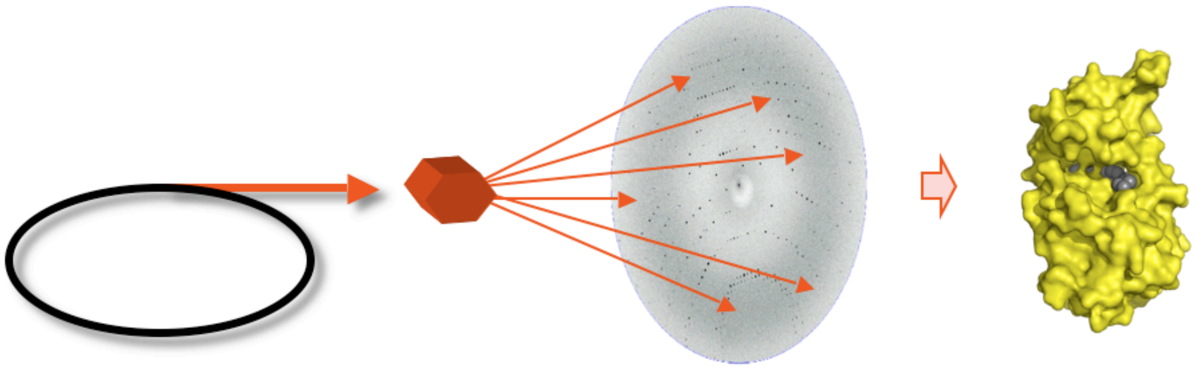
Data in Protein Crystallography

Thomas R. Schneider
Project Coordinator EMBL@PETRA3

Standard Data Formats for Experiments with Photons, Neutrons, Ions
DESY 27/10/2010

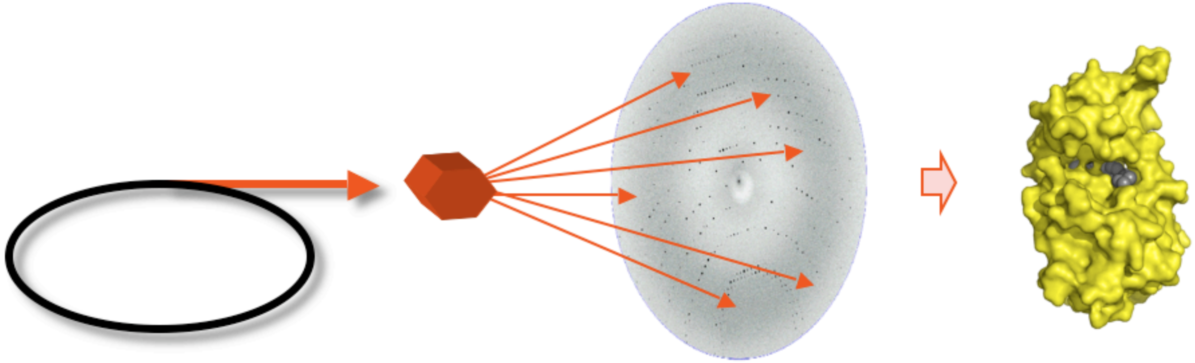
EMBL

Macromolecular Crystallography (MX)



- **Raw Data:** 2D diffraction images collected from a sample rotating slowly around a chosen axis.
- **Extra:** Many crystals have to be measured(100-1000ds) potentially spatially resolved; compatible data sets will be combined later on.
- **Boundary conditions:**
  - Users use various synchrotrons
  - Users are not physicists
  - 'Upstream' and 'Downstream' data are very important as well

The Plan

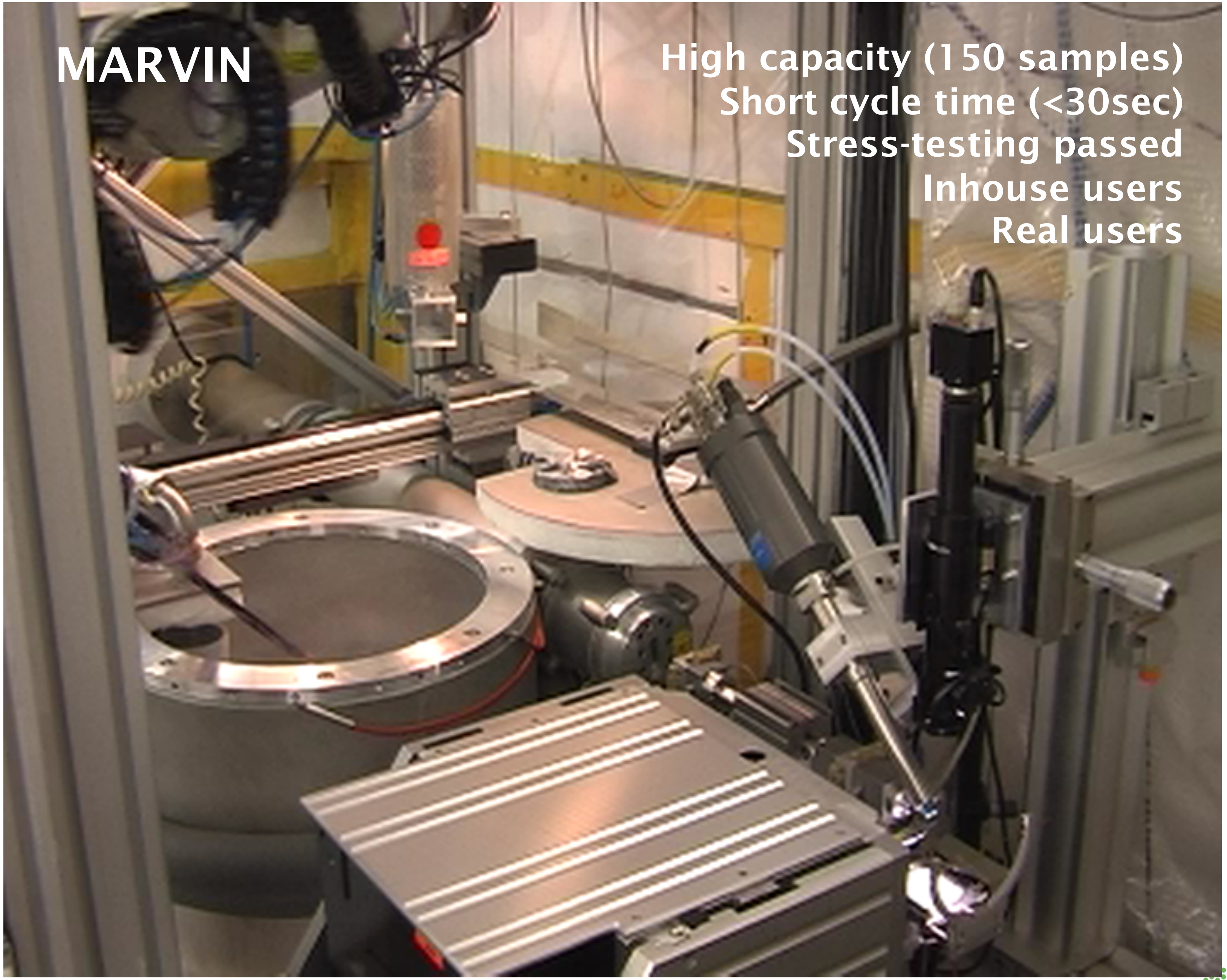


1. **Procedures producing data**
2. **Some approaches in use in MX**

Raw image data

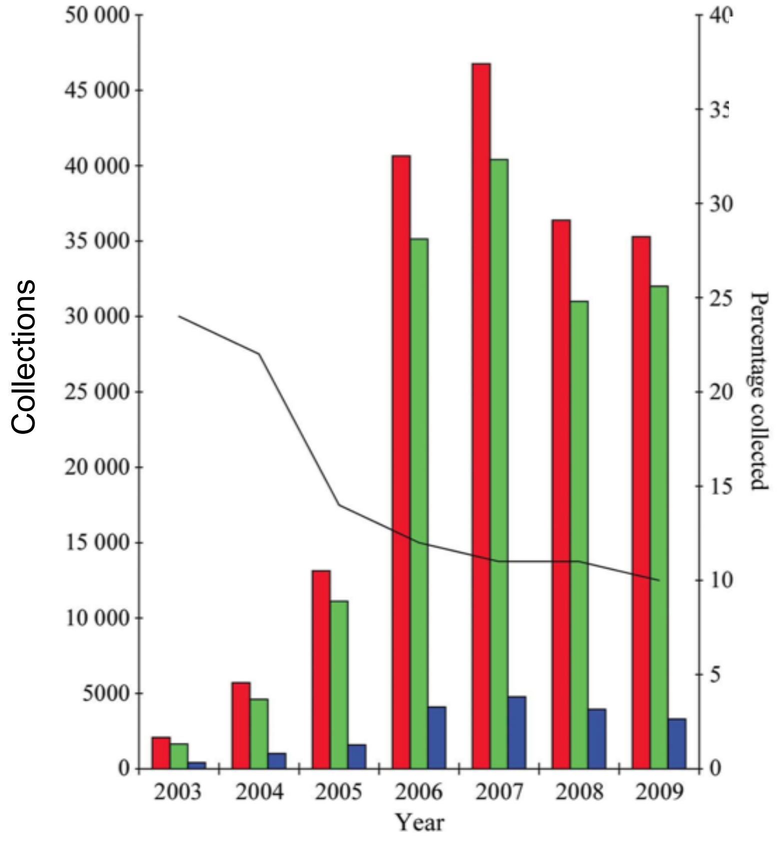
- A typical data set collected on a PILATUS 6M detector using the shutter-less rotation method can be collected in **2.5 min.** and produces ca. **4.5 GB** of raw data. 'In phase' processing is possible using local clusters.
- On a well equipped beamline, in principle, 250 data sets can be collected in 24 hours. -> **1 TB of raw data / day.**
- New technologies are already / will be soon available.
  - Frame-shift CCDs. Frame rate -> 100 Hz
  - Next generation PADs. Frame rate -> kHz, more pixels
  - Next generation amorphous Se detectors
- On 3rd generation sources, MX is not flux-limited -> more samples per time are possible.
- In principle there is CBF (Crystallographic Binary File Format, http://www.esrf.eu/computing/Forum/imgCIF/cbf_definition.html, cif-type header + binary image), however it can be difficult to motivate detector manufacturers to obey the rules (abuse of free format fields …).

MARVIN



High capacity (150 samples)
Short cycle time (<30sec)
Stress-testing passed
Inhouse users
Real users

Evaluation vs. Data Collection



🟥 Collections

🟩 Evaluations <4 images

🟦 Data sets >30 images

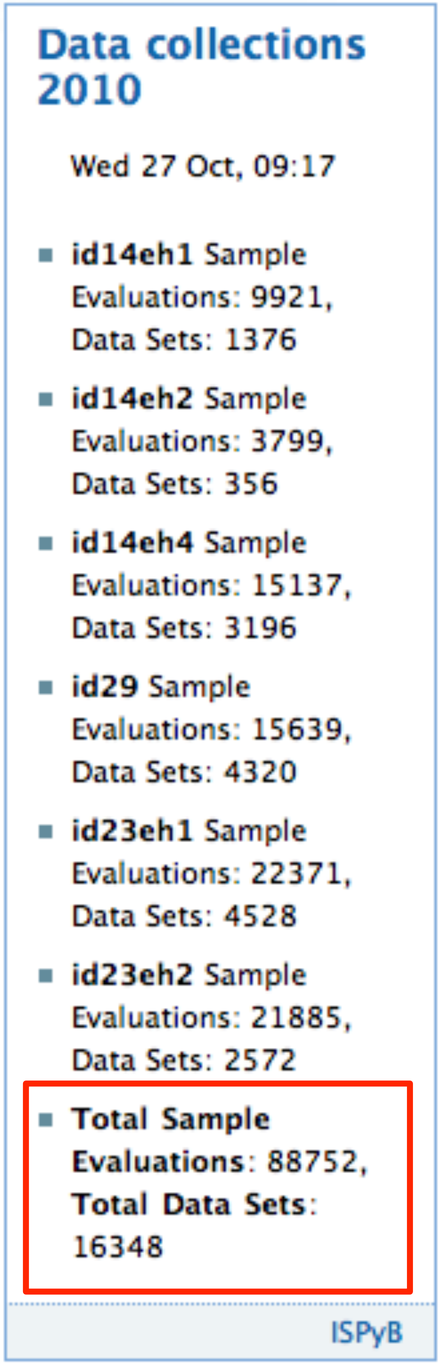
— Percentage collected

ID14-1/2/3

Bowler et al. (2010)
Acta Cryst D66:855

Sheer Volume

- http://www.esrf.eu / UsersAndScience/ Experiments/MX



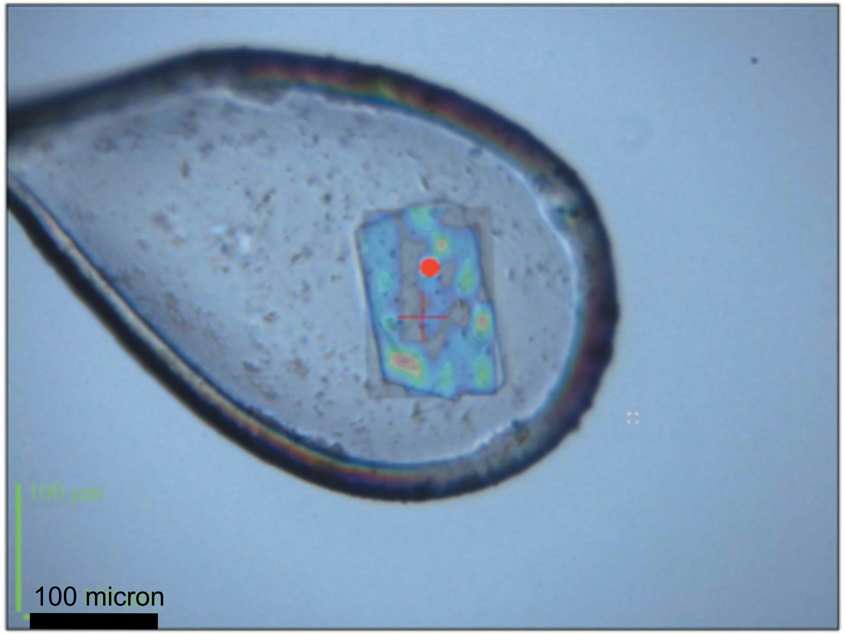
Data collections 2010

Wed 27 Oct, 09:17

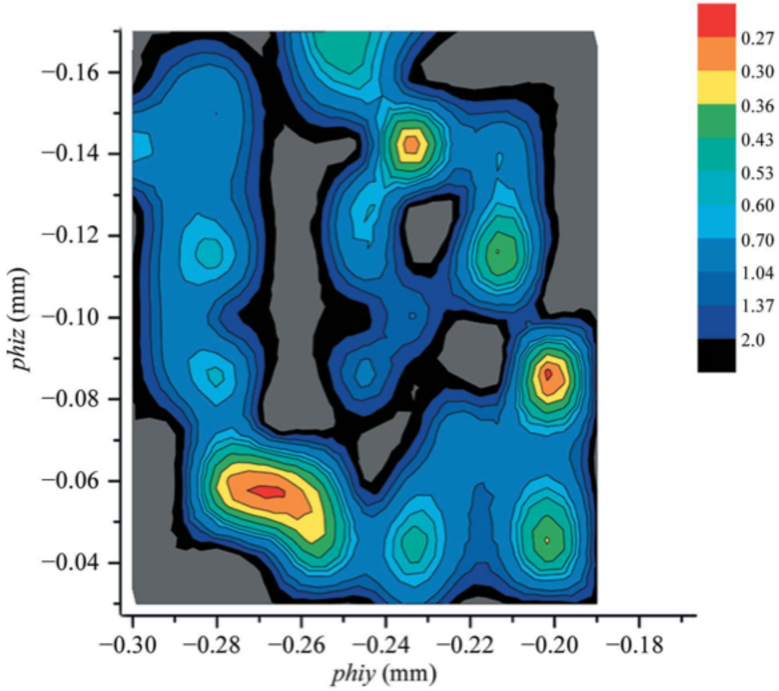
- **id14eh1** Sample Evaluations: 9921, Data Sets: 1376
- **id14eh2** Sample Evaluations: 3799, Data Sets: 356
- **id14eh4** Sample Evaluations: 15137, Data Sets: 3196
- **id29** Sample Evaluations: 15639, Data Sets: 4320
- **id23eh1** Sample Evaluations: 22371, Data Sets: 4528
- **id23eh2** Sample Evaluations: 21885, Data Sets: 2572
- **Total Sample Evaluations**: 88752, **Total Data Sets**: 16348

ISPyB

Fine Grained Sample Evaluation



Beta1-andregenic GPCR



Rastering in 15 micron steps
+ scoring

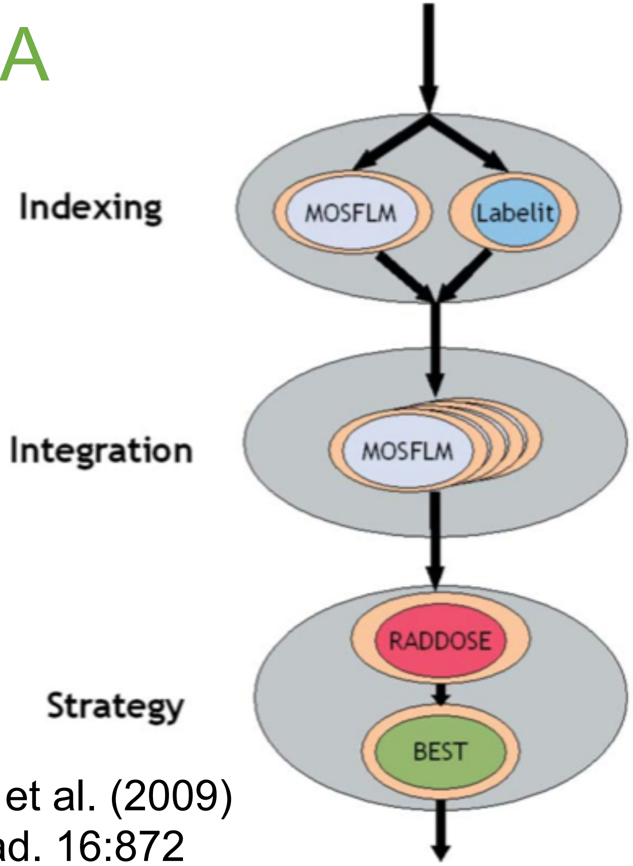
- 'Diffraction Cartograph' Bowler et al. Acta Cryst D (2010) D66:855

Data Collection Strategy

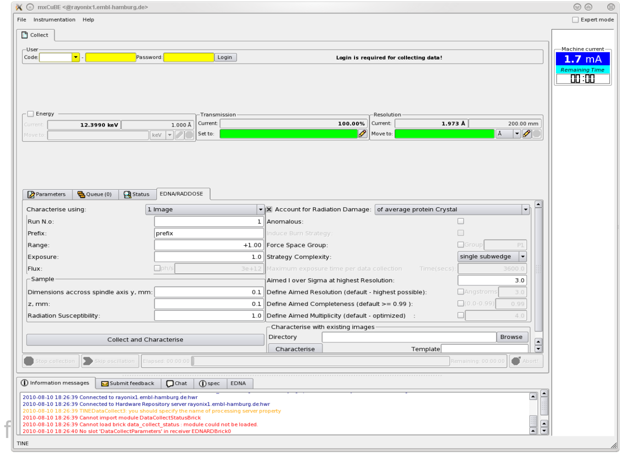
- Results from sample characterization have to be analysed and a strategy has to be created to obtain the best possible data for the purpose (phasing, ligand structures ….) while fighting radiation damage
- BEST (Gleb Bourenkov, EMBL-HH; Sasha Popov, ESRF-GR)
- Conditions for 'clever' data collection
  - Evaluation parameters and resulting data are stored
  - Crystals can be remounted reproducibly (within X micron, within Y degree) -> 'NewPin'
- The 'real' data collection could actually take place on a different beamline / synchrotron

Workflows / Data Models / EDNA

- Framework for implementing pipelines for online-data analysis.
- At the heart is the data model for the specific process. This can be defined e.g. in XML schema definition (XSD).
- Establishing a data model for kappa-goniometry with one flat detector has been a difficult exercise.
- EMBL-HH has written an mxcube-brick for integrating EDNA into the beamline user interface.



Incardorna et al. (2009)
J. Sync. Rad. 16:872



User

Code: - Password:

Login is required for collecting data!

Machine current: 1.7 mA

Remaining Time 00:00

□ Energy

Current: 12.3990 keV 1.000 Å

Move to: keV

Transmission

Current: 100.00%

Set to:

Resolution

Current: 1.973 Å 200.00 mm

Move to: Å

Characterise using: 1 Image

Run N.o: 1

Prefix: prefix

Range: +1.00

Exposure: 1.0

Flux: □ ph/s 3e+12

Sample

Dimensions accross spindle axis y, mm: 0.1

z, mm: 0.1

Radiation Susceptibility: 1.0

☒ Account for Radiation Damage: of average protein Crystal

Anomalous: □

Induce Burn Strategy: □

Force Space Group: □ Group P1

Strategy Complexity: single subwedge

Maximum exposure time per data collection Time(secs): 3600.0

Aimed I over Sigma at highest Resolution: 3.0

Define Aimed Resolution (default - highest possible): □ Angstroms 3.0

Define Aimed Completeness (default >= 0.99): □ (0.0-0.99) 0.99

Define Aimed Multiplicity (default - optimized) : □ 4.0

Characterise with existing images

Directory

Template

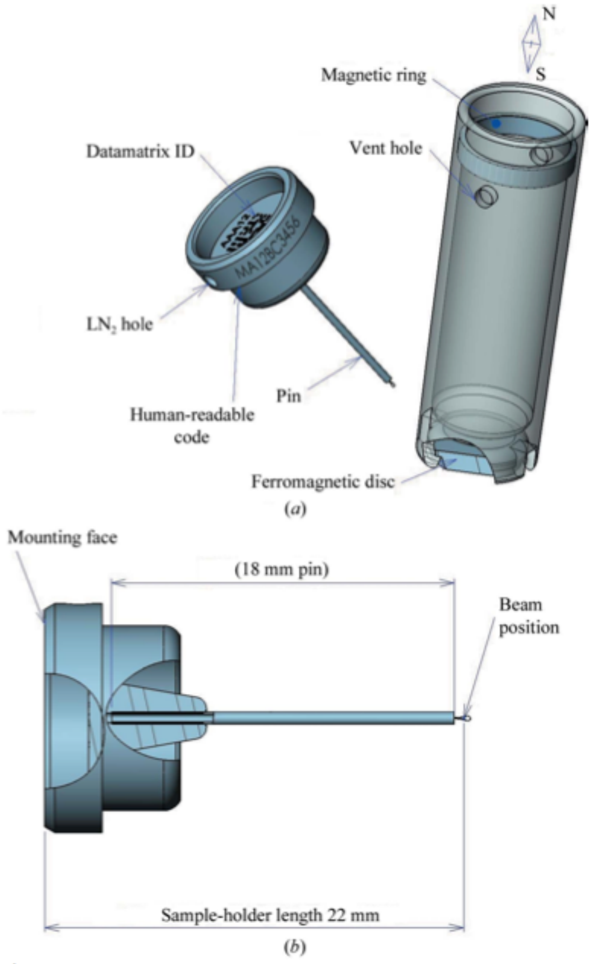
Elapsed: 00:00:00

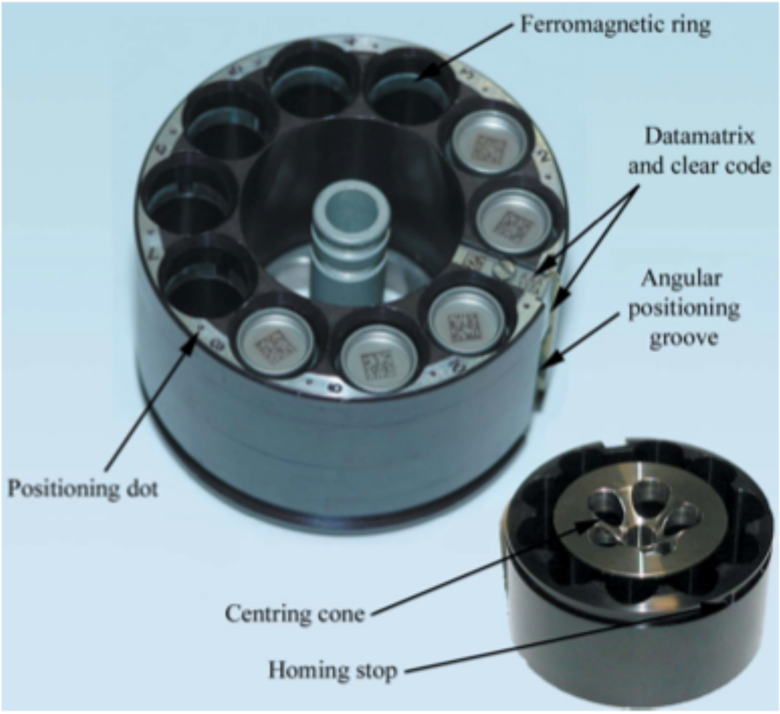
Remaining: 00:00:00

```
2010-08-10 18:26:39 Connected to rayonix1.embl-hamburg.de:hwr
2010-08-10 18:26:39 Connected to Hardware Repository server rayonix1.embl-hamburg.de:hwr
2010-08-10 18:26:39 TINEDataCollect3: you should specify the name of processing server property
2010-08-10 18:26:39 Cannot import module DataCollectStatusBrick
2010-08-10 18:26:39 Cannot load brick data_collect_status : module could not be loaded.
2010-08-10 18:26:40 No slot 'DataCollectParameters' in receiver EDNARDBrick0
```

Sample Management is needed for 100/1000s of samples

The SPINE standard sample holder





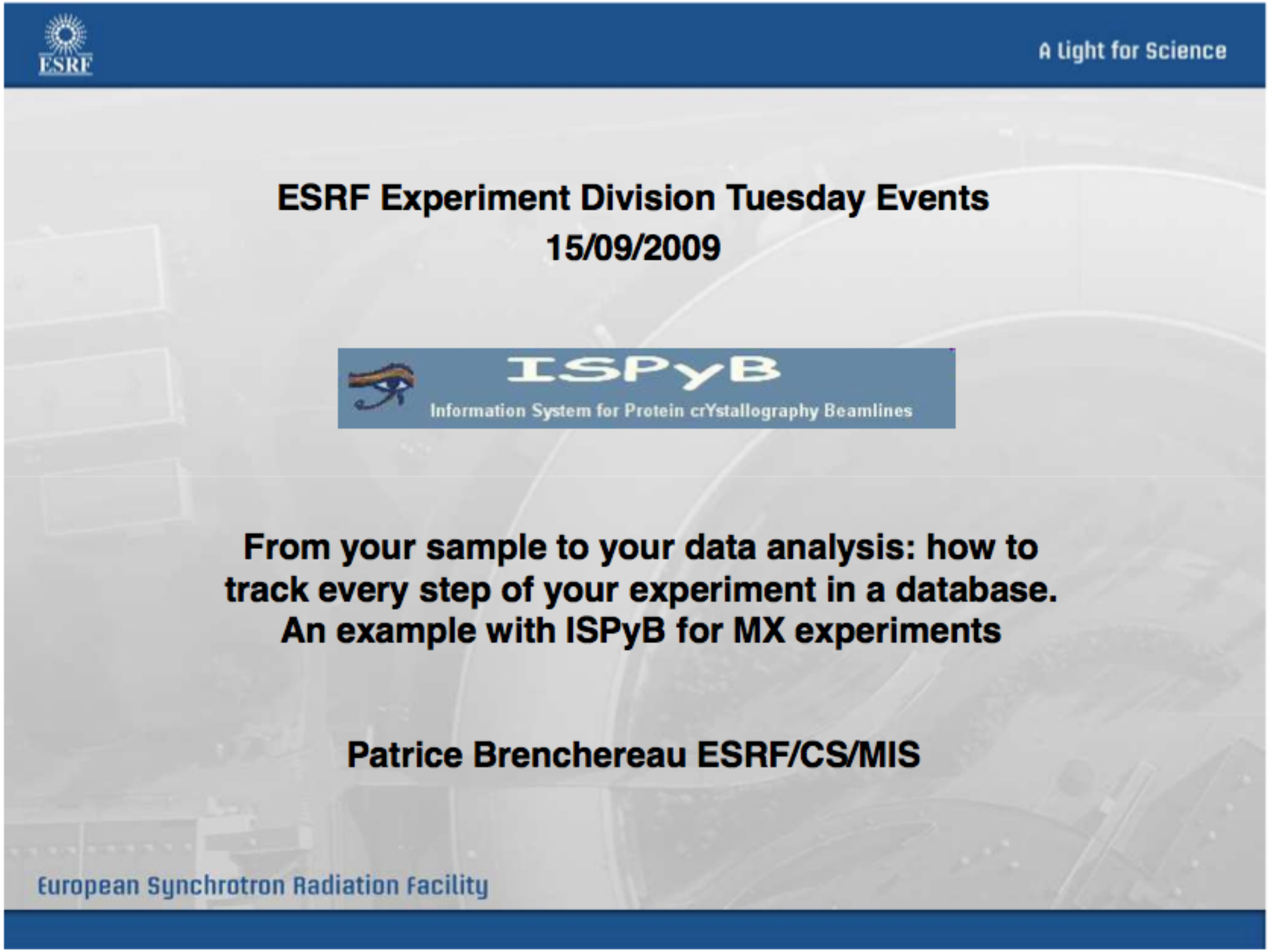
Cipriani et al. (2006)
Acta Cryst. D62:1251

Sample Management

- Items
  - Samples (barcode, protein acronym, crystal form)
  - Containers / baskets (barcode)
  - Dewars (barcode, courier tracking number)
  - Shipments (set of dewars)
  - Pile of shipments (Fri @ ESRF)
- Physical location
  - Where is the crystal?
  - Where is the dewar…

http://www.esrf.eu/UsersAndScience/Experiments/MX/How_to_use_our_beamlines/ISPYB/ISPyB_090915_01%20_00.pdf

iSpyB



ESRF

A Light for Science

ESRF Experiment Division Tuesday Events
15/09/2009

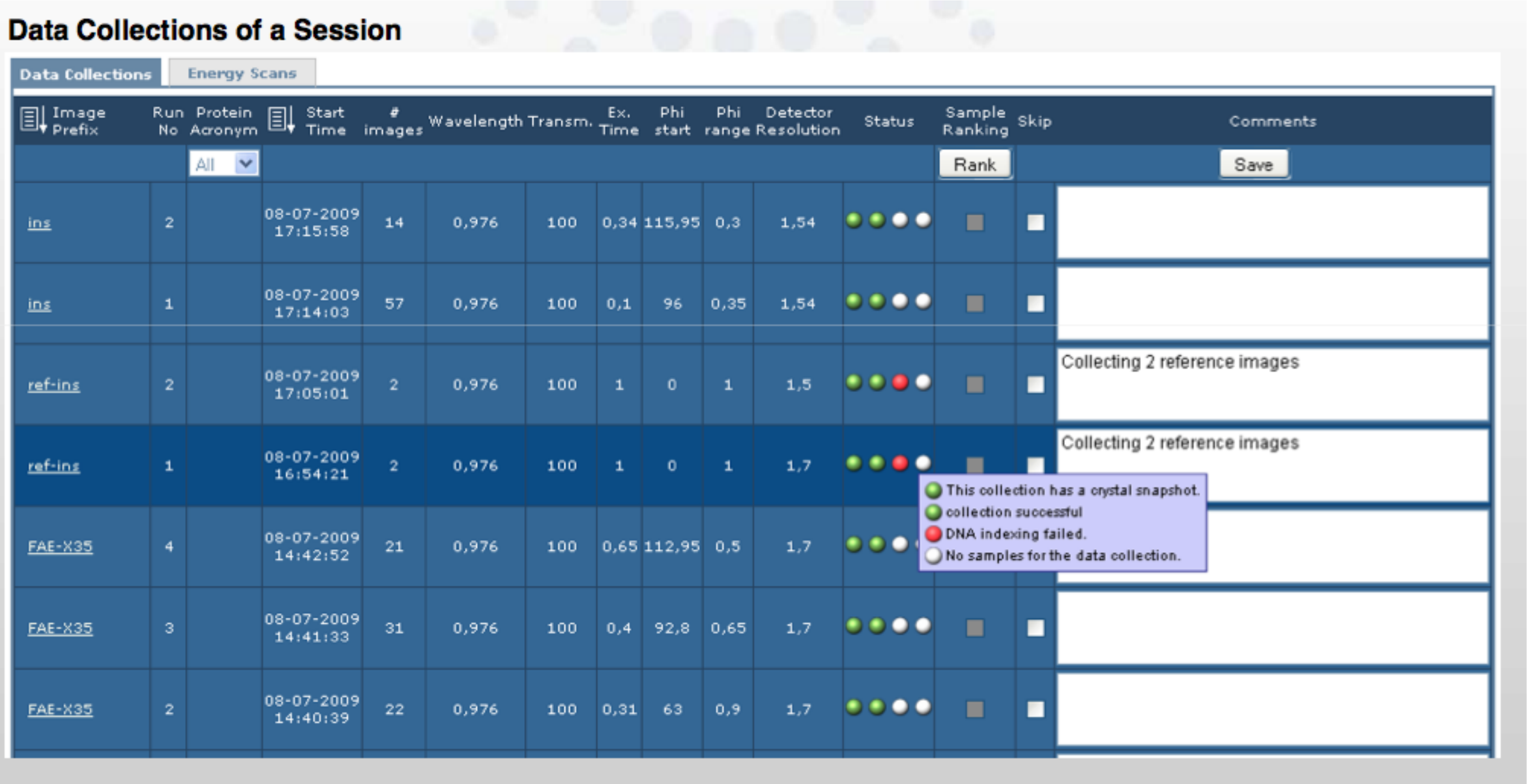
ISPyB
Information System for Protein crYstallography Beamlines

From your sample to your data analysis: how to track every step of your experiment in a database. An example with ISPyB for MX experiments

Patrice Brenchereau ESRF/CS/MIS

European Synchrotron Radiation Facility

iSpyB – Data Collection



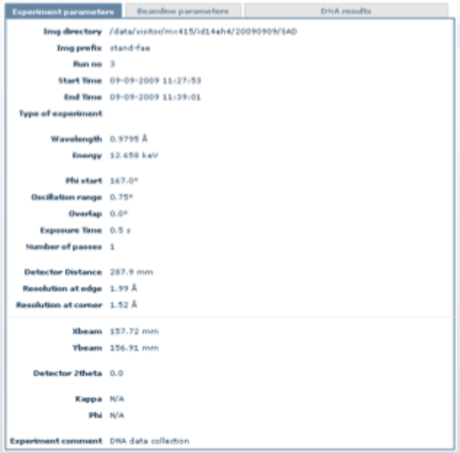
Data Collections of a Session

Image Prefix	Run No	Protein Acronym	Start Time	# images	Wavelength	Transm.	Ex. Time	Phi start	Phi range	Detector Resolution	Status	Sample Ranking	Skip	Comments
		All										Rank		Save
ins	2		08-07-2009 17:15:58	14	0,976	100	0,34	115,95	0,3	1,54	🟢🟢⚪⚪	■	□	
ins	1		08-07-2009 17:14:03	57	0,976	100	0,1	96	0,35	1,54	🟢🟢⚪⚪	■	□	
ref-ins	2		08-07-2009 17:05:01	2	0,976	100	1	0	1	1,5	🟢🟢🔴⚪	■	□	Collecting 2 reference images
ref-ins	1		08-07-2009 16:54:21	2	0,976	100	1	0	1	1,7	🟢🟢🔴⚪	■	□	Collecting 2 reference images
FAE-X35	4		08-07-2009 14:42:52	21	0,976	100	0,65	112,95	0,5	1,7	🟢🟢⚪			
FAE-X35	3		08-07-2009 14:41:33	31	0,976	100	0,4	92,8	0,65	1,7	🟢🟢⚪⚪	■	□	
FAE-X35	2		08-07-2009 14:40:39	22	0,976	100	0,31	63	0,9	1,7	🟢🟢⚪⚪	■	□	

🟢 This collection has a crystal snapshot.
🟢 collection successful
🔴 DNA indexing failed.
⚪ No samples for the data collection.

iSpyB – Data Collection

Parameters & Results



Img directory	/data/visitor/mx415/id14eh4/20090909/SAD
Img prefix	stand-fae
Run no	3
Start Time	09-09-2009 11:27:53
End Time	09-09-2009 11:39:01
Type of experiment	
Wavelength	0.9795 Å
Energy	12.658 keV
Phi start	167.0°
Oscillation range	0.75°
Overlap	0.0°
Exposure Time	0.5 s
Number of passes	1
Detector Distance	287.9 mm
Resolution at edge	1.99 Å
Resolution at corner	1.52 Å
Xbeam	157.72 mm
Ybeam	156.91 mm
Detector 2theta	0.0
Kappa	N/A
Phi	N/A
Experiment comment	DNA data collection

Crystal Snapshots

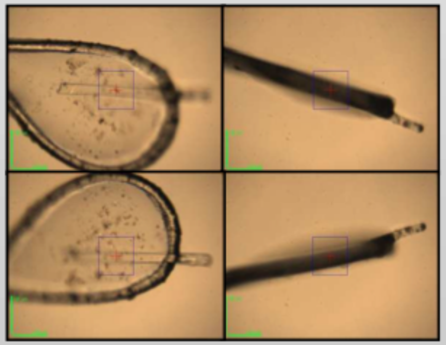
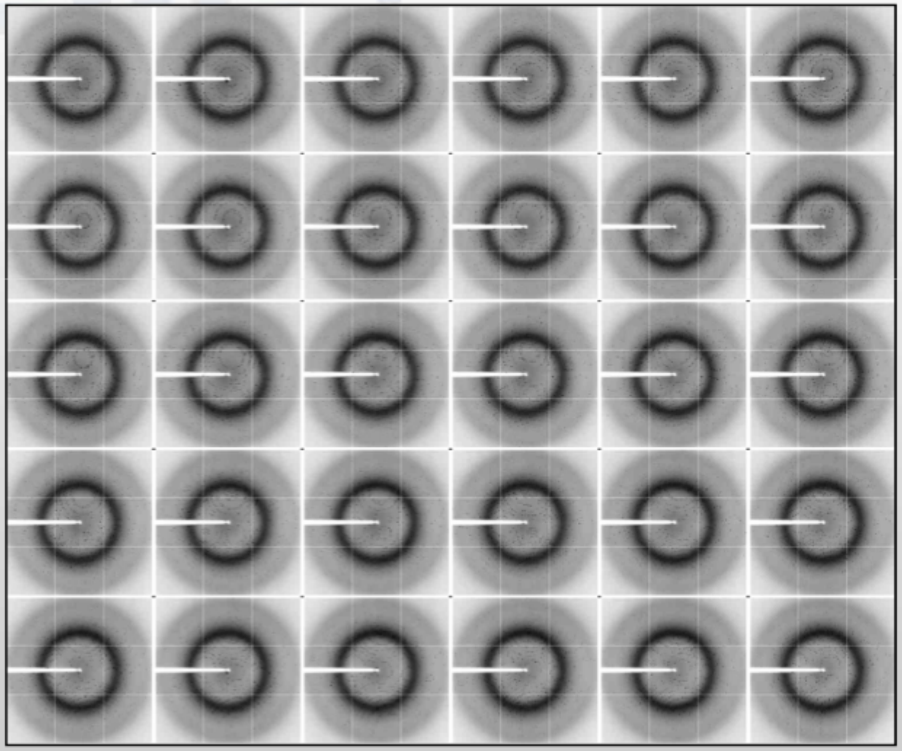
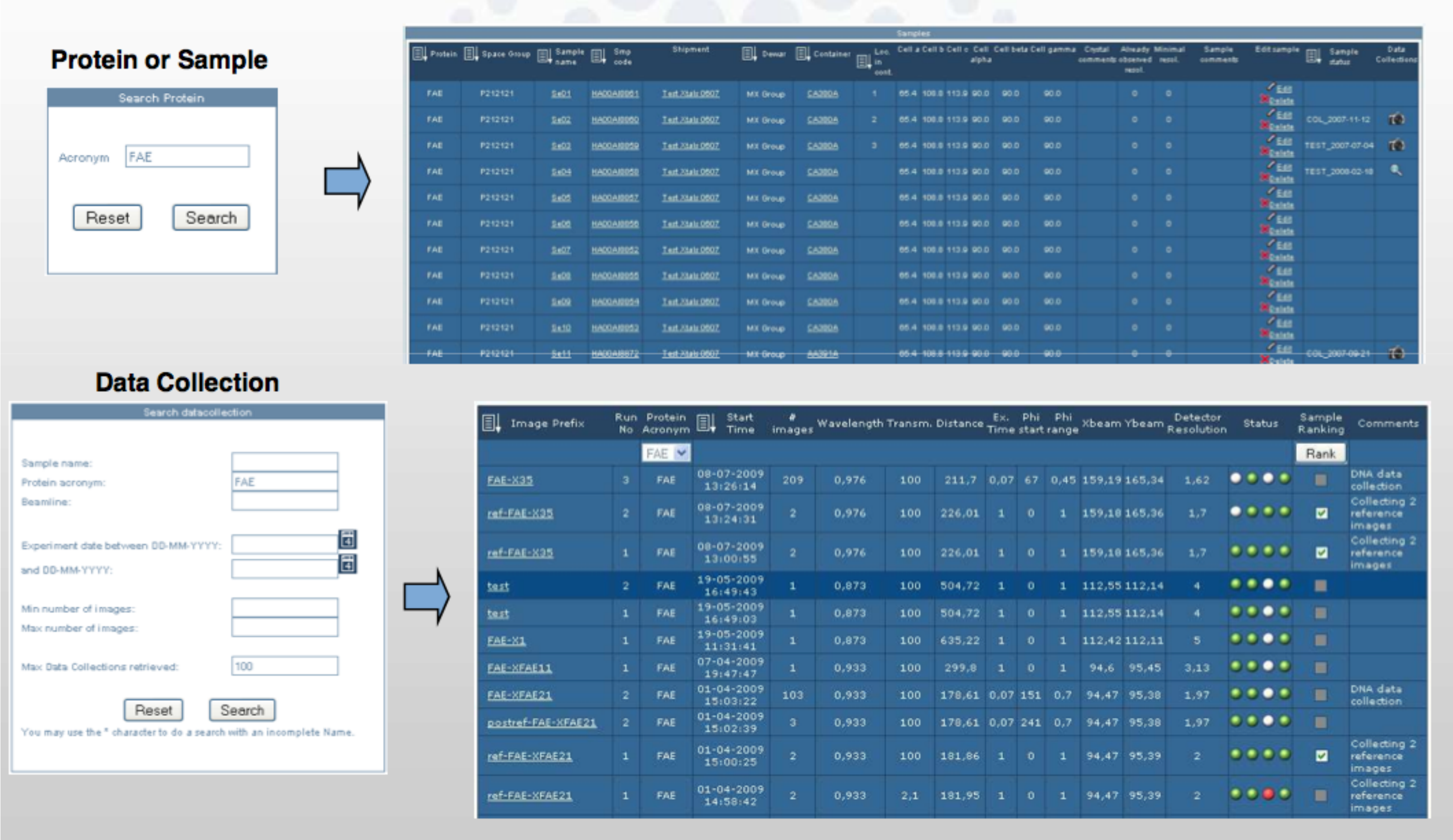


Image thumbnails (image wall)



iSpyB – Data Mining



Protein or Sample

Search Protein

Acronym: FAE

Reset | Search

Samples

Protein	Space Group	Sample name	Smp code	Shipment	Dewar	Container	Loc. in cont.	Cell a	Cell b	Cell c	Cell alpha	Cell beta	Cell gamma	Crystal comments	Already observed resol.	Minimal resol.	Sample comments	Edit sample	Sample status	Data Collections
FAE	P212121	Se01	HA00AI8061	Test Xtals 0607	MX Group	CA380A	1	65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete		
FAE	P212121	Se02	HA00AI8060	Test Xtals 0607	MX Group	CA380A	2	65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete	COL_2007-11-12	
FAE	P212121	Se03	HA00AI8059	Test Xtals 0607	MX Group	CA380A	3	65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete	TEST_2007-07-04	
FAE	P212121	Se04	HA00AI8058	Test Xtals 0607	MX Group	CA380A		65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete	TEST_2008-02-18	
FAE	P212121	Se05	HA00AI8057	Test Xtals 0607	MX Group	CA380A		65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete		
FAE	P212121	Se06	HA00AI8056	Test Xtals 0607	MX Group	CA380A		65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete		
FAE	P212121	Se07	HA00AI8052	Test Xtals 0607	MX Group	CA380A		65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete		
FAE	P212121	Se08	HA00AI8055	Test Xtals 0607	MX Group	CA380A		65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete		
FAE	P212121	Se09	HA00AI8054	Test Xtals 0607	MX Group	CA380A		65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete		
FAE	P212121	Se10	HA00AI8053	Test Xtals 0607	MX Group	CA380A		65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete		
FAE	P212121	Se11	HA00AI8072	Test Xtals 0607	MX Group	AA391A		65.4	108.8	113.9	90.0	90.0	90.0		0	0		Edit Delete	COL_2007-09-21	

Data Collection

Search datacollection

Sample name:

Protein acronym: FAE

Beamline:

Experiment date between DD-MM-YYYY:

and DD-MM-YYYY:

Min number of images:

Max number of images:

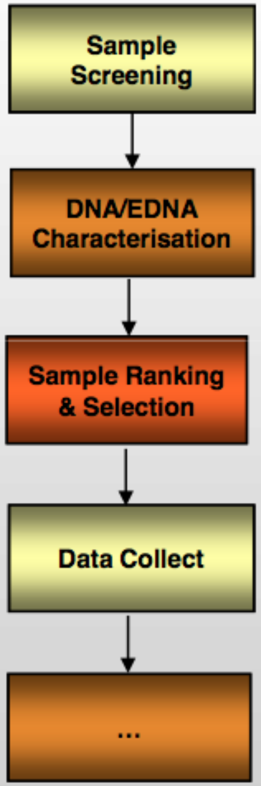
Max Data Collections retrieved: 100

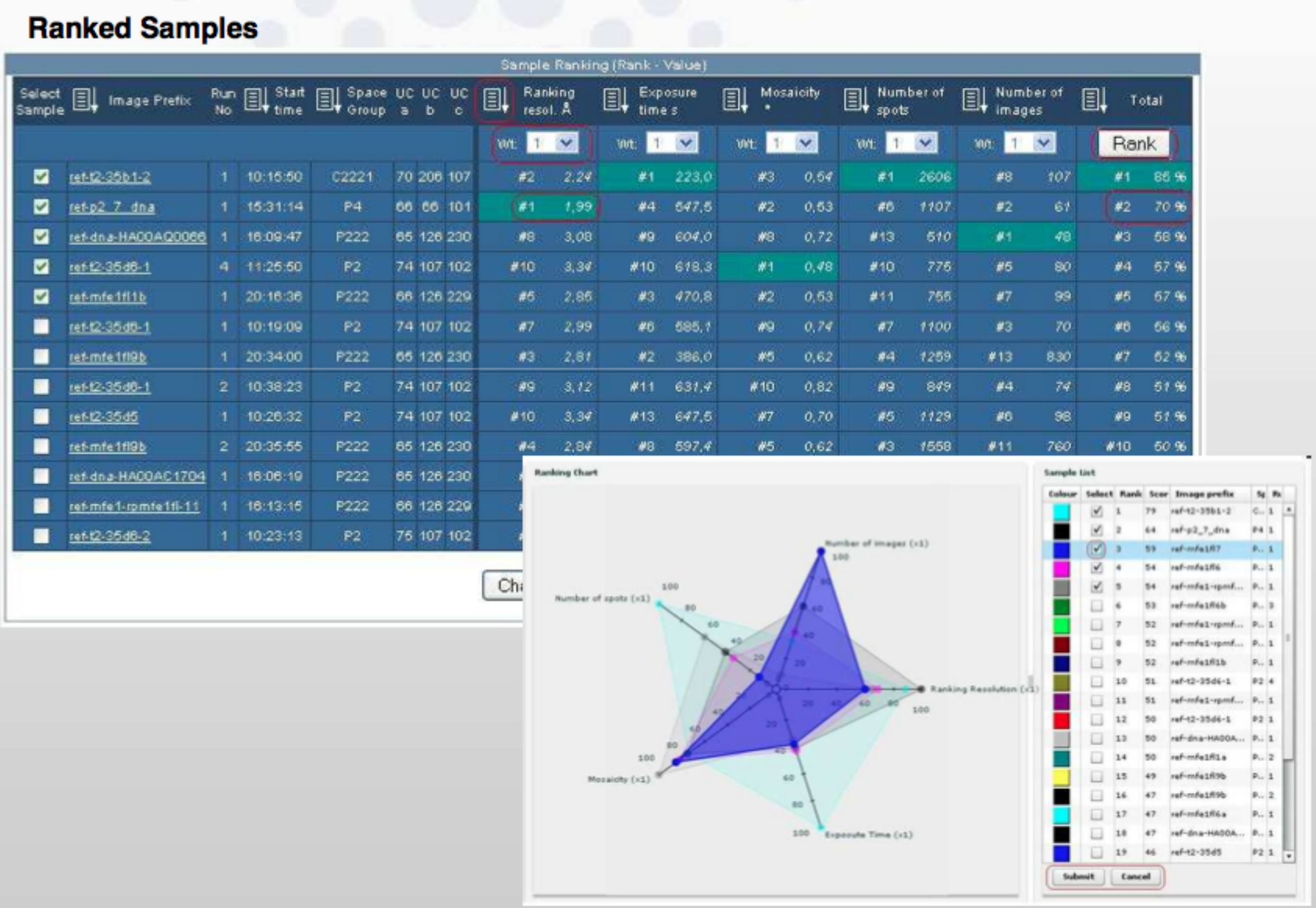
Reset | Search

You may use the * character to do a search with an incomplete Name.

Image Prefix	Run No	Protein Acronym	Start Time	# images	Wavelength	Transm.	Distance	Ex. Time	Phi start	Phi range	Xbeam	Ybeam	Detector Resolution	Status	Sample Ranking	Comments
		FAE													Rank	
FAE-X35	3	FAE	08-07-2009 13:26:14	209	0,976	100	211,7	0,07	67	0,45	159,19	165,34	1,62	○●○●	□	DNA data collection
ref-FAE-X35	2	FAE	08-07-2009 13:24:31	2	0,976	100	226,01	1	0	1	159,18	165,36	1,7	○●●●	☑	Collecting 2 reference images
ref-FAE-X35	1	FAE	08-07-2009 13:00:55	2	0,976	100	226,01	1	0	1	159,18	165,36	1,7	●●●●	☑	Collecting 2 reference images
test	2	FAE	19-05-2009 16:49:43	1	0,873	100	504,72	1	0	1	112,55	112,14	4	●●○●	□	
test	1	FAE	19-05-2009 16:49:03	1	0,873	100	504,72	1	0	1	112,55	112,14	4	●●○●	□	
FAE-X1	1	FAE	19-05-2009 11:31:41	1	0,873	100	635,22	1	0	1	112,42	112,11	5	●●○●	□	
FAE-XFAE11	1	FAE	07-04-2009 19:47:47	1	0,933	100	299,8	1	0	1	94,6	95,45	3,13	●●○●	□	
FAE-XFAE21	2	FAE	01-04-2009 15:03:22	103	0,933	100	178,61	0,07	151	0,7	94,47	95,38	1,97	●●○●	□	DNA data collection
postref-FAE-XFAE21	2	FAE	01-04-2009 15:02:39	3	0,933	100	178,61	0,07	241	0,7	94,47	95,38	1,97	●●○●	□	
ref-FAE-XFAE21	1	FAE	01-04-2009 15:00:25	2	0,933	100	181,86	1	0	1	94,47	95,39	2	●●●●	☑	Collecting 2 reference images
ref-FAE-XFAE21	1	FAE	01-04-2009 14:58:42	2	0,933	2,1	181,95	1	0	1	94,47	95,39	2	●●🔴●	□	Collecting 2 reference images

iSpyB – Sample Ranking





Ranked Samples

Sample Ranking (Rank - Value)

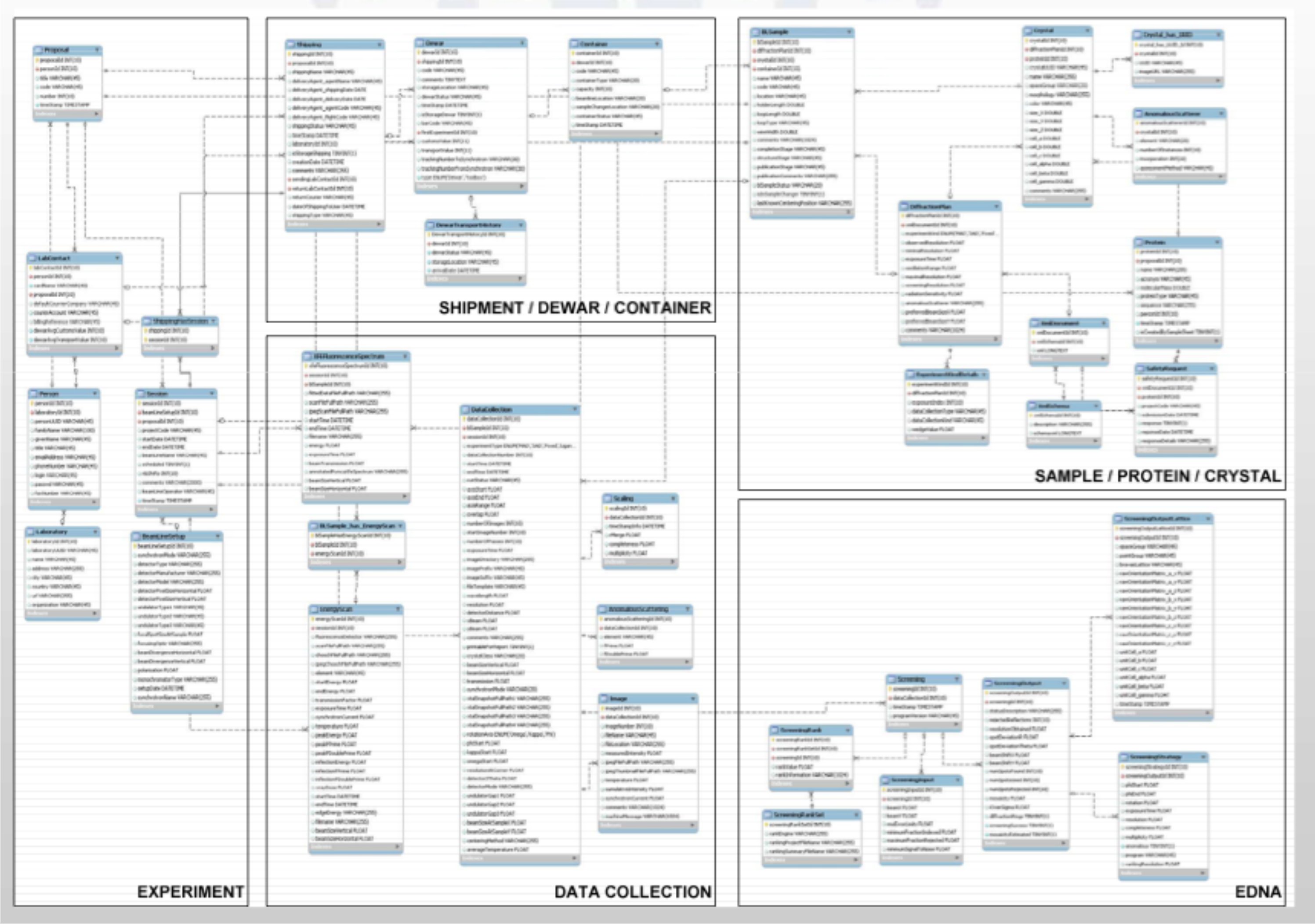
Select Sample	Image Prefix	Run No	Start time	Space Group	UC a	UC b	UC c	Ranking resol. Å	Exposure time s	Mosaicity °	Number of spots	Number of images	Total
								Wt: 1	Wt: 1	Wt: 1	Wt: 1	Wt: 1	Rank
☑	ref-t2-35b1-2	1	10:15:50	C2221	70	206	107	#2 2,24	#1 223,0	#3 0,54	#1 2606	#8 107	#1 85 %
☑	ref-p2_7_dna	1	15:31:14	P4	66	66	101	#1 1,99	#4 547,5	#2 0,53	#6 1107	#2 61	#2 70 %
☑	ref-dna-HA00AQ0066	1	16:09:47	P222	65	126	230	#8 3,08	#9 604,0	#8 0,72	#13 510	#1 48	#3 58 %
☑	ref-t2-35d6-1	4	11:25:50	P2	74	107	102	#10 3,34	#10 618,3	#1 0,48	#10 775	#5 80	#4 57 %
☑	ref-mfe1fl1b	1	20:16:36	P222	66	126	229	#5 2,85	#3 470,8	#2 0,53	#11 755	#7 99	#5 57 %
□	ref-t2-35d6-1	1	10:19:09	P2	74	107	102	#7 2,99	#6 585,1	#9 0,74	#7 1100	#3 70	#6 56 %
□	ref-mfe1fl9b	1	20:34:00	P222	65	126	230	#3 2,81	#2 386,0	#5 0,62	#4 1259	#13 830	#7 52 %
□	ref-t2-35d6-1	2	10:38:23	P2	74	107	102	#9 3,12	#11 631,4	#10 0,82	#9 849	#4 74	#8 51 %
□	ref-t2-35d5	1	10:26:32	P2	74	107	102	#10 3,34	#13 647,5	#7 0,70	#5 1129	#6 98	#9 51 %
□	ref-mfe1fl9b	2	20:35:55	P222	65	126	230	#4 2,84	#8 597,4	#5 0,62	#3 1558	#11 760	#10 50 %
□	ref-dna-HA00AC1704	1	16:06:19	P222	65	126	230						
□	ref-mfe1-rpmfe1fl-11	1	16:13:15	P222	66	126	229						
□	ref-t2-35d6-2	1	10:23:13	P2	75	107	102						

Ranking Chart

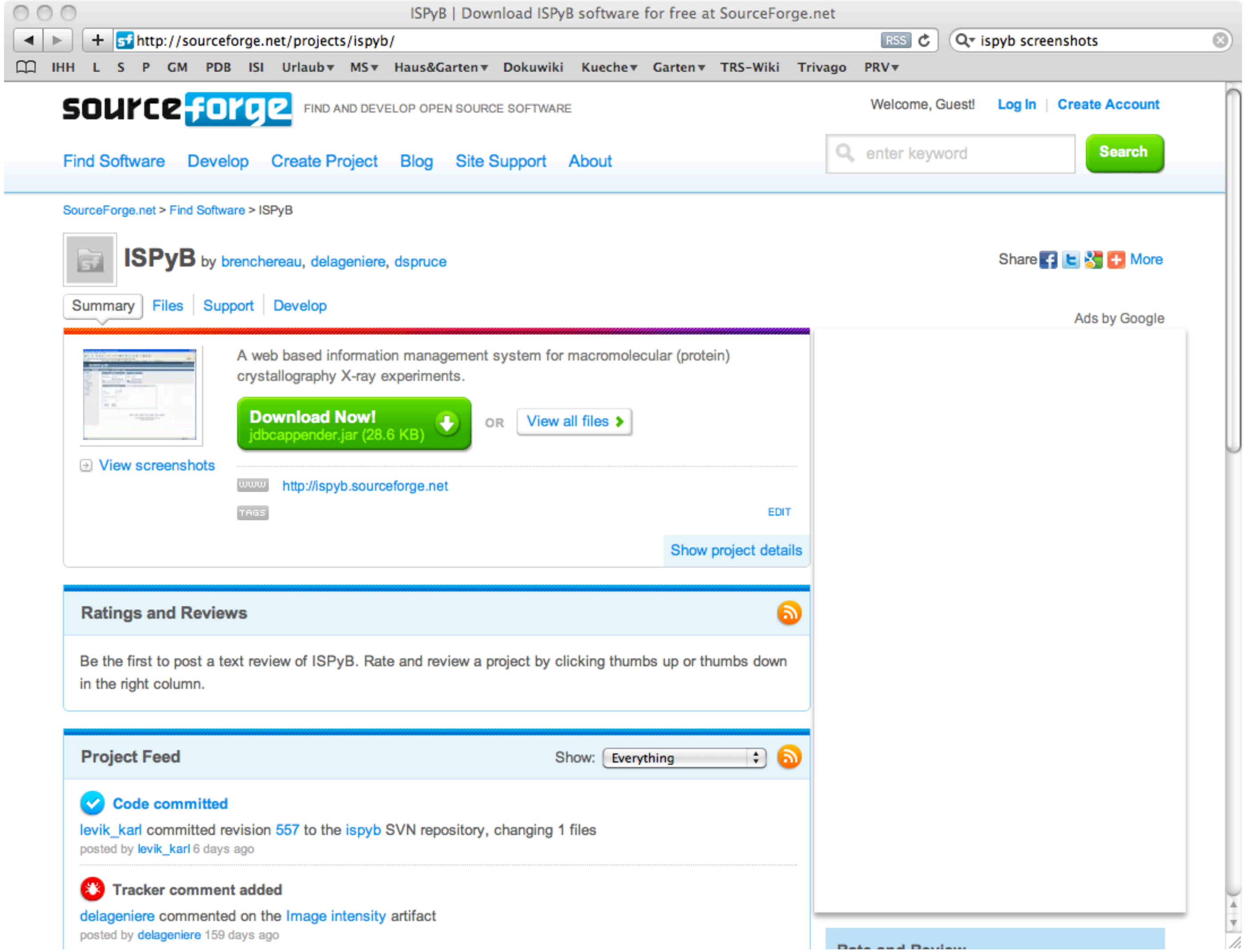
Sample List

Colour	Select	Rank	Scor	Image prefix	Sg	Rn
🟦	☑	1	79	ref-t2-35b1-2	C..	1
⬛	☑	2	64	ref-p2_7_dna	P4	1
🟦	☑	3	59	ref-mfe1fl7	P..	1
🟪	☑	4	54	ref-mfe1fl6	P..	1
⬜	☑	5	54	ref-mfe1-rpmf...	P..	1
🟩	□	6	53	ref-mfe1fl6b	P..	3
🟩	□	7	52	ref-mfe1-rpmf...	P..	1
🟥	□	8	52	ref-mfe1-rpmf...	P..	1
🟦	□	9	52	ref-mfe1fl1b	P..	1
🟨	□	10	51	ref-t2-35d6-1	P2	4
🟪	□	11	51	ref-mfe1-rpmf...	P..	1
🟥	□	12	50	ref-t2-35d6-1	P2	1
⬜	□	13	50	ref-dna-HA00A...	P..	1
🟦	□	14	50	ref-mfe1fl1a	P..	2
🟨	□	15	49	ref-mfe1fl9b	P..	1
⬛	□	16	47	ref-mfe1fl9b	P..	2
🟦	□	17	47	ref-mfe1fl6a	P..	1
⬛	□	18	47	ref-dna-HA00A...	P..	1
🟦	□	19	46	ref-t2-35d5	P2	1

iSpyB – Data Model



iSpyB is open source



ISPyB | Download ISPyB software for free at SourceForge.net

http://sourceforge.net/projects/ispyb/

ispyb screenshots

IHH L S P GM PDB ISI Urlaub MS Haus&Garten Dokuwiki Kueche Garten TRS-Wiki Trivago PRV
sourceforge FIND AND DEVELOP OPEN SOURCE SOFTWARE
Welcome, Guest! Log In | Create Account
enter keyword
Search
Find Software Develop Create Project Blog Site Support About
SourceForge.net > Find Software > ISPyB
Share More
Summary Files Support Develop
Ads by Google

ISPyB by brenchereau, delageniere, dspruce

A web based information management system for macromolecular (protein) crystallography X-ray experiments.

Download Now! jdbcappender.jar (28.6 KB) OR View all files

View screenshots

www http://ispyb.sourceforge.net

TAGS EDIT

Show project details

Ratings and Reviews

Be the first to post a text review of ISPyB. Rate and review a project by clicking thumbs up or thumbs down in the right column.

Project Feed Show: Everything

Code committed

levik_karl committed revision 557 to the ispyb SVN repository, changing 1 files

posted by levik_karl 6 days ago

Tracker comment added

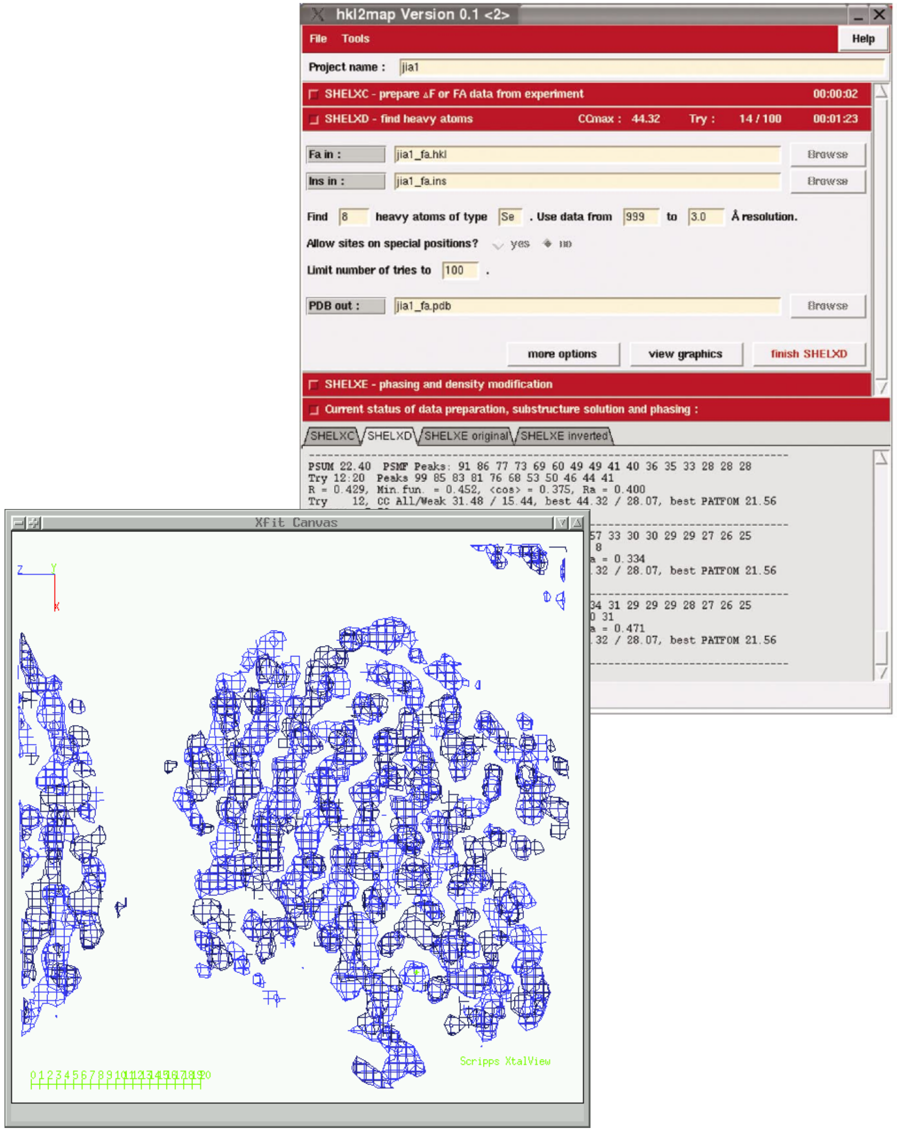
delageniere commented on the Image intensity artifact

posted by delageniere 159 days ago

EMBL

Data Evaluation

- 'Standard cases'
  - HKL2MAP will produce an electron density map in 3-5 min.
  - AutoRickshaw will produce a 3D-model in some hours
- Difficult cases
  - Many dead ends
- Inclusion of HKL2MAP results into iSpyB has been attempted.



Project name : jia1

□ SHELXC - prepare ΔF or FA data from experiment 00:00:02

□ SHELXD - find heavy atoms CCmax : 44.32 Try : 14 / 100 00:01:23

Fa in : jia1_fa.hkl

Ins in : jia1_fa.ins

Find 8 heavy atoms of type Se . Use data from 999 to 3.0 Å resolution.

Allow sites on special positions? ○ yes ● no

Limit number of tries to 100 .

PDB out : jia1_fa.pdb

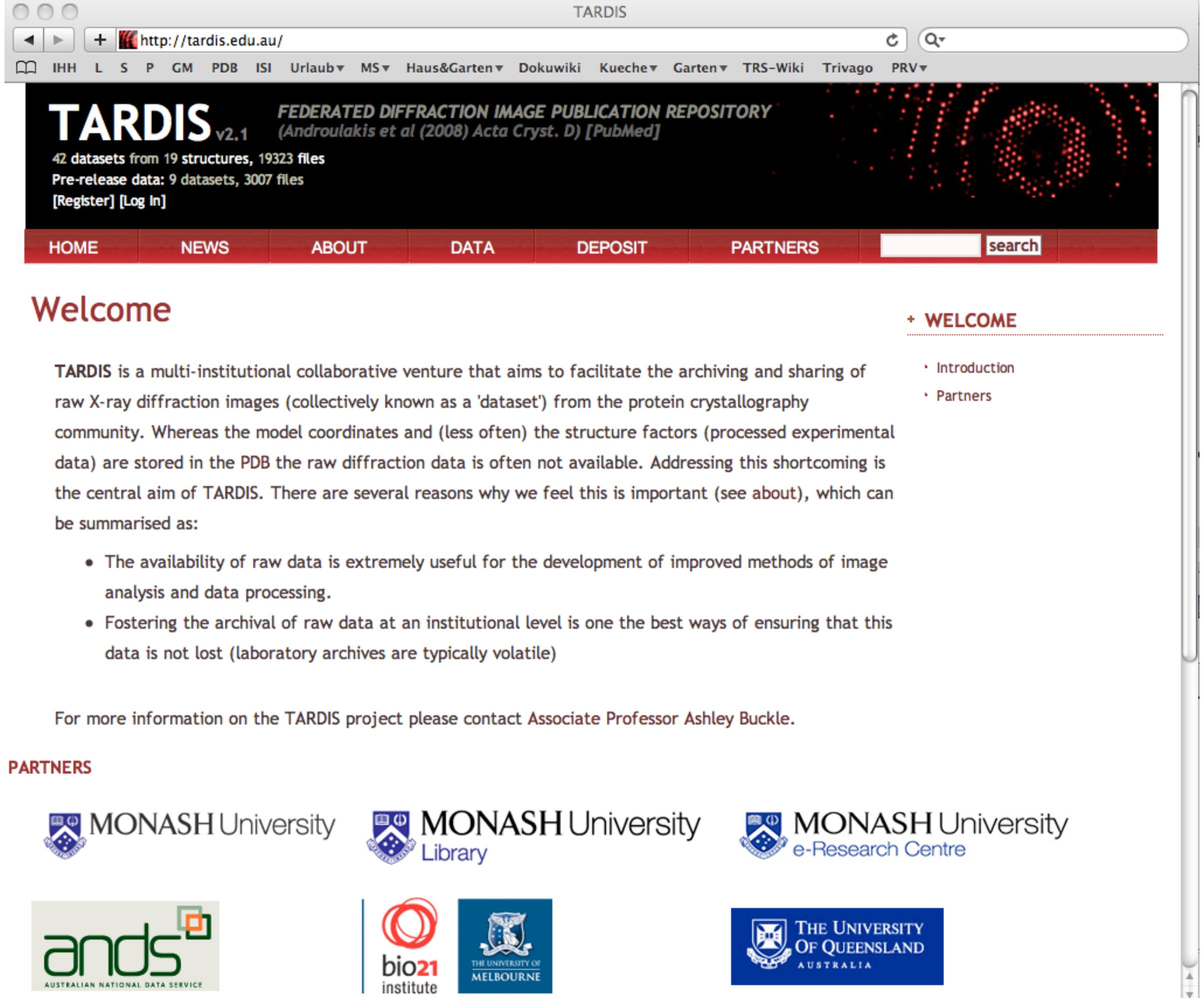
□ SHELXE - phasing and density modification

□ Current status of data preparation, substructure solution and phasing :

```
------------------------------------------------------------------------
PSUM 22.40  PSMF Peaks: 91 86 77 73 69 60 49 49 41 40 36 35 33 28 28 28
Try 12:20  Peaks 99 85 83 81 76 68 53 50 46 44 41
R = 0.429, Min.fun. = 0.452, <cos> = 0.375, Ra = 0.400
Try    12, CC All/Weak 31.48 / 15.44, best 44.32 / 28.07, best PATFOM 21.56
                              ------------------------------------------
                              57 33 30 30 29 29 27 26 25
                               8
                              a = 0.334
                              .32 / 28.07, best PATFOM 21.56
                              ------------------------------------------
                              34 31 29 29 29 28 27 26 25
                              0 31
                              a = 0.471
                              .32 / 28.07, best PATFOM 21.56
                              ------------------------------------------
```

TARDIS

- http://tardis.edu.au
- http://www.monash.edu.au/news/newsline/story/1608, http://mpegmedia.abc.net.au/rn/podcast/2010/07/fte_20100715_0850.mp3



TARDIS
http://tardis.edu.au/
IHH L S P GM PDB ISI Urlaub MS Haus&Garten Dokuwiki Kueche Garten TRS-Wiki Trivago PRV
TARDIS v2.1
FEDERATED DIFFRACTION IMAGE PUBLICATION REPOSITORY
(Androulakis et al (2008) Acta Cryst. D) [PubMed]
42 datasets from 19 structures, 19323 files
Pre-release data: 9 datasets, 3007 files
[Register] [Log In]
HOME
NEWS
ABOUT
DATA
DEPOSIT
PARTNERS
search
WELCOME
Introduction
Partners

Welcome

TARDIS is a multi-institutional collaborative venture that aims to facilitate the archiving and sharing of raw X-ray diffraction images (collectively known as a 'dataset') from the protein crystallography community. Whereas the model coordinates and (less often) the structure factors (processed experimental data) are stored in the PDB the raw diffraction data is often not available. Addressing this shortcoming is the central aim of TARDIS. There are several reasons why we feel this is important (see about), which can be summarised as:

- The availability of raw data is extremely useful for the development of improved methods of image analysis and data processing.
- Fostering the archival of raw data at an institutional level is one the best ways of ensuring that this data is not lost (laboratory archives are typically volatile)

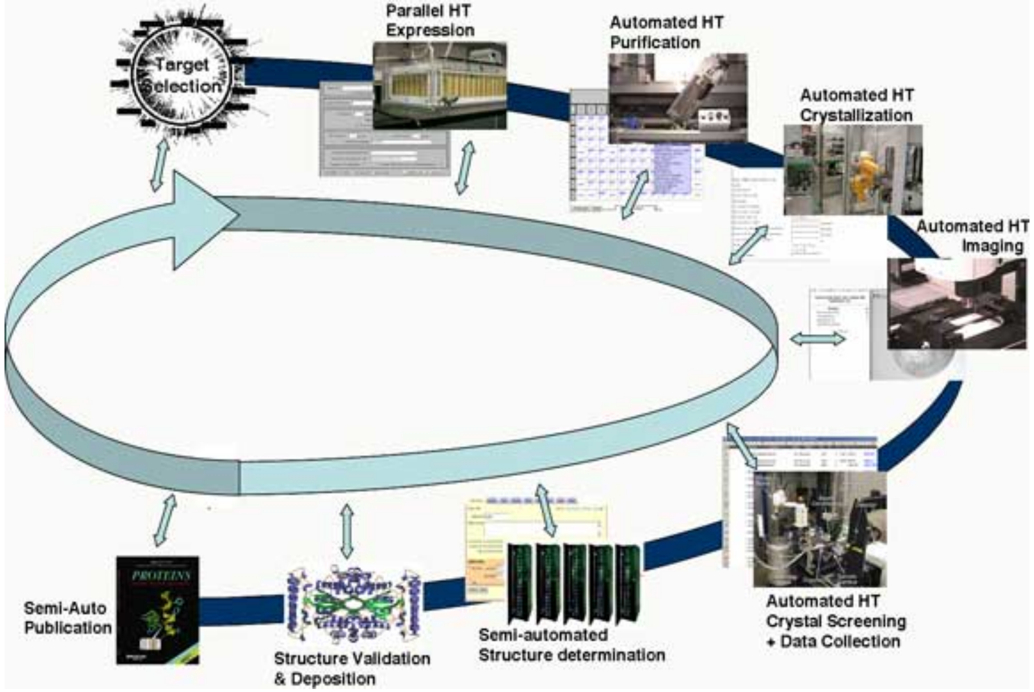
For more information on the TARDIS project please contact Associate Professor Ashley Buckle.

PARTNERS

MONASH University
MONASH University Library
MONASH University e-Research Centre
ands AUSTRALIAN NATIONAL DATA SERVICE
bio21 institute
THE UNIVERSITY OF MELBOURNE
THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Structural Genomics projects

- Go from gene to structure
- Well-defined workflows including experiments at synchrotrons
- Examples:
  - Joint Center for Structural Genomics www.jcsg.org
  - Structural Genomics Consortium www.thesgc.org


Target Selection
Parallel HT Expression
Automated HT Purification
Automated HT Crystallization
Automated HT Imaging
Automated HT Crystal Screening + Data Collection
Semi-automated Structure determination
Structure Validation & Deposition
Semi-Auto Publication
PROTEINS

http://www.nigms.nih.gov/Initiatives/PSI/Centers/JCSG.htm

Long-Term Archiving of raw data

- Reasons
  - Potential value for re-evaluation with new technology / fresh brain
  - 10 years obligatory documentation
  - Traceability in case of possible fraud
- Who owns the raw data / meta data?
  - The PI? For how long?
  - The synchrotron?
  - The public?
- Role management
  - Assignment of roles to users
  - Identification of users and their roles
  - mobility between groups/institutions
  - orphaned data (retirement etc.)

- dCache@DESY plus an interface (Frank Schluenzen, Ilya Agapov)

Some additional points

- 'Religious' approach to work flows (different labs have different religions) -> flexibility / intuitivity
- Confidentiality / Visibility / Practicality issues with centralized servers.
- Practical work with users:
  - Much of the practical work is done by inexperienced PhD students
  - More experienced scientists only sporadically collect data (at different sources)
  - If meta-data collection is not done automatically, it will not be done (in academia at least)
  - Data Management should be totally transparent (drag-and-drop …)
- Licensing issues with data handling systems (GPL, LGPL, open source, unknown, …)

EMBL@PETRA3

- **Goal**: 3D structural information about biological macromolecules and their complexes at the highest possible resolution (see ribosome).
- **'Integrated Facility for Structural Biology'** will offer user access to:
  - 1 SAXS (small angle x-ray scattering) beamline at PETRA III
  - 2 MX (macromolecular crystallography) beamlines at PETRA III
  - High-Throughput Crystallization Facility (1 Mio experiments, largest in Europe)
  - Sample preparation and characterization
  - (Remote) Data Evaluation
    - arpWarp, ATSAS, autoRickshaw, …
- Remote Access where requested and possible
- Industrie is welcome – confidentiality is important

MX – Different data at different steps

- Protein Production -> 'PIMS' et al. (>2nd inc, >10y)
- Protein Characterization -> 'PIMS' et al.
- SAXS experiment -> 'iSpyB' (nascent)
- Crystallization -> 'Crims' et al. (>5y)
- Crystallographic Testing and Data Collection -> 'ispyB' (2nd inc, >5y)
- Data Processing -> 'ispyB'
- Data Evaluation -> 'ispyB'
- Results -> Protein Data Bank (+20y)
- Long time Archive -> ???

Summary

- Synchrotron-based crystallography is an advanced technique with established work-flows.
- Despite this, in most places in Europe (except ESRF and DIAMOND) data management is still done by transporting hard disks.
- This system will not be viable in the future:
  - Robotics will further increase through-put
  - Fast detectors will increase through-put and data volume
  - Requirements for documentation will be enforced
- Data from related experiments / methods (also wet-lab) need to be stored and managed ideally in an integrated manner. Consistent meta data / connectivity between data are very important.
- Resources needed for design and implementation are substantial
- **Working together with other facilities is important both from the user side (homogeneity) and the supplier side (synergies).**