Janet UK meeting

Diamond experience with Project Moonshot

York, 5 April 2012

Bill Pulford

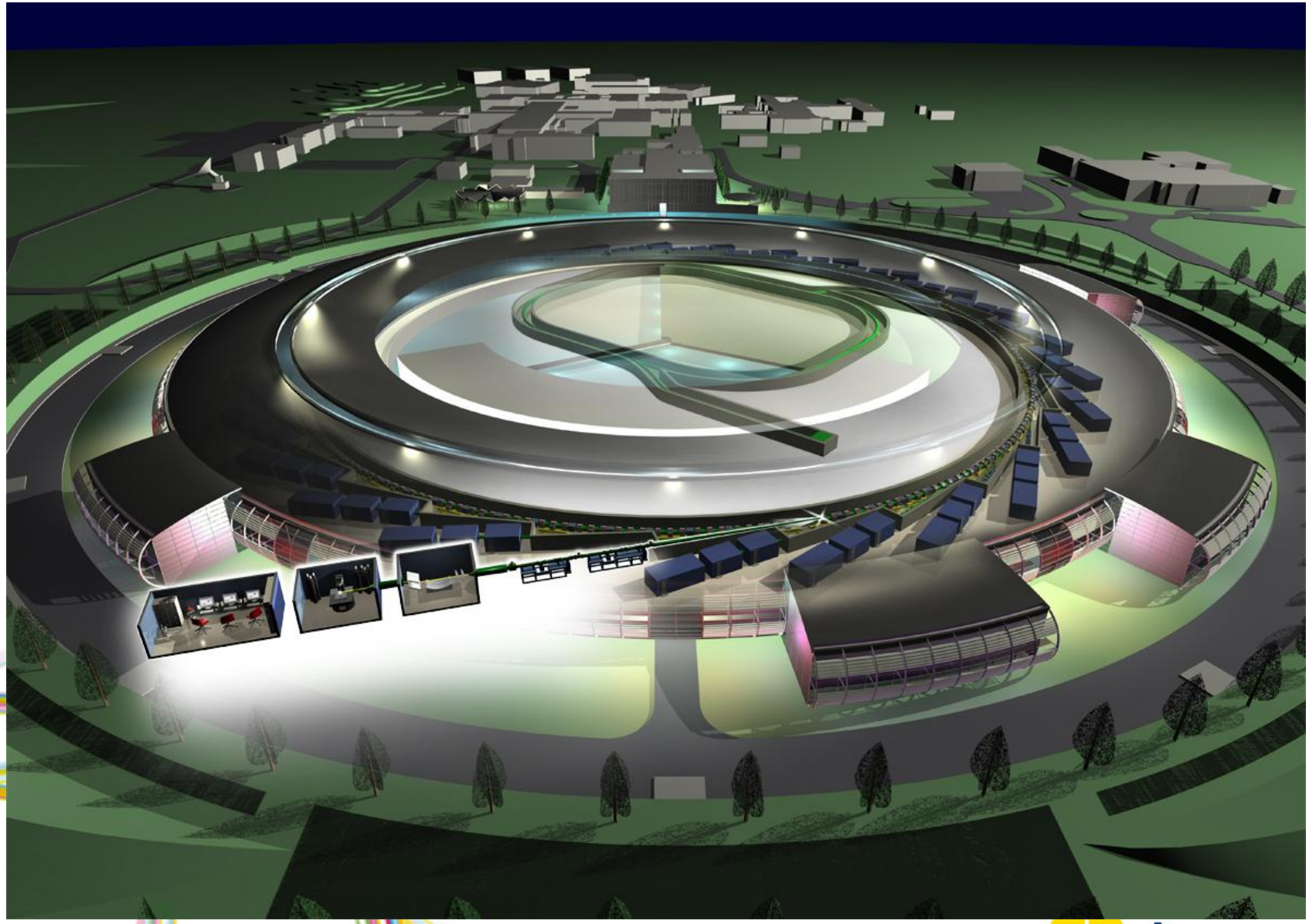
Head of Data Acquisition and Scientific Computing

Diamond Light Source

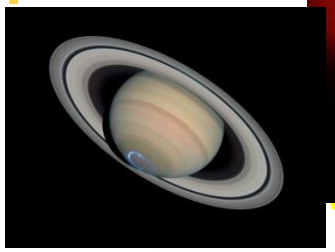
Three Linked Contributions

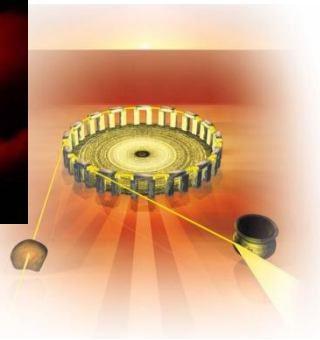
Diamond Light Source	Brief Overview
PANData	Brief description of aims
MOONSHOT	How the above come together with Janet and Moonshot.

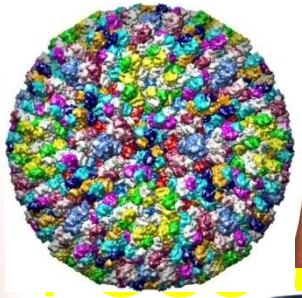
Diamond



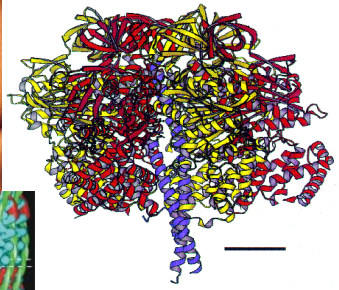
pandata europe

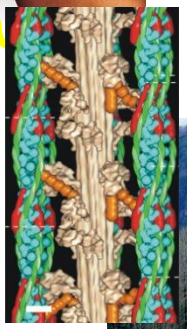


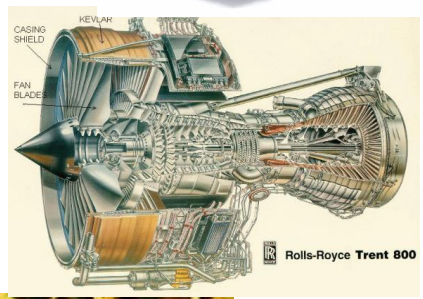


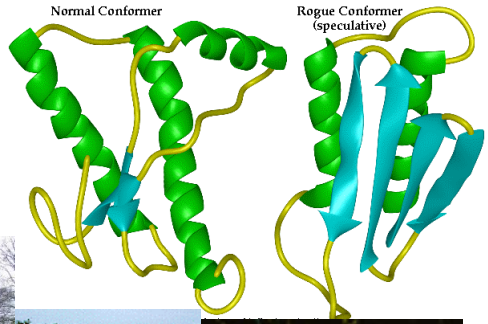


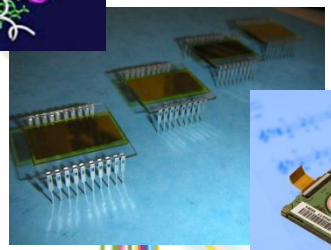
janet

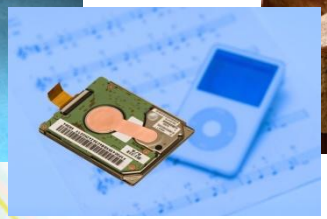


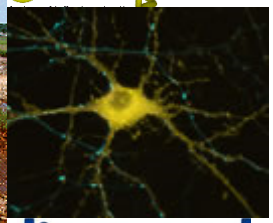








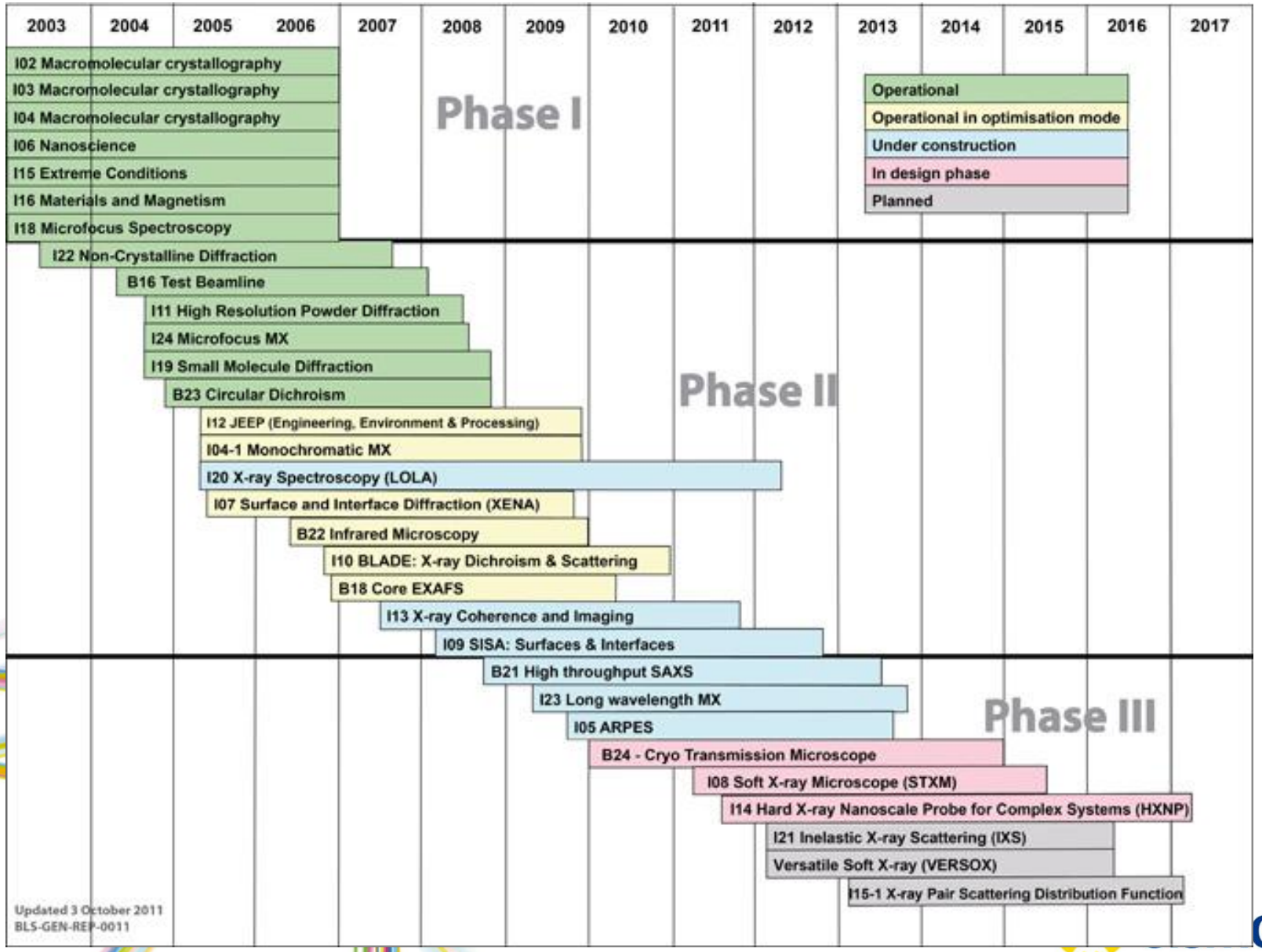




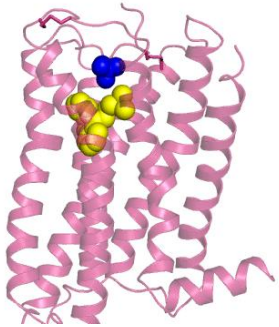
diamond

Diamond Light Source in reality

The science done at Diamond

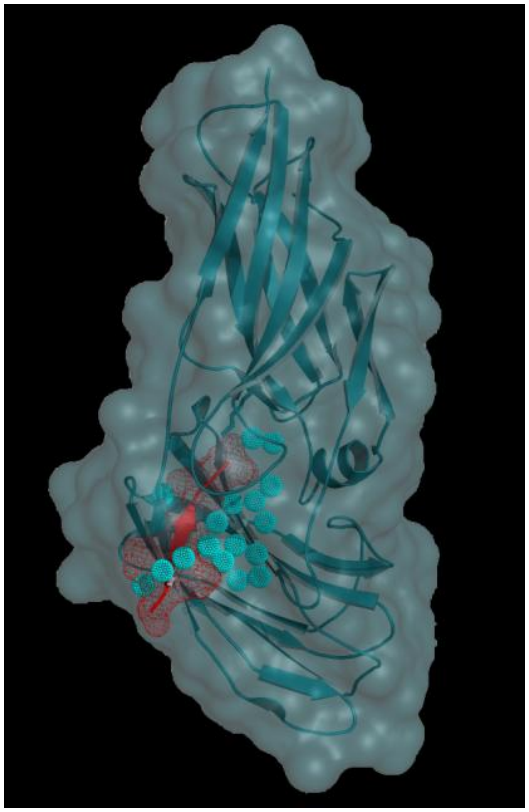


Health and disease

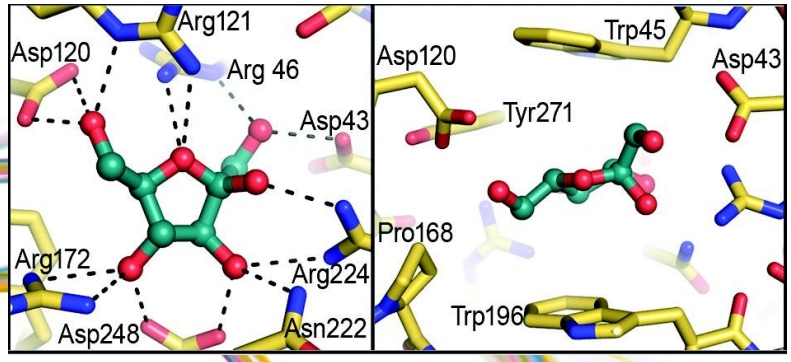


Structure of the
Histamine H1 receptor

Understand rejection in
hip implants



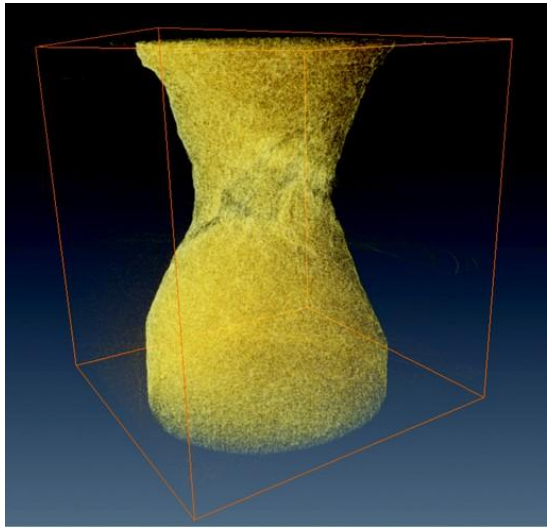
New drug target for
Candida infection



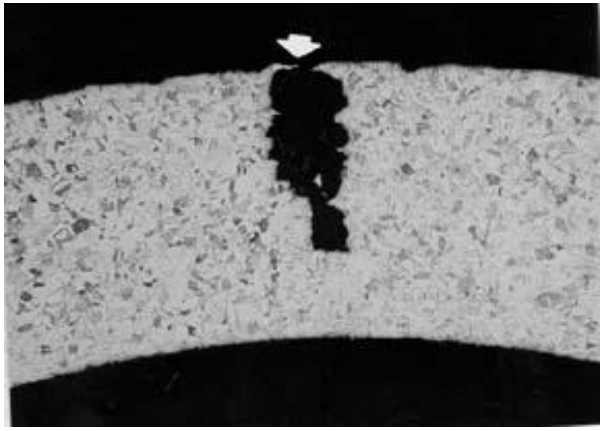
Sugar binding in gut flora

Improving nutritional
quality in wheat

Engineering and manufacturing

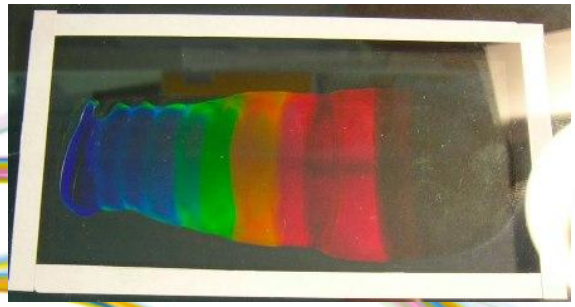


Casting aluminium

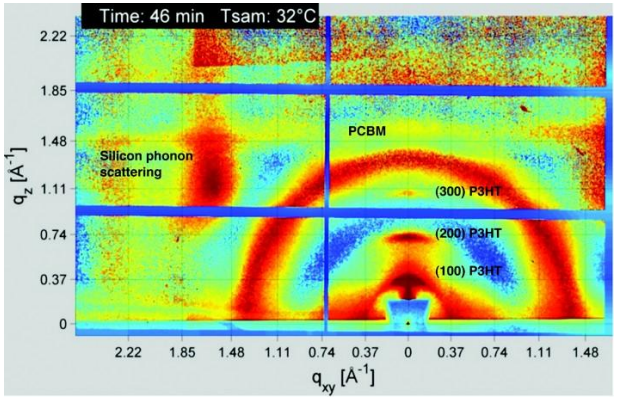


Understanding the corrosion process

Pharmaceutical manufacture and processing

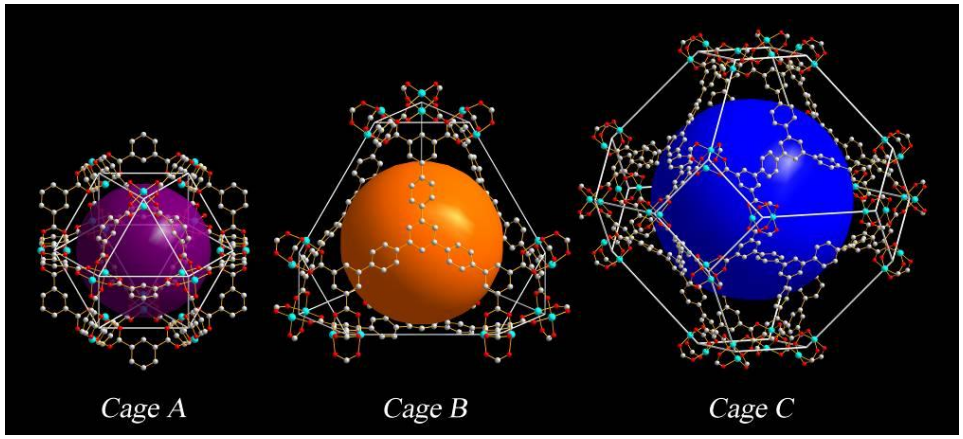


Tunable polymers

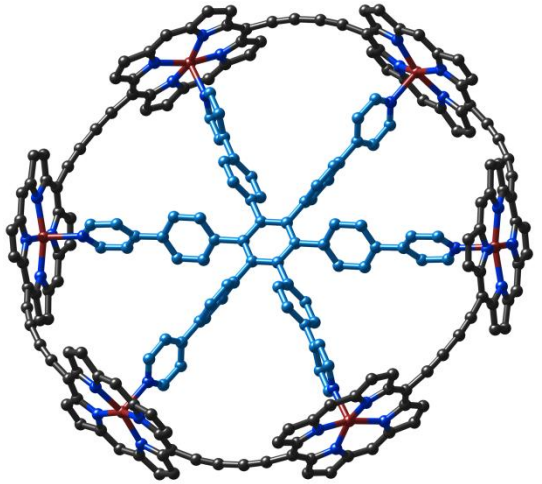


Organic photovoltaics

New materials

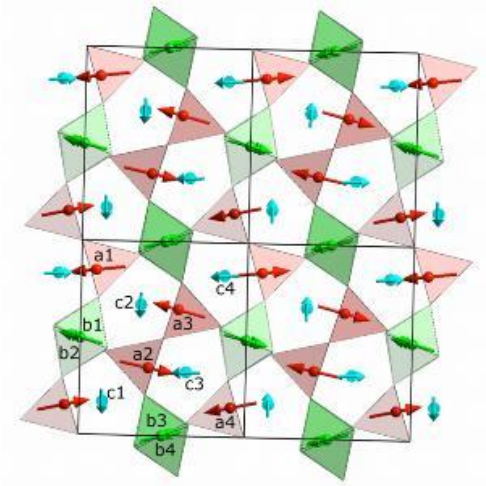


MOFs for hydrogen storage



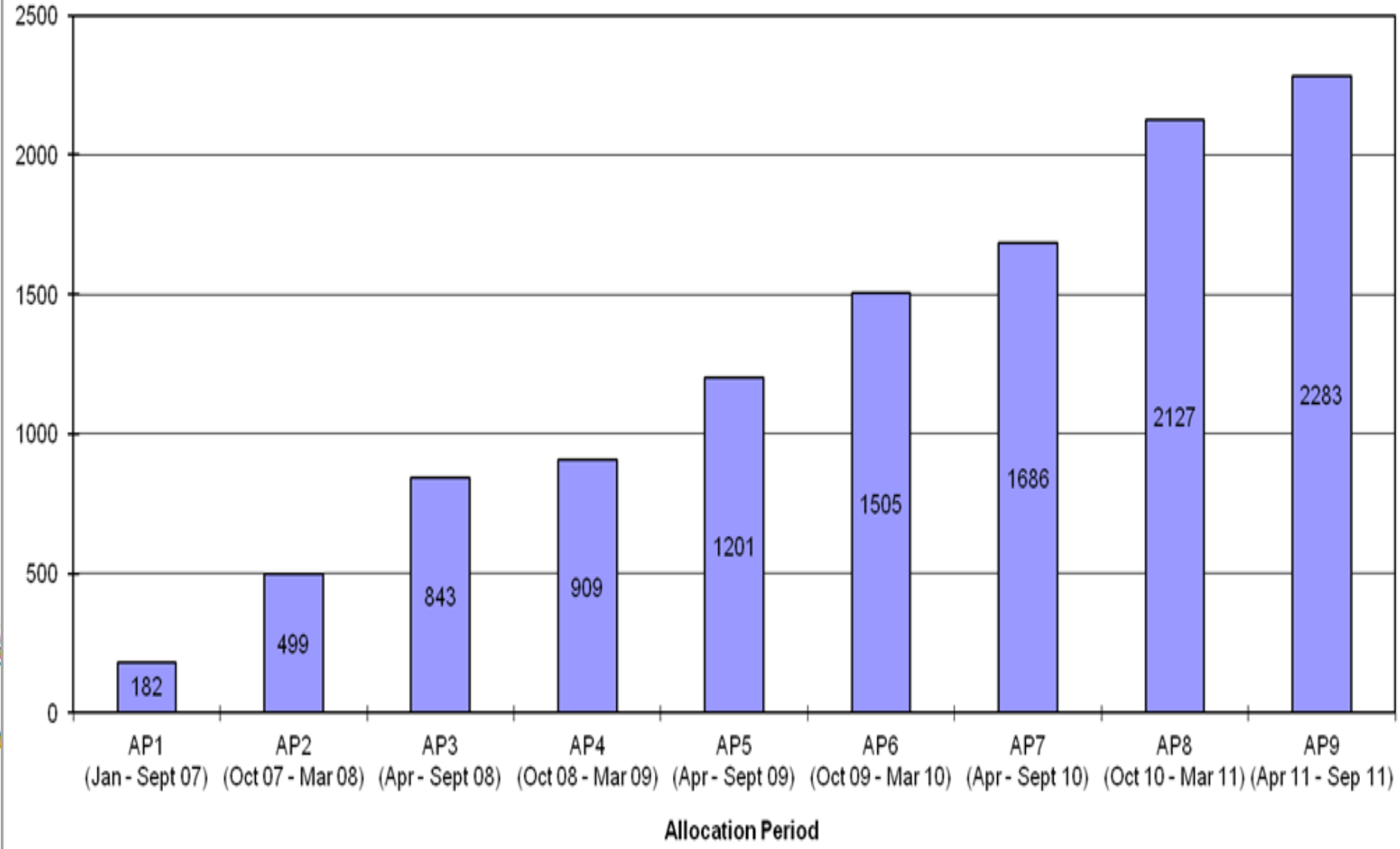
Harry's wheel – complex templates for new materials

Bio-mimetics



Multiferroics – electronic storage and memory

Number of External Users on Experiments



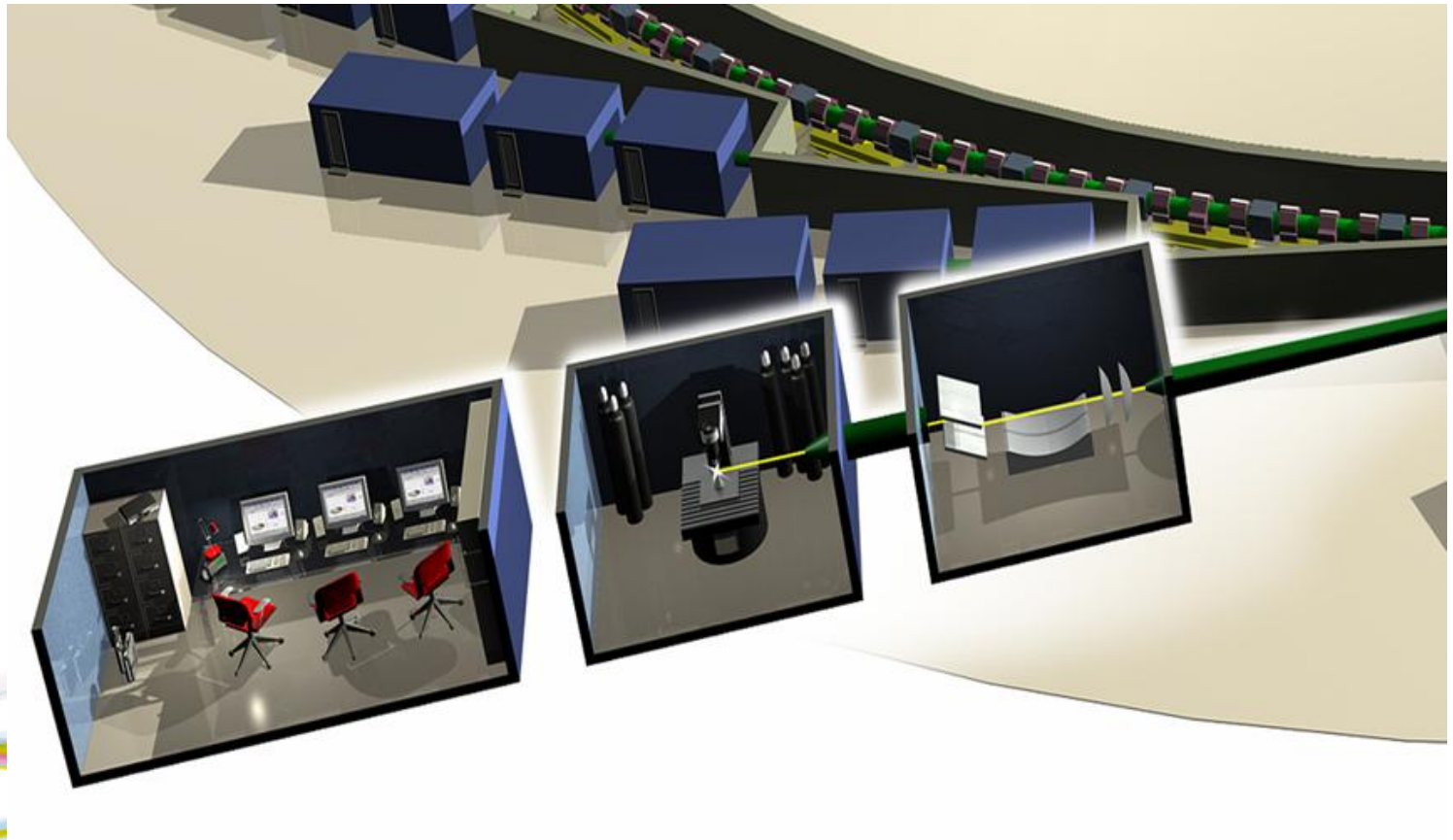
Overall Mission for Data Acquisition and Analysis

- Our overall mission is that users in all of our supported scientific disciplines be able to acquire data using the most flexible and capable tools that we can provide and then leave Diamond having at least evaluated their quality but ideally also having been able to perform the appropriate analysis.
- These capabilities are provided by a number of very advanced tools particularly Generic Data Acquisition(GDA) for acquisition and Scientific Data Analysis workbench (SDA) for analysis.
- These tools coupled with highly evolved software workflows, bespoke parallelized applications and powerful an underlying architecture of processing systems and networks enable us to provide our experimenters with some of the most advanced Acquisition and Analysis available.

Most automated processes require Single Sign On.

In the year between 1-Jan-2011 and 1-Jan-2012 we had 1726 experimental visits and 4976 external experimenters of whom 1976 where unique. Currently we have Data Volumes and Numbers of files (ICAT) ~ 228Tb/95,000,000

Overall Experimental Layout



PANData – a Brief Description

PaNdata brings together thirteen major world class European research infrastructures to create a fully integrated, pan-European, information infrastructure supporting the scientific process. PaNdata launched two projects supported by the European Commission to achieve this long term goal.

PANDATA consists of two parts:

1. Europe was a pure support activity, laying the foundation for a federated data infrastructure by developing a policy and software framework. This project has been successfully concluded in November 2011. Diamond were responsible for the selection of the Authentication system.

2. Open Data Infrastructure (PaNdata ODI) project has started to work on the implementation of a federated data infrastructure, in particular on software and data catalogues, user identities and authentication systems or optimized data analysis methods, to mention some of the core topics. PaNdata ODI runs until September 2014

The PANData initiative

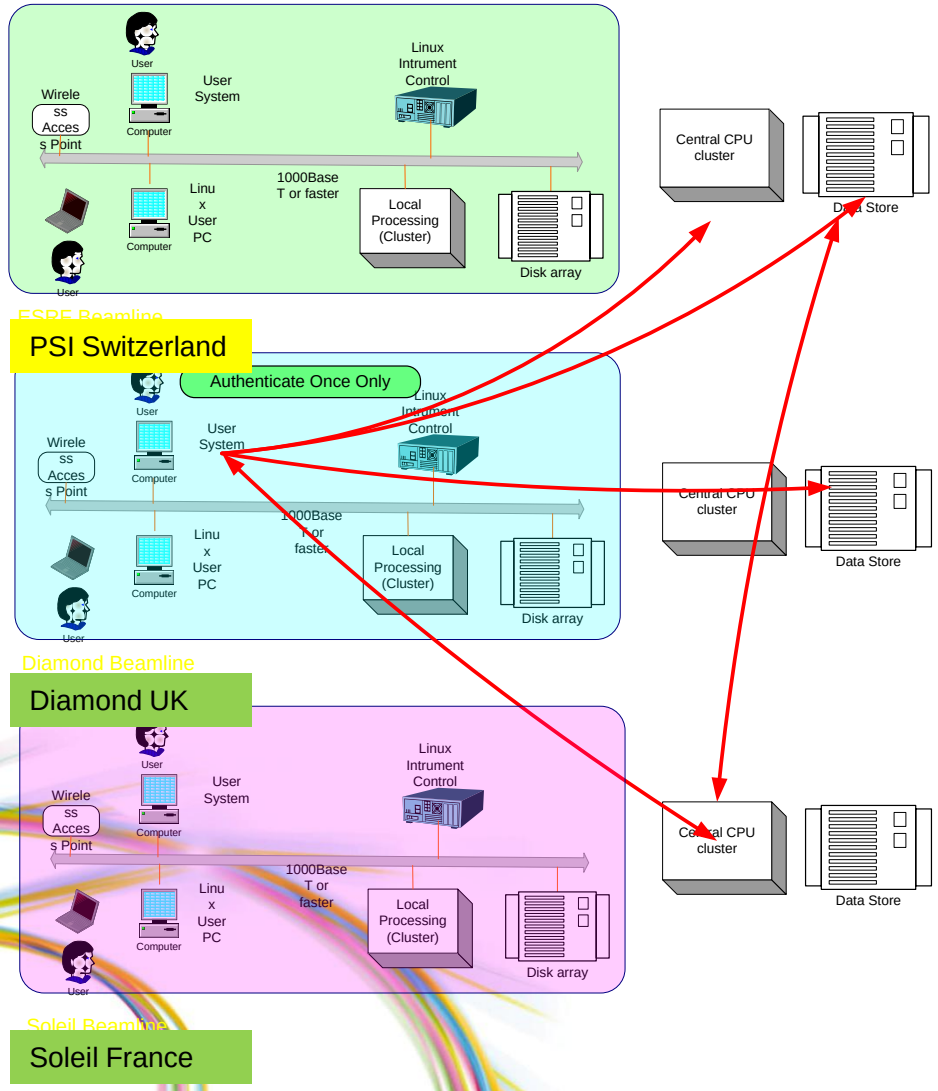
Five Photon and Neutron sites in Europe are in PANData:

- ISIS + DIAMOND
- SINQ + SLS
- ILL + ESRF
- HMI + BESSY, now the HZB
- LLB + SOLEIL
- (+ DESY, ELETTRA, and ALBA)

In addition a number of other European and Commercial projects are interested PANData; notably:

CRISP (Other scientific faciities including CERN, X-Ray Free Electron Laser (XFEL), ORCID and OpenAire+(Publishers),Biostruct and Callipso(Biomolecules)

Example Use of PANDATA and AAI

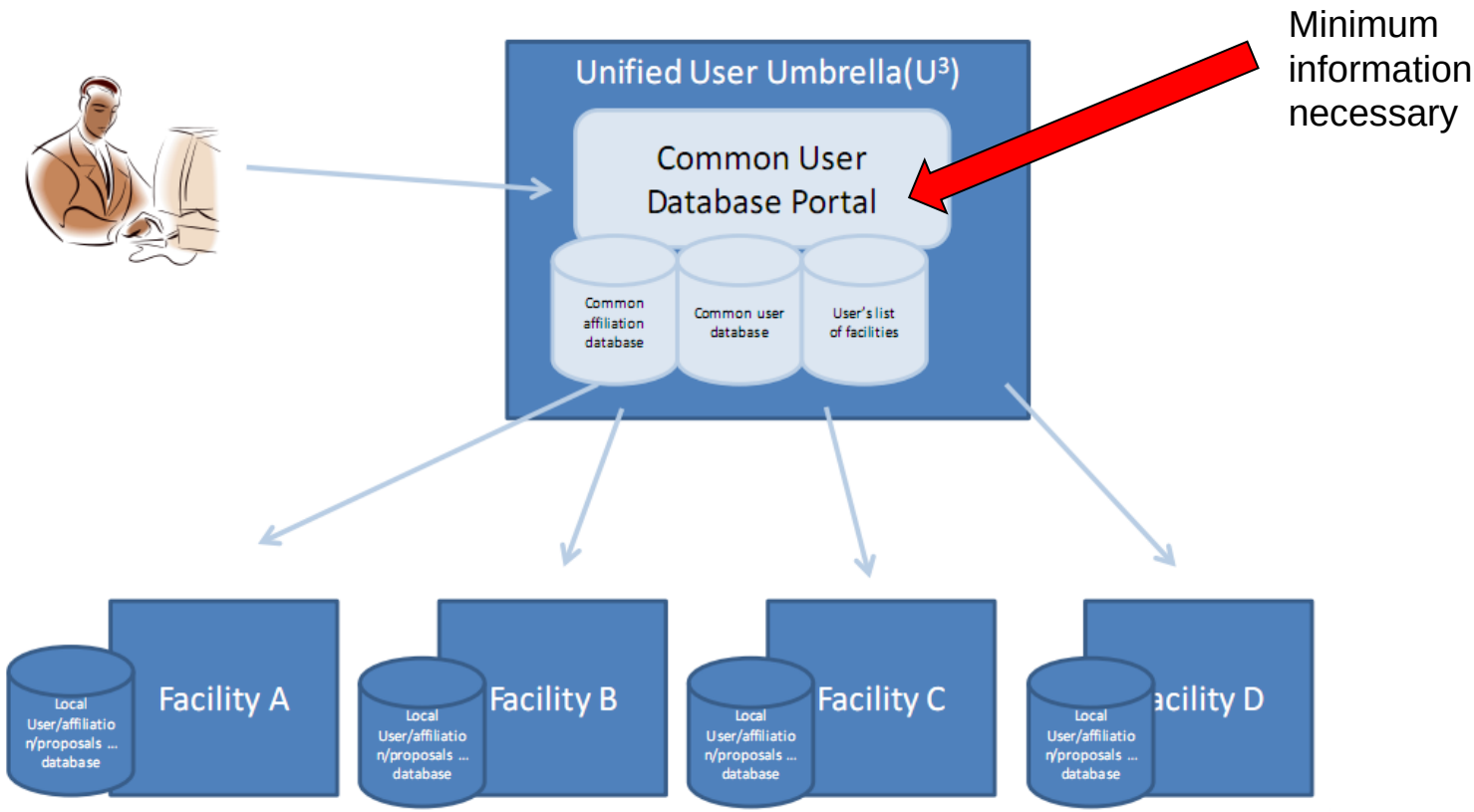


A user can perform the entire experimental process from proposal submission through data acquisition, data analysis to publication using one set of credentials.

The diagram shows data acquisition at one facility while consulting data at another and using CPU resources of a third.

The Umbrella Project Authentication

The minimum user information possible is stored centrally to avoid Data Protection issues. The Authentication is done by the user logging into the Umbrella central site – currently umbrella.psi.ch – to generate a Shibboleth token. Authorization is delegated to the facility site.



Umbrella Extended Functionality

- *Umbrella Plus*
  - Proposal-based user administration
    - Linking via Umbrella to local WUOs: includes full user services
    - Remote file access, remote experiment access + …
  - Non-proposal-based user administration
    - HEP-type operation (very long-term proposals)
    - Small facilities (e.g. university labs, …)
    - May have need for user db, but not for the rest
    - Umbrella + stripped-down version of a WUO
      - Core user db
      - Shibboleth communication
      - Green / red lamp at the output
- *Umbrella Bio*
  - Currently 2 decoupled user review/access schemes
  - Combine Umbrella + BioStruct

Recommendations from Pandata Europe WP 4 (AAI)

1. An authentication mechanism based on Umbrella should be implemented as widely as possible across participating facilities.

2. A technical solution needs to be provided to allow interactive sessions on computing resources at the facilities' sites. Initial coverage must include Linux and Windows (optionally Macintosh) systems in order for any complete adoption of the authentication system.

3. An important architectural advantage is offered by using a system such as the Jasig Central Authentication Service (CAS) where most of the internal Authentication and Authorization issues are covered by this one system thereby obviating the need to modify dependent systems.

Integration with Project Moonshot - Basic Requirements

- The deliverables of project Moonshot look to be appropriate to address the Roadmap requirement. Notes:
  a. In essence this would be to use the Umbrella shibboleth credentials to set up interactive sessions for systems including Linux, Macintosh and Windows; this would provide a fairly complete solution for proposal submission through data acquisition to data analysis with one single set of credentials.
- The major component would be the modified openssh provided with Moonshot. Ideally this would be modified to accept shibboleth credentials and then allow ssh sessions on Linux and Macintosh (probably). I understand that Windows authentication follows the same principles. Notes:
  a. An optional but very useful extension would be the availability of a Pluggable Authentication Module (PAM) that would address the issue of collaborating facilities having to support multiple authentication systems particularly during transitional phases.
- The ideal deliverable would be a package managed installation such as an rpm or msi file that could be installed by a Systems Administrator and then be configured for IdP, optional CAS and other items that the author has not considered yet.

Profitable interactions between developers

radiusd module problem			
radiusd module problem	Brian Abram <[log in to unmask]>	Tue, 22 Nov 2011 10:33:53 +0000	157 lines
Re: radiusd module problem	Brian Abram <[log in to unmask]>	Tue, 22 Nov 2011 11:14:28 +0000	143 lines
Re: radiusd module problem	Roland Hedberg <[log in to unmask]>	Tue, 22 Nov 2011 15:34:48 +0100	50 lines
Refactoring the moonshot-ui software			
Re: Refactoring the moonshot-ui software	Pete Fotheringham <[log in to unmask]>	Mon, 28 Nov 2011 14:45:33 +0000	45 lines
relocation error: /opt/moonshot/sbin/gss-server: symbol gss_pname_to_uid, version gssapi_krb5_2_MIT not defined			
relocation error: /opt/moonshot/sbin/gss-server: symbol gss_pname_to_uid, version gssapi_krb5_2_MIT not defined	Brian Abram <[log in to unmask]>	Wed, 9 Nov 2011 12:44:54 +0000	204 lines
shibboleth/RequestUnsupported			
shibboleth/RequestUnsupported	Brian Abram <[log in to unmask]>	Fri, 25 Nov 2011 12:06:15 +0000	115 lines
shibboleth/RequestUnsupported	Brian Abram <[log in to unmask]>	Fri, 25 Nov 2011 12:50:30 +0000	104 lines
Re: shibboleth/RequestUnsupported	Roland Hedberg <[log in to unmask]>	Fri, 25 Nov 2011 16:15:50 +0100	23 lines
Re: shibboleth/RequestUnsupported	Cantor, Scott <[log in to unmask]>	Fri, 25 Nov 2011 18:19:23 +0000	25 lines
Re: shibboleth/RequestUnsupported	Josh Howlett <[log in to unmask]>	Fri, 25 Nov 2011 18:25:12 +0000	29 lines
Re: shibboleth/RequestUnsupported	Brian Abram <[log in to unmask]>	Fri, 25 Nov 2011 18:55:17 +0000	56 lines

FreeRADIUS - Working with multiple SAML IdPs and profiles

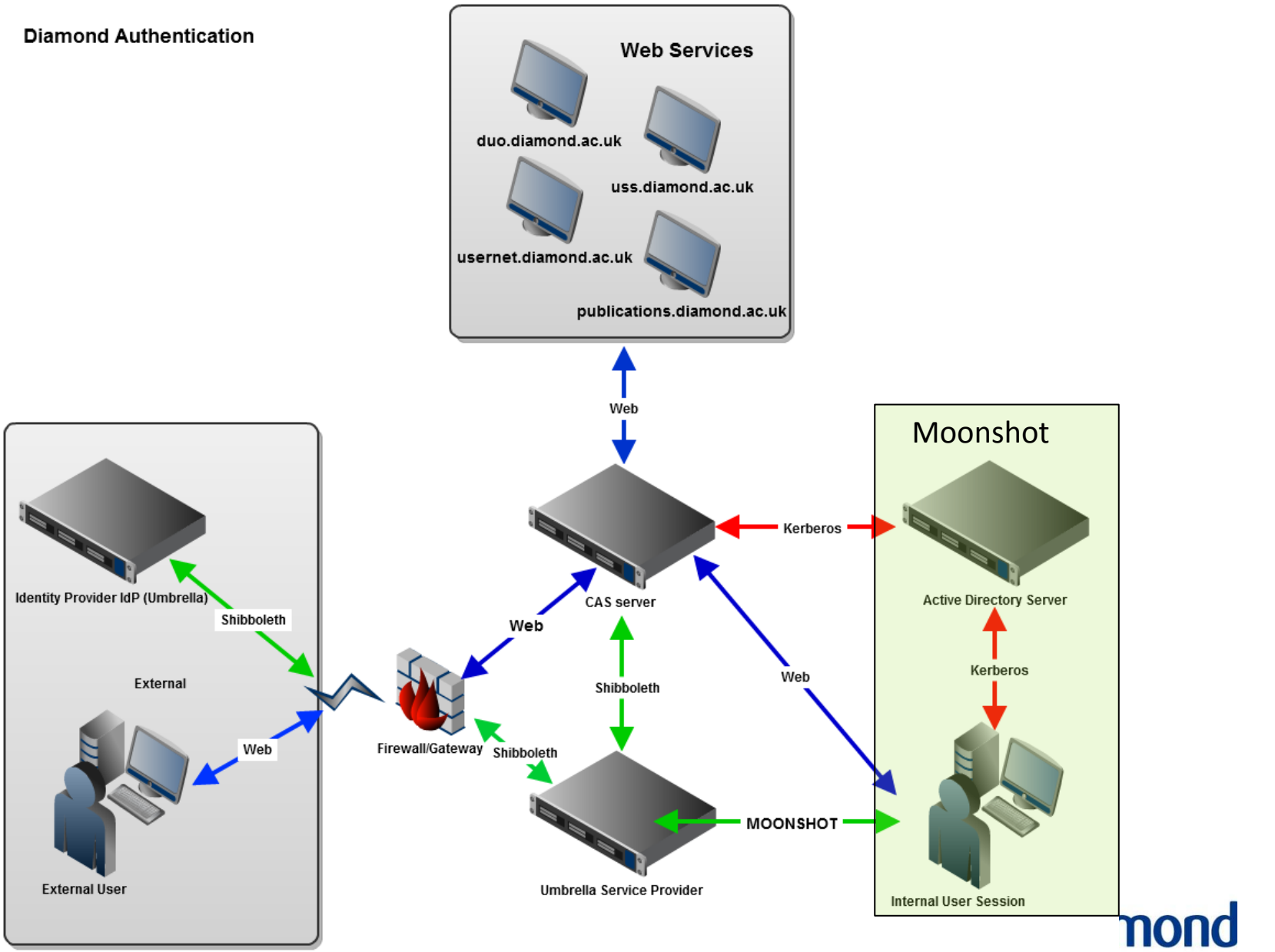
During development, testing and deployment it is often necessary to switch between different configurations. Here we identify the files that make up a profile in order to understand what would make it easier to switch between profiles with the minimum effort.

Code (Freeradius -2.0.0.8 RH6)	Comment
aa_config.py, ecp_config.py	This is the master configuration file, which determines which overall profile is required. This could also be known as idp_profile_config. Propose merging these into a single file, using a conditional for AA/ECP -> config.py
pysaml_config.py sp.xml	This contains details about the service provider. It is unlikely to change on a real deployment. It is not thought to be necessary to have more than one copy of this.CHECK: Is that true when switching from AA to ECP? It is required for generating SP.XML which is sent to the IdP administrator
metadata.xml	One of these is required for each IdP. CHECK:For a specific IdP can the same file be used for both AA and ECP? Propose naming these uniquely and setting the appropriate name in config.py just before running RADIUS. -> psi-idp-metadata.xml , cu-idp-metadata.xml
freeradius_aa.py, freeradius_ecp.py	One for AA and ECP. Propose using a wrapper to decide which calls are required according to the profile -> freeradius_aaecp.py
modules_python_aa, modules_python_ecp	One of these is required for each variant. It tells RADIUS which function variant to call Propose a single generic module and delegate this selection to freeradius_aaecp.py ->modules_python

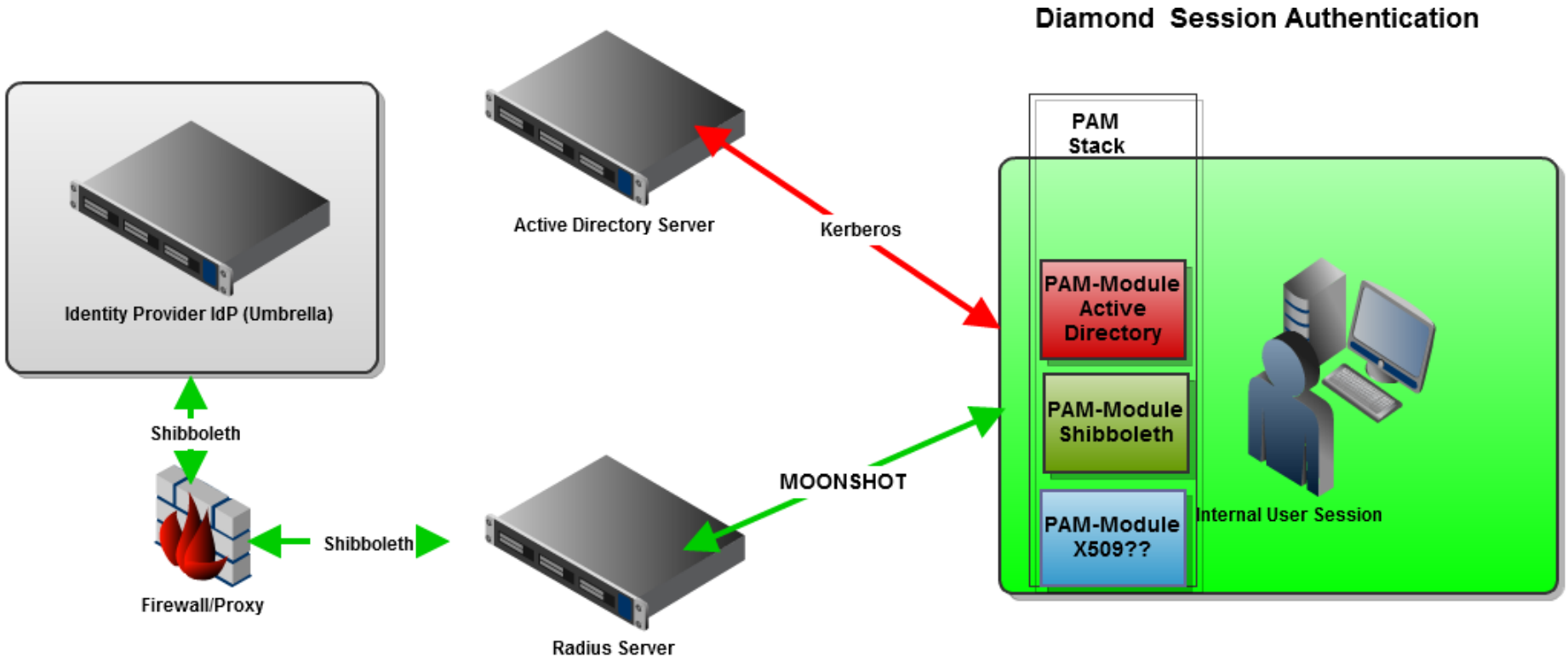
Setup notes for Moonshot implementation on the IdP:

On the IdP we had to enable following profile configurations in the relying party configuration for this specific Moonshot SP: SAML2ECPProfile
SAML2AttributeQueryProfile and SAML2ArtifactResolutionProfile
Make sure that all assertions and responses are signed, e.g. signResponses="always" signAssertions="always"

Potential Future Authentication Strategy



User Session Authentication with Moonshot



Issues for consideration

- Current systems can be difficult to define the duration for which users can be authenticated.
  a. X509 certificates may be useful to help with this problem
- A mechanism of persistence or caching is necessary to avoid loss of service due to network unavailability.
- PAM stacks. User X509 certificates
- Must support Windows, Macintosh and Linux
- Should be very simple to install and configure (consider rpm or msi)
- The overall network infrastructure should provide the necessary support.

Interesting:

- Compatible with any Microsoft initiative if possible.
- EduRoam
- Mobile devices

Acknowledgements:

a) Brian Abram (Diamond/Janet UK)
b) Bjoern Abt (PSI)
c) Roland Hedburg
d) Josh Howlett (Janet UK)
e) Rhys Smith (Cardiff)

Industrial

Dept.

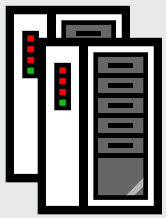
Academic

Dept.

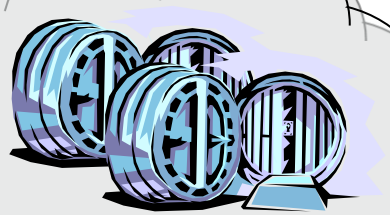
Site

Facility 1

Interf



High Power Cluster



Data Repository

Site

Facility 2

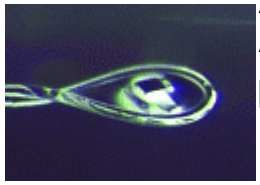
Cloud

Site

Publisher ORCID.O penaire

Back

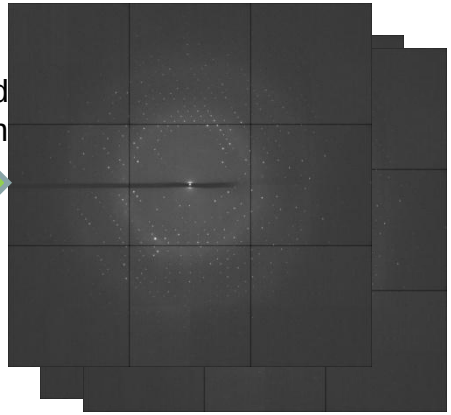
High Speed automated data processing for MX



Mounted macromolecular crystal

Automated Acquisition

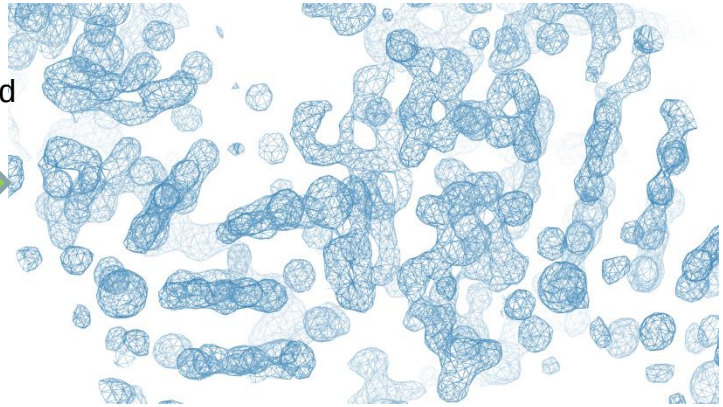
3 mins



Raw Diffraction data

Automated Analysis

2 mins



Electron Density map of protein/DNA (The 3-D structure)

Biological Macromolecule to structure in less than 5 minutes!

In more detail

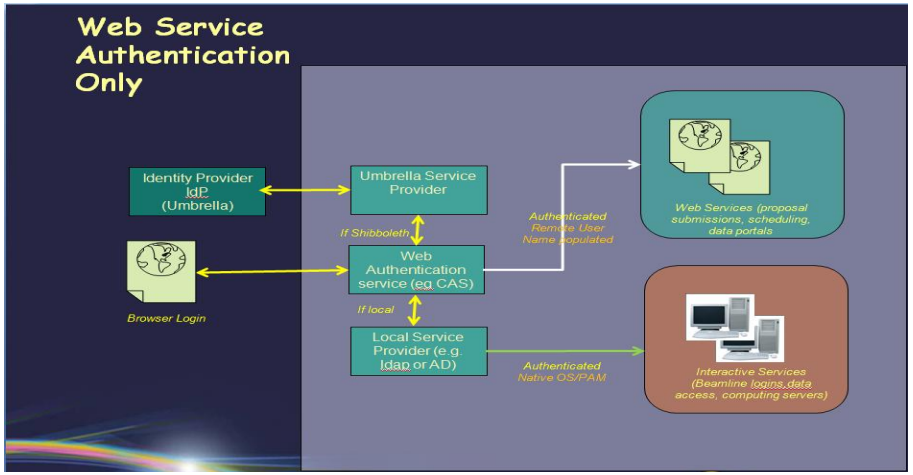
Raw diffraction data recorded in less than 3 minutes on Diamond MX beamline I02, as part of ongoing research into DNA / ligand structures, joint Ph.D. place between Diamond and Reading University (James Hall).

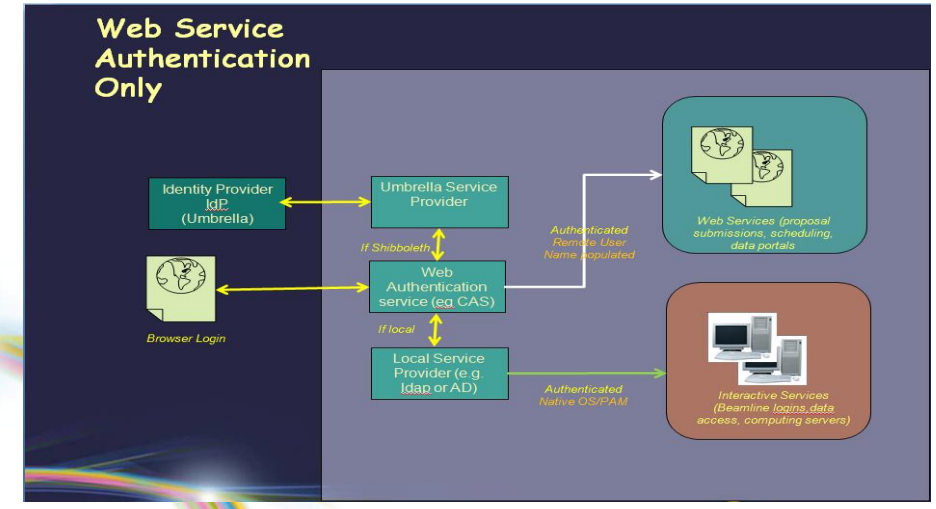
Automatically calculated map derived purely from experimental information, requiring a search over a number of parameters. Results were obtained less than two minutes after the data collection was completed, a step which would once have taken hours!

In both the processing and phase calculation extensive use was made of parallel processing, making use of one to two hours processing time in under two minutes.

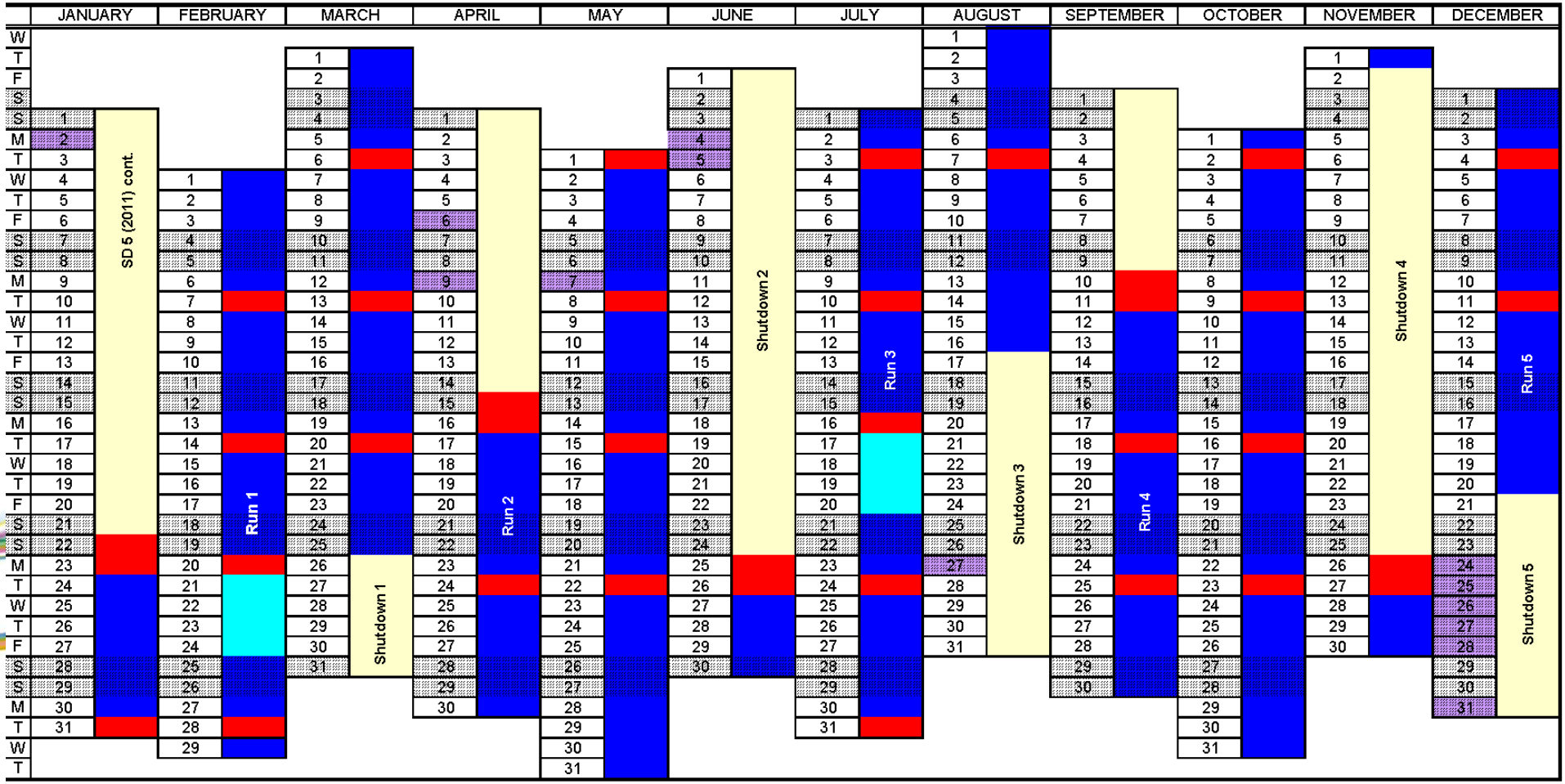
Thank you

Current and Future Visions

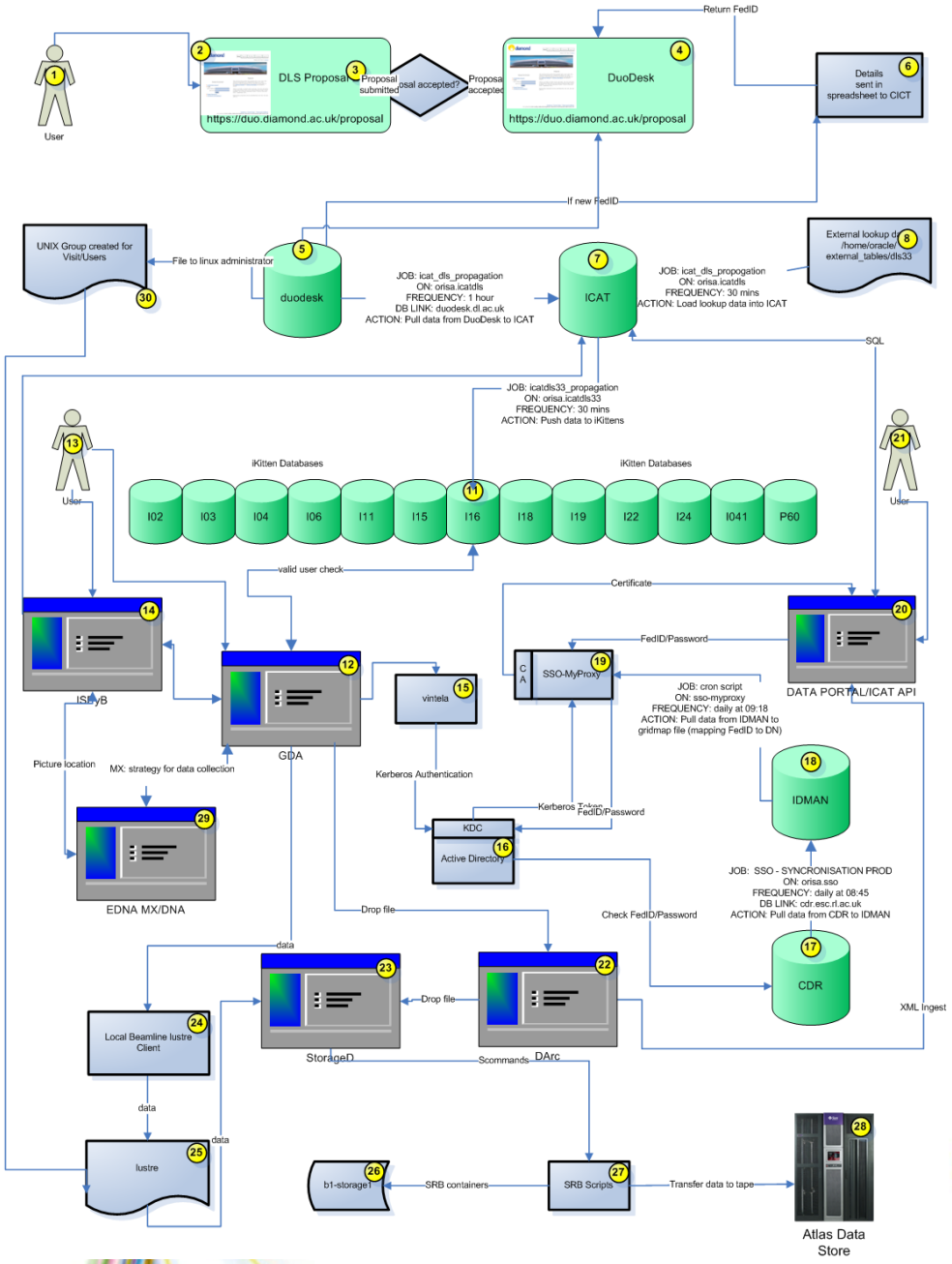




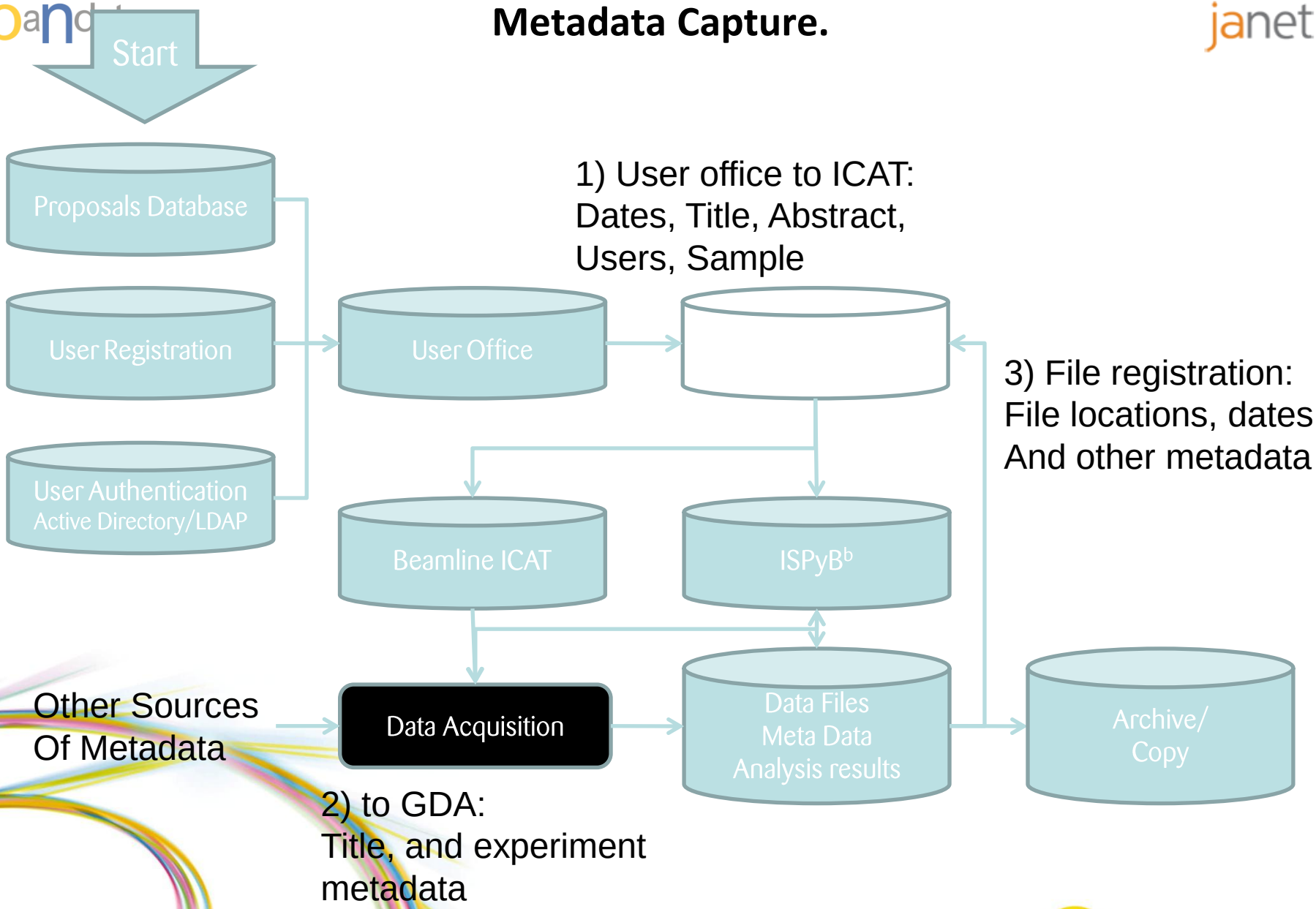
2012 Operations calendar



Diamond e-Infrastructure

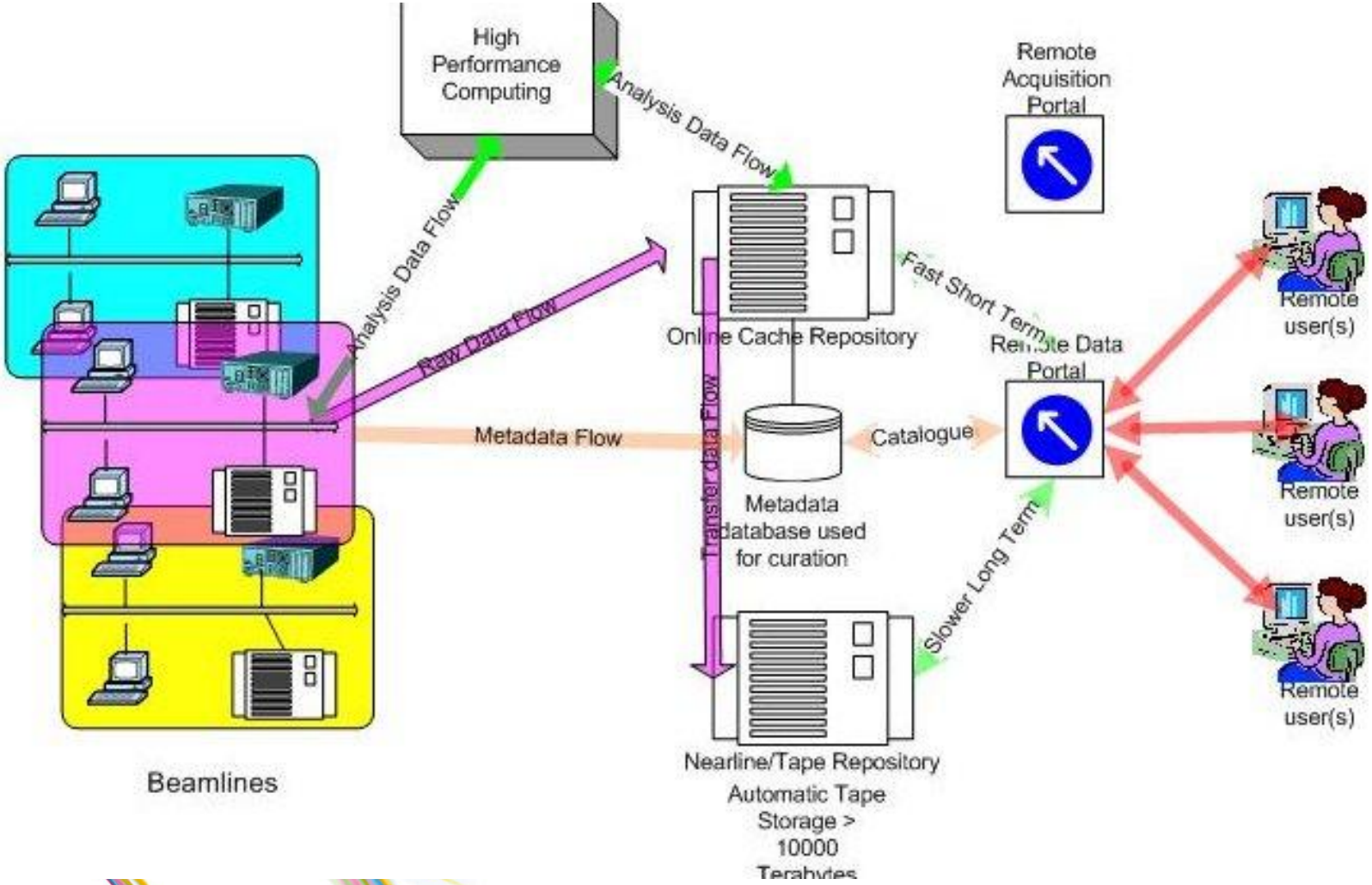


Metadata Capture.



[a] ICAT: http://code.google.com/p/icatproject/
[b] ISPyB: http://sourceforge.net/projects/ispyb/

Example Use of PANDATA and AAI

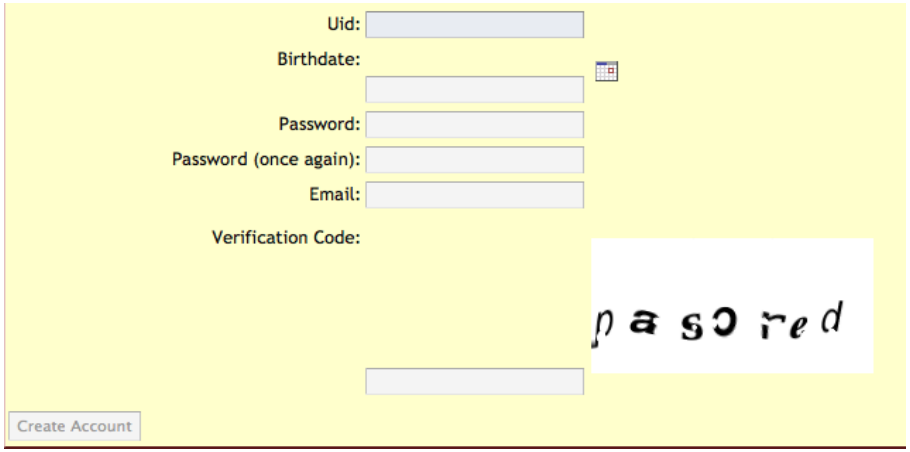


Requirements for cross facility AAI

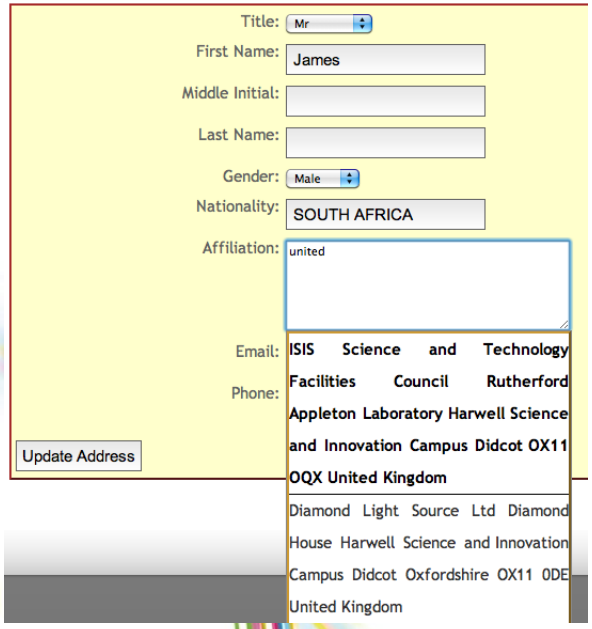
- The aim of this project is to provide a mechanism for uniquely identifying users of large scientific facilities and associated communities irrespective of their method of access. For the purposes of the report this authenticator would be given the name ESP or European Scientific Pass. (Consider ESI for European Scientific Identifier)
- All users of the major facilities will need only one username/password combination to access any of the facilities.
- The local user administration offices should be the principal controllers of authorisation.
- Users must be able to update their own passwords or other personal data using a “Bank Type” web application.
- The process of gaining a network identifier would require proof of identity but must be accomplished in a few minutes.
- These credentials or an automatically generated certificate or token will allow access to any computing technology given the correct authorization.
- The authorization will be performed locally by the facility involved based on the single unique identifier derived from 1-3.
- The AAI system would extend to scientists and engineers supporting data acquisition and analysis at their facilities. The authentication and authorization implicit in the system provides the necessary access security requested in the outline document.

Umbrella User Create/Update

Account creator



Account Updater



The EAA system …
is a distributed infrastructure for handling authentication at a super facility level.
will allow coexistence with the authentication mechanisms of the existing WUO tools.
stands for European Authentication and Authorization
handles user accounts.
allows Single Sign-On.
is the “workhorse” of the Umbrella system.
is non-invasive.

Overall Mission for Data Acquisition and Analysis

Our overall mission is that users in all of our supported scientific disciplines be able to acquire data using the most flexible and capable tools that we can provide and then leave Diamond having at least evaluated their quality but ideally also having been able to perform the appropriate analysis. These capabilities are provided by a number of tools particularly Generic Data Acquisition(GDA) for acquisition and Scientific Data Analysis workbench (SDA) for analysis. These coupled with highly evolved software workflows, bespoke parallelized applications and powerful an underlying architecture of processing systems and networks enable us to provide our experimenters with some of the most advanced Acquisition and Analysis available.

Ideally only logging in once!

The software developments in the past year have evolved rapidly from the solid base established in 2011 to allow increasing complex and potentially difficult experiments and then enable the experimenters to perform data evaluation and even detailed analysis during their visit to Diamond.

In the year between 1-Jan-2011 and 1-Jan-2012 we had 1726 experimental visits and 4976 external experimenters of whom 1976 where unique. Currently we have Data Volumes and Numbers of files (ICAT) ~ 228Tb/95,000,000