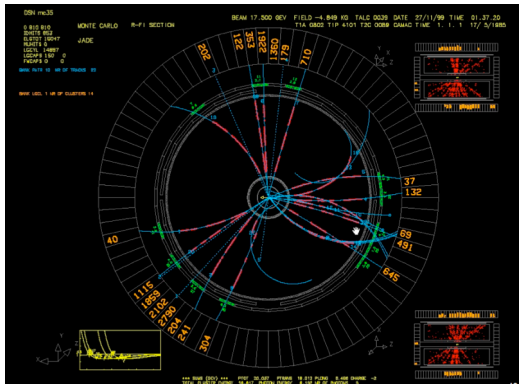
In NLO QCD: $\alpha_S(35\ \text{GeV}) = 0.14 \pm 0.02$
No indication of a running α_S signature

In re-summed NNLO QCD: $\alpha_S(M_Z) = 0.1172 \pm 0.0051$
Significant evidence of running α_S and asym. freedom

David South Data Preservation in HEP

OPEN ACCESS TO DATA – Data Preservation

- Recent re-analysis of JADE data is such a case (S. Bethke, J. Olsson et al.)
  - Conversion of old data format done in 2005 and 2008 by Jan
  - Successful revitalisation and validation of complete JADE reconstruction and simulation software and event display
  - Involved conversion, translation and some rewriting of original code



- There were several interesting JADE anecdotes along the way, including:
  - A hand-typed recovery of a luminosity / calibration file from a (green) paper copy found in Jan's DESY office
  - An old version of BOSlib79 found at the Tokyo computing centre
  - Original JADE MC 9-track tapes found in a Heidelberg University cupboard
- Only through careful documentation will we avoid such things

David South Data Preservation in HEP

OPEN ACCESS Freely available online

PLoS one

Empirical Study of Data Sharing by Authors Publishing in PLoS Journals

Caroline J. Savage, Andrew J. Vickers*

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America

Abstract

Background: Many journals now require authors share their data with other investigators, either by depositing the data in a public repository or making it freely available upon request. These policies are explicit, but remain largely untested. We sought to determine how well authors comply with such policies by requesting data from authors who had published in one of two journals with clear data sharing policies.

Methods and Findings: We requested data from ten investigators who had published in either PLoS Medicine or PLoS Clinical Trials. All responses were carefully documented. In the event that we were refused data, we reminded authors of the journal's data sharing guidelines. If we did not receive a response to our initial request, a second request was made. Following the ten requests for raw data, three investigators did not respond, four authors responded and refused to share their data, two email addresses were no longer valid, and one author requested further details. A reminder of PLoS's explicit requirement that authors share data did not change the reply from the four authors who initially refused. Only one author sent an original data set.

Conclusions: We received only one of ten raw data sets requested. This suggests that journal policies requiring data sharing do not lead to authors making their data sets available to independent investigators.

OPEN ACCESS TO DATA – Compliance

OPEN ACCESS Freely available online

PLoS one

Empirical Study of Data Sharing by Authors Publishing in PLoS Journals

Caroline J. Savage, Andrew J. Vickers*

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America

Abstract

Background: Many journals now require authors share their data with other investigators, either by depositing the data in a public repository or making it freely available upon request. These policies are explicit, but remain largely untested. We sought to determine how well authors comply with such policies by requesting data from authors who had published in one of two journals with clear data sharing policies.

Methods and Findings: We requested data from ten investigators who had published in either PLoS Medicine or PLoS Clinical Trials. All responses were carefully documented. In the event that we were refused data, we reminded authors of the journal's data sharing guidelines. If we did not receive a response to our initial request, a second request was made. Following the ten requests for raw data, three investigators did not respond, four authors responded and refused to share their data, two email addresses were no longer valid, and one author requested further details. A reminder of PLoS's explicit requirement that authors share data did not change the reply from the four authors who initially refused. Only one author sent an original data set.

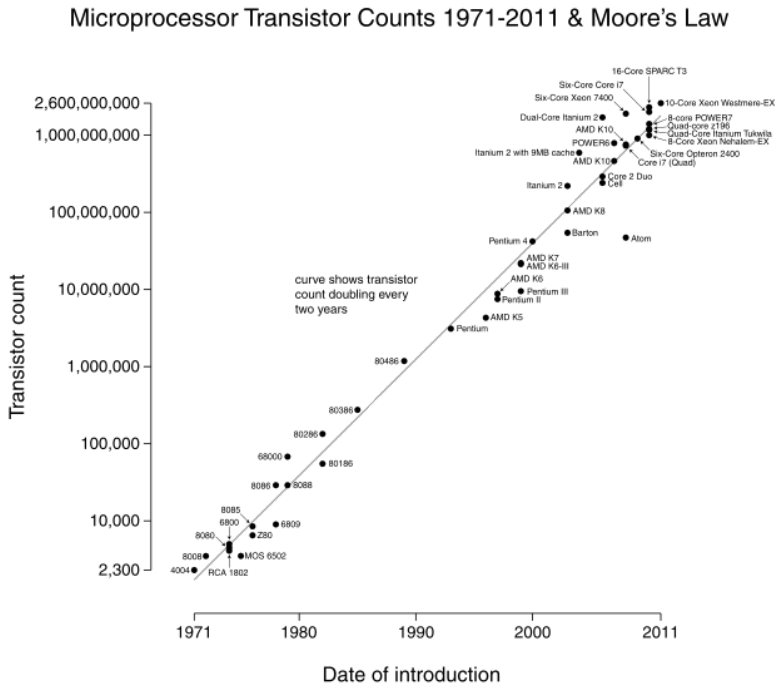
Conclusions: We received only one of ten raw data sets requested. This suggests that journal policies requiring data sharing do not lead to authors making their data sets available to independent investigators.

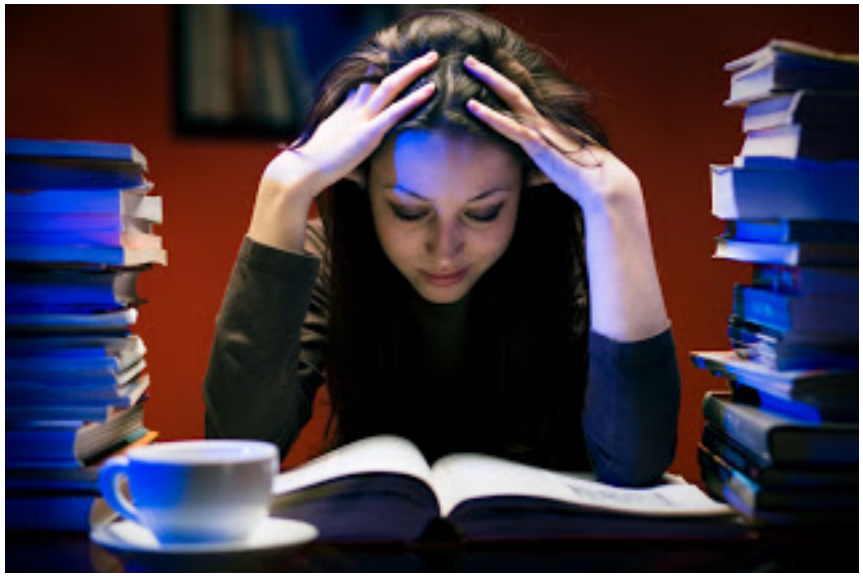
- Requested data from ten investigators who had published in either PLoS Medicine or PLoS Clinical Trials. Journals with an Open Data Mandate for their publications.
- In the event of refusal, authors reminded of the journal's data sharing guidelines.
- Three investigators did not respond, four authors responded and refused to share their data, two email addresses were no longer valid, and one author requested further details.
- Only one author sent an original data set.

Summary

- Best Practice has to be enshrined in Policy.
- Policies will only be as good as they are enforceable. Those that rely solely on enlightened self interest will struggle to succeed in a heterogeneous environment.
- Standard formats and protocols should always be favored over propriety solutions. This does not mean that choosing a standard is easy.
- Open Access is gaining ground both politically and in practice. How this will affect you is still unclear.

> Remember Hardware is Cheap, and Easily Replaced; People are Neither Nor.





- http://www.helmholtz-lsdma.de
- http://pan-data.eu
- http://royalsociety.org/policy/projects/science-public-enterprise/report/
- http://data-archive.ac.uk
- http://oa.helmholtz.de
- Savage CJ, Vickers AJ (2009) Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLoS ONE 4(9): e7078. doi: 10.1371/ journal.pone.0007078