



ECM27 Workshop on Data Diffraction Deposition



TOC

Facility environment

- Research at large facilities
- ✤ IT requests
- Facilities and users
- **EU** projects
 - PaNdata and CRISP
 - MI3 and CALIPSO
 - Biostruct X
 - Urgent issues
- Authentication / Authorization
 - Umbrella
 - Federated Identity Management
- Conclusion



Research at large facilities I

Photon facilities

- Synchrotrons and Free Electron Lasers (FELs)
- Produce light of highest brightness
- Typical range from infra-red to Xrays
- About 15 synchrotrons in EU (ESRF + national)

FELs, even 10³ to 10⁶ times brighter

- SLAC/Stanford, DESY/Hamburg, FEL/Spring-8/Japan, PSI/Villigen
- Membrane proteins; microscopic movies of chemical reactions

Neutron facilities

- Complementary
- Similar user community
- Wide range of research areas
 - Archaeology, chemistry, materials science, life sciences, physics …

Small teams, visit for

- Few hours (structural biology) to
- Few weeks (superconductivity, nano investigations)



Research at large facilities II

In EU over 30'000 visiting users /y

- ✤ Large overbooking (≥3:1), low chance to be accepted
- Important to minimize administrative load (local user offices)

On-site visits

- Short duration
- In part spontaneous (keep that attraction)
- Part-time users
- Fedex-type experiments

Decentralized structure (compare e.g. to CERN)

- Manifold research fields
- Several facilities, trans-facility experiments

National character of facilities

Report to national governments (with few exceptions)



What are the IT requests? I

Huge datasets

- Novel 2D detectors, quantum leap in data quality, but also data volumes
- Multi-image techniques (tomography, lens-less imaging)
- Molecular movies at FELs
- 'Petabyte' 'normal' unity; time over for 'hard-disk in the trouser pocket'
- Many talk about *storing* data, but must also to talk about *handling*, need for new strategies

Trans-facility experiments

- Standardize proposal procedures on EU scale
- Standardize metadata

Remote, non-local data access

- Analyze data remotely at facility
- Combine datasets taken at different facilities [Umbrella(PSI)+ICAT(STFC)?]
- Combine different data types (raw, derived, published)
- Clouds (commercial, community-centered)



What are the IT requests? II

Remote experiment access

- Basic: passive online access to measured data
- Advanced: active control [Umbrella(PSI)+Moonshot(STFC)?]

International identity

- ✤ Unique
- Persistent
- User friendly

Online, On-the-fly data analysis

- Are the experimental parameters right?
- Filtering?

PR Issues

- Improve corporate identity
- Improve public lobbying



But ...

There is no free money lying around

- Within institutes large facilities are competing with other excellent projects
- Even more projects coming up (e.g. FELs)
- In 1st order total sum resources at best constant
- Resources for IT not always at top of popularity scale
- So, would have to
 - o shift money from other requests (detectors)
 - o shift manpower

Way out

- Simplify procedures
 - o Consequences on resources
- Need to archive all that data?
 - o Filters
 - o Triggers
 - o ... come back to that later

Look out for synergies

o EU projects



Sociology of facilities and users

- Progress possible only, if facilities and users collaborate
- Commonalities and differences

Organizational structure

- Facilities
 - o Well structured
- Users
 - o Loose collaborations

Coupling to infrastructure

- Facilities
 - Long-term commitment of resources, setting of priorities, financial responsibility
- Users
 - o Limited, mainly just users

Long-term relation and interest to BL

- Facilities
 - o Yes
- Users
 - o Very limited

Selection of experiments

Facilities

Scientific orientation

- Facilities
 - o According to resources, focused
- ✤ Users
 - o Very flexible, wide range
- Reporting to
 - Facilities
 - Facility management, national government
 - ✤ Users
 - o International community
- **Figure of merit**
 - Facilities
 - o Publications
 - ✤ Users
 - o Publications



User and Beamline Scientists

On the one hand service

- Provide support, expert knowledge
 - o Extreme mode: Fedex-type experiments (but caveat)

On the other hand need support from users

- Prioritization of new developments
 - o Resource competition with other facility projects
- Justification towards facility management
- Increased need for IT contacts before (!) measurement
 - Resource optimization
 - Setup of filters / triggers
- Publications
 - Adequate citations
 - o Figure of merit also for BL scientists and facilities



TOC

- Facility environment
 - Research at large facilities
 - ✤ IT requests
 - ✤ Facilities and users

EU projects

- PaNdata and CRISP
- MI3 and CALIPSO
- Biostruct X
- Urgent issues
- Authentication / Authorization
 - Umbrella
 - Federated Identity Management
- Conclusion



PaNdata ODI

- **PaNdata Open Data Infrastructure**
- Proposal to construct and operate a sustainable data infrastructure for European Photon and Neutron laboratories. This will enhance all research done in the neutron and photon communities by making scientific data accessible allowing experiments to be carried out jointly in several laboratories.
- **Formed in 2008**
- PaNdata collaboration: 13 major world class European Research Infrastructures to construct and operate a common data infrastructure for the European Neutron and Photon large facilities.
- In 2010: begin of a Support Action which is focusing on standardization activities in the areas of:
 - ✤ data policy,
 - user information exchange,
 - scientific data formats,
 - interoperation of data analysis software,
 - integration and cross-linking of research outputs.



PaNdata ODI Work Packages

UWP3, User Catalogue and AAA Service (PSI)

 To deploy, operate and evaluate a system for pan-European user identification across the participating facilities

U WP4, Data catalogue Service (ELETTRA)

- This work package will deploy, operate and evaluate a generic catalogue of scientific data across the participating facilities and promote its integration with other catalogues beyond the project.
- o Specifically, we will:
 - 1. Develop the generic software infrastructure to support the interoperation of facility data catalogues,
 - 2. Deploy this software to establish a federated catalogue of data across the partners,
 - 3. Provide data services based upon this generic framework which will enable users to deposit, search, visualize, and analyze data across the partners' data repositories,
 - 4. Evaluate this service from the perspective of facility users,
 - 5. Manage jointly the evolution of this software and the services based upon it,
 - 6. Promote the take up of this technology and the services based upon it beyond the project.

WP5, Virtual Laboratories (DESY)

To deploy a set of integrated end-to-end user and data services supporting three specific techniques: (1) Structural 'joint refinement' against X-ray & neutron powder diffraction data, (2) simultaneous analysis of SAXS and SANS data for large scale structures, (3) access to tomography data exemplified through paleontological samples.



PaNdata Work Packages

WP6, Provenance (STFC), start m7

 To develop a conceptual framework, which can record and recall the data continuum, and especially the analysis process, and to provide a software infrastructure which implements that model to record analysis steps hence enabling the tracing of the derivation of analyzed data outputs.

WP7, Preservation (ILL), start m10

 To incorporate models and tools oriented towards long-term data preservation into the PaNdata infrastructure, focusing on several aspects considered of benefit: an OAIS-based infrastructure; persistent identifiers; and certification of authenticity and integrity.

WP8, Scalability (DIAMOND)

 To develop a scalable data processing framework combining parallel file systems with a parallelized standard data format (Nexus, HDF5) to permit applications to make most efficient use of dedicated multi-core environments and to permit simultaneous ingest of data from various sources, while maintaining the possibility for real-time data processing.



PaNdata collaborators

- Joachim Metge
- - Michael Hagelstein
- DESY
 - Frank Schluenzen, Rolf Treusch, Jan-Peter Kurz, Ulrike Lindemann
- DIAMOND
 - Bill Pulford
- Fermi/Elettra
 - Cecilia Blasetti, Ornela Degiacomo, Giorgio Paolucci

ESRF

- Rudolf Dimper, Dominique Porte, Stefan Schulze
- **HZB**
 - Thomas Gutberlet, Dietmar Herrendoerfer, Olaf Schwarzkopf

- Jean-Francois Perrin, F. Festivi
- - Tom Griffin
- MaxLAB
 - Ulf Johansson

D PSI

- Bjoern Abt, Stephan Egli, Stefan Janssen, Mirjam van Daalen, Heinz J Weyer
- **Soleil**
 - Frederique Fraissard
- STFC
 - Juan Bicarregui, Anthony Gleeson, Brian Matthews



CRISP

- **Name:** Cluster of Research Infrastructures and Synergies in Physics (CRISP)
- Purpose is to create synergies and develop common solutions for an initial group of eleven ESFRI-PPs (European Strategy Forum on Research Infrastructure preparatory phase) projects in the field of Physics, Astronomy, and Analytical Facilities.

Ultimate aim is

- To supply the best service to the rapidly growing and largely diversified user community, and
- To ensure that the large investments made at the national and international levels result in significant progress in science.
- Key topics identified within these challenges have been clustered into Topic Groups:
 - Accelerators,
 - Instruments & Experiments,
 - Detectors & Data Acquisition,
 - Information Technology & Data Management.



CRISP IT Work Packages

WP16, Common User Identity System (PSI)

 Develop and deploy a pan-European system for unique identification (Authentication and authorization infrastructure: AAI) of users at the infrastructures of the participating RIs EuroFEL (PSI), ESRF, ESS, FAIR (GSI), ILL, and XFEL for the management of local and remote access to facilities, experiments, data, and IT resources.

WP17, Metadata Management and Data Continuum (ILL)

 The main objectives of this work package are (1) to choose and implement metadata management and metadata mining services and (2) to establish an environment permitting a data continuum from raw data to publications across the participating RIs ILL, ESRF, SLHC at CERN, and EuroFEL (DESY).

U WP18, High-speed Data Recording (EU XFEL)

 The objective of this work package is to provide solutions for (1) high-speed recording of data to permanent storage and archive, and (2) optimized and secured access to data using standard protocols for the RIs XFEL, ESRF, EuroFEL (DESY), ESS, ILL, and SKA (UOXF.DB).

WP19, Distributed Data Infrastructure (CERN)

 Analyze the existing distributed data infrastructures from the network and technology perspective. Plan and experiment their evolution to support the expanding data management needs of the set of participating research infrastructures. SLHC at CERN, EuroFEL (DESY), FAIR (GSI), ELI (MTA-SZTAKI) and SKA (UOXF.DB) participate to all tasks.



CRISP IT collaborators

Laurence Field

DESY

 Frank Schluenzen, Rolf Treusch, Jan-Peter Kurz, Ulrike Lindemann

ESRF

 Rudolf Dimper, Dominique Porte, Stefan Schulze

ESS

Stig Skelboe

GANIL

GSI

Peter Malzacher

🗆 ILL

Jean-Francois Perrin, F. Festivi

C XFEL

- Krzysztof Wrona
- Description PSI
 - Bjoern Abt, Stephan Egli, Stefan Janssen, Mirjam van Daalen, Heinz J Weyer



Other important FP7 projects I

Gamma Facility-oriented, I3 (Integrated Infrastructure Initiatives)

NMI3, Neutron Scattering and Muon Spectroscopy

- Facilitate the pan-European coordination of neutron scattering and muon spectroscopy research activities, by integrating all research infrastructures in these fields within the European Research Area. NMI3 is a consortium of 18 partner organizations from 12 countries, including 8 facilities.
- Transnational Access –gives European users access to all of the relevant European research facilities and hence the possibility to use the best adapted infrastructure for their research.
- Joint Research Activities NMI3 fosters collaborations focusing on specific R&D areas to develop techniques and methods for the next-generation instrumentation. These collaborations are transnational and involve all European facilities and academic institutions with experts and know-how in the relevant fields.
- <u>Education</u> By offering funding for schools and workshops and producing educational and dissemination resources, NMI3 aims to train future generations of users.

CALIPSO, same for Synchrotron and FEL facilities

- Coordinated access to Lightsources to promote standards and optimization; all large EU facilities.
- o Also trans-national access, JRAs



Other important FP7 projects II

Research-field-oriented

- Biostruct X, Structural Biology
 - Provides integrated transnational access via 44 European installations in four key areas of structural biology:
 - Macromolecular X-ray crystallography (MX)
 - Small angle X-ray scattering (SAXS)
 - X-ray imaging (XI)
 - Protein production and high-throughput
 - Crystallization (PP&HTX).
 - o Offers:
 - Access to facility and experimental station
 - Automated sample handling
 - Remote experimental control (optional)
 - Online sample purification (optional)
 - Online data processing and interpretation software
 - Access to associated infrastructure sites, laboratory facilities, and computational facilities.
 - Data processing and analysis software



Potential operational conflicts

EU support via CALIPSO / NMI3

- Support fits research facility structure
 - o Support control via facility-local Proposal Review Committees
 - o But CALIPSO would have needed 30M, got <10M

EU support via Biostruct X

- Research at one specific facility only part of larger proposal
 - o Measurement seen in wider context
 - o Decision on support already before coming to facility
 - o Attractive concept, but severe management problems

Issue not yet solved

- Duplication of user databases (< 30'000 users annually)
- Duplication of
 - o User side: proposals
 - o Facilities & Biostruct: scientific ranking and committees
- Competence conflicts
 - o Who decides upon research direction?
 - o The EU takes the easy road
- But important to find a solution
 - o Will very probably not be the last case



Umbrella and BioStruct





b) BioStruct as present present







Umbrella and BioStruct II







Urgent Issues for Facility-User Cooperation

Common Data Policy

- Data preservation, public / restricted access; embargo period (R. Dimper, C. Nave)
- 🖵 Common Data Format
 - ✤ NEXUS, HDF5

Metadata standardization

Electronic logbook, reanalyze data, trans-facility experiments

Data handling

- Remote Data access
- Remote experiment access
- Analysis centers, pre-analysis, common software
- Analysis at facility vs. analysis at home
- Online, on-the-fly analysis (triggers & filters), never filter?
- Data continuum, 'living publication' (Helliwell et al.)
 - Publication together with data, registration of publications, X-referencing
- Authentication
 - See next slides

All these topics require substantial resources. Facilities need user feedback on priorities



User ID, Authentication, Authorization

Need for User ID

- EU-wide, trans-facility
- Persistent

Basis for practically all new developments

Element in all EU projects discussed

Properties required

- Technical
 - o State of the art protocols, e.g. Shibboleh (hackers!)

Management

- o Fit to characteristics of community
 - Cooperation and(!) competition
- o Respect confidentiality and autonomy requirements

Character

o Slim, very limited resources



Umbrella as solution

Incorporate confidentiality aspects

- High competition, especially structural biology
- Time-window structured access to experiments and data

Rely on existing local user office structure

- ✤ Great experience
- DIY (Do It Yourself) operation
 - o Users: manage their personal entries
 - o User offices: supervising; manage authorizations

Base system on professional authentication standard

- Shibboleth, federated Single-Sign-On System (SAML), widely used
- Special photon / neutron user federation
- Only one identity provider
- Supervising by local User Offices

Concept

- Unique user identification on EU (transfacility) scale
- Hybrid information storage
- No automatic cross-facility information exchange
- Waterproof but slim data protection system



The Umbrella Concept



ECM27, Bergen DDD Workshop August 6, 2012



The Umbrella Concept



ECM27, Bergen DDD Workshop August 6, 2012



Hybrid concept (central and federated)

Answer to conflicting requests:

- ≻Efficient technology
- ➤Confidentiality
- Consequent distinction of authentication and authorisation





UPS characteristics

Present situation:

- Heavy administrative load on users
- No synchronization in call for proposals
- No EU proposal standard
- Start always from scratch in spite of iterative character

Umbrella answer: subdivision into different parts

- ✤ Statistical
- Facility
- General (science)

Umbrella solution characteristics

- Federated proposal storage at facilities
- Compatibility with existing proposal handling
- Federated hybrid user database
- No Cross / trans-facility actions
- User: significant reduction of administrative load
- Facilities: no change in proposal handling work flow
- Proposals are key elements for remote data access



Remote data access, concept proposed

- **Embargo vs. post-embargo period**
 - Here only embargo (most critical, confidentiality)
- Standard access rights rule
 - No chance for manual central authorization
 - 1'000s of experiments, 10'000s of users
- Identity by Umbrella
 - Unique, EU-wide user authentication
- Keep Role of proposal as organizing element
 - Users convene for a short time slot for performing an experiment
 - Principal investigator / main proposer
 - Who participates in experiment, has access right to data
 - Proposal officially accepted by facility, PI is official contact
 - PI defines who participates in the experiment







Umbrella collaborators

- □ ALBA (**P**)
 - Joachim Metge
- DESY (C+P)
 - Frank Schluenzen, Rolf Treusch, Jan-Peter Kurz, Ulrike Lindemann
- DIAMOND (P)
 - Bill Pulford
- Fermi/Elettra (P)
 - Cecilia Blasetti, Ornela Degiacomo, Giorgio Paolucci
- EMBL HH / Biostruct X
 - Johannes Schmidt
- ESRF (C+P)
 - Rudolf Dimper, Dominique Porte, Stefan Schulze
- European XFEL (C)
 - Krzysztof Wrona
- Friedrich Miescher Institut
 - Dean Flanders, Roger Schmidt
- **GSI (C)**
 - Peter Malzacher, Almudena Montiel

- □ HZB **(P)**
 - Thomas Gutberlet, Dietmar Herrendoerfer, Olaf Schwarzkopf
- □ *I LL (C+P)*
 - ✤ Jean-Francois Perrin, F. Festivi
- □ ISIS (**P**)
 - Tom Griffin
- IPJ (Poland)
 - Robert Nietubic
- MaxLAB
 - Ulf Johansson
- □ *PSI* (C+P)
 - Bjoern Abt, Stephan Egli, Stefan Janssen, Markus Knecht, Mirjam van Daalen, Heinz J Weyer
- Soleil (P)
 - Frederique Fraissard
- □ STFC (**P**)
 - Anthony Gleeson





Umbrella Technical Team

Umbrella Management Team

Facility		Management	Technical
Alba	Ρ	J. Metge	S. Vicente
DESY	P+C	F. Schluenzen	J.P. Kurz, U. Lindemann
DIAMOND	Ρ	B. Pulford	B. Pulford
Elettra	Р	G. Paolucci, C. Blasetti	F. Bille
EMBL HH& Biostruct X		J. Schmidt	J. Schmidt
ESRF	P+C	D. Porte	S. Schulze
European XFEL	С		
FMI		D. Flanders	R. Schmidt
GSI	С	P. Malzacher, K. Schwarz	A. Montiel Gonzales
HZB	Р	Th. Gutberlet	A. Tomiak
ILL	Ρ	JF. Perrin	F. Festivi
ISIS & STFC	Ρ	T. Griffin	A. Wilson
PSI	P+C	S. Janssen D. Feichtinger	M. Knecht
Umbrella team	P+C	B. Abt, M. Van Daalen H.J. Weyer (lead)	B. Abt (lead) M. Van Daalen H.J. Weyer

ECM27, Bergen DDD Workshop August 6, 2012



Range of authentication /access control

- Present discussions
 - Only at facilities
- **Given Setup**
 - Interest in extending to simple system
 - o At home institution
 - o Clouds

Discussion needed bw facilities and users



Federated Identity Management

History

- Started by IT leaders of EIROforum (European laboratories)
- Lead by CERN
- Search for a common federated AAI system
- Wide range of research communities (HEP, Life sciences, Humanities, P/N facility users, Climate research)
- Activities
 - Draft FIM paper
 - Past workshops (CERN, RAL, Taipei, Nymegen)
 - Upcoming workshops (Washington (fall)?, PSI (spring 2013))

Next steps

- One academic identity system?
 - Many different requirements (library-type -> research facility)
- Federated system?
 - o Bridging, flexible interface definitions



FIM and New vistas (1)

Bridging, different federations

There will always be many federations

- o Banks, airlines, medical sector, government sector, academic, Facebook, Google, ...
- CRISP
 - o Partly topic of WP16 (PSI and GSI)
- Different options how to deal with
 - o No answer, islands
 - Too dangerous, do not trust
 - o Fully transparent
 - Risky
 - o Bridging
 - User can e.g. bring her/his attributes from Facebook





FIM and New vistas (2)

- Bridging, different federations
 - * ...

New media, how do we deal with them

- Support or 'You are entering the wilderness'
- Fora, Facebook
 - o Facility operated, 'info trees' (EuroFEL, CALIPSO), Wikis
 - There is a need, but labor intensive
 - o Commercial, User driven (Facebook, Google+)
 - Researchers: info exchange
- Clouds
 - o Community driven
 - Helix Nebula, High interest in further development
 - o Commercial
 - Users: analysis, publ. preparation (repl. for email)
 - Let them just do or give support and coordinate?



Conclusion

Several EU initiatives interesting for users

- Approach is to see all issues related to experimental data in one common view
- Access support
 - o Optimize resources
- New developments, trends
 - o Facilities, detectors, new IT-tools

Trans-facility actions

- o First step: cooperation of IT responsibles from different facilities
- o Next steps: cooperation with users
- o Extremely exciting ideas on data continuum in this workshop
- o But realization possible only if based upon a solid IT basis
 - Trans-facility aspects
 - Exploiting of synergies
 - Common voice towards decision makers

Cooperation and feedback between facilities and users essential

o IUCr represetative as guest at PaNdata?



Thank you

ECM27, Bergen DDD Workshop August 6, 2012

Heinz J Weyer, PSI 39