DNA Full meeting
held at the ESRF on November 4th 2003

Participants: Andrew Leslie[1], Harry Powell[1], Colin Nave[2], Graeme Winter[2], Andrew Thompson[3], Eric Girard[3], Venkataraman Parthasarathy[4], Elspeth Gordon[5], Sean McSweeney[5], Darren Spruce[5], Olof Svensson[5], Didier Nurizzo[5], Vicente Rey-Bakaikoa[5], Jens Meyer[5], Pierre Legrand[6], Florent Cipriani[6], Ludovic Launer[7]

1) Medical Research Council, Cambridge, UK
2) Daresbury Laboratory, UK
3) Synchrotron Soleil, Gif-sur-Yvette, France
4) EMBL Hamburg, Germany
5) ESRF, Grenoble, France
6) EMBL Grenoble, France
7) BM 14 CRG at the ESRF, MRC France

These were the items on the agenda of this meeting:

- Report from developers' meetings held since the last full meeting (Olof)
- Status of project and demo of current version (Graeme)
- Online integration, status and perspectives – MOSFLM (Graeme) and XDS (Pierre)
- Setting up a database or collection of reference images (Harry)
- Causal graphs – how to represent knowledge (Graeme)
- Sample changers (Sean)
- Scoring schemes for multiple samples (Andrew L)
- Automation – what information will come with a Fedexed sample (Ludovic)
- LIMS (Darren)
- Relationship with other projects (Colin)
- Management of the project (Sean)
- Next meeting

The most important item of the agenda was "Management of the project", therefore I'll start these minutes with this item and come back to the other items later.

1. Management of the project

1.1 DNA project specification - DNA 1.0: A system that can collect "good data" automatically.

We defined at this point both a DNA project mission and a task list with priorities. Here is the project specification for release 1.0:

The system should be reliable, robust and be provided with extensive documentation.

More precisely, the system should:

- work with single wavelength data (MAD will be considered for DNA 2)
- be able to do screening and ranking automatically (hence it should work with sample changers)
- automatically determine the optimum exposure time (BEST)
- be able to do space group determination
- integrate and scale collected data
- run offline
- run on any site (however the distribution version will run only on Linux computers)

The input to the system should be both manual and via a beamline LIMS (the eHTPX data model should be used). The output of the system should be a scaled data set and a summary (certification) of the experiment.

1.2 List of priorities

We decided that DNA 1.0 should be available by the next full meeting which is planned to take place in Cambridge in June 2004 (see Section 11). In order to achieve this goal we decided to make a first public release (limited distribution) with the functionality as it is implemented in the DNA system at the time of the meeting with the addition of the possibility to run the system offline. This release will be made available before Christmas 2003.

Sean listed two items that are of paramount importance and hence have the highest priority:

- Beam position. Almost all failures of indexing using DNA have been tracked back to a faulty beam position. Therefore each facility should make sure that the beam position that is used by DNA is the correct one.

- Local issues not in DNA. Sean thinks that most of the failures of DNA on ID 14 were caused by actors or events outside the core DNA development, e.g. DNA-ProDC communication, network problems etc.

Sean and Colin are responsible for making sure that the two items listed above are not causing problems to DNA.

We also decided on the following priority list:

Priority	Description	Responsible
1	Web page documentation. All developers should work on improving the documentation. SK was appointed to coordinate this work.	OS, DS, GW, HP, KA and SK
1	First public release (DNA 0.x), limited distribution. GW is responsible for making a distribution that contains both the DNA system and all components needed for running it (CCP4, Java and Python)	GW
1	Testing. Soleil and EMBL Hamburg should install the 0.x release as soon as it's available and test it using both test images in the distribution as well as other images.	AT, VP
1	Archive of reference images. The DNA system can now copy reference images to a central location. After some days of use on ID 14 we have already got reference images from 50 samples. OS will set up a web page describing these images and make a DVD available for distribution.	OS
1	Space group determination, integrating and scaling: It was agreed to follow two approaches: 1) AWL will provide a script for Graeme to incorporate it in the scheduler for running MOSFLM / REINDEX / SORTMTZ / SCALA and tell him which parameters should be extracted from the SCALA logfile and used as a score for evaluating the correct Laue group. 2) GW and OS will help PL to integrate the XDSAUTO (see Section 4) Python wrapper into the scheduler.	AWL, GW, OS and PL
2	The BEST program should be integrated in the DNA system in order to optimally determine the exposure time. HP will work together with Sasha (EMBL Hamburg).	HP
2	Scoring scheme (See Section 7)	AWL, PL, DL person
2	Diffraction plan usage (See Section 8)	LL
3	Parallelisation and reorganisation of the software (See Section 2)	GW, OS
3	LIMS (See Section 9)	DS, LL, GW
3	GUI driveable externally	SK

2. Report from developers' meetings held since the last full meeting

Olof gave a short report of the two developers' meetings that have taken place since the last full meeting. Since there are minutes from both these meetings I don't repeat their contents here but just provide the corresponding URLs:

- June 11th - 12th at DL:
  http://www.dna.ac.uk/minutes/110603/notes.html

- September 23rd – 26th at ESRF:
  http://www.dna.ac.uk/minutes/230903/minutes.pdf

Two points from the last DNA developers meeting where discussed in detail – DNA re-organisation and parallellisation.

3. Project status and demo of current version

Olof reported that the DNA system is installed on the ID14 stations EH1, EH2 and EH3, and has been working not perfectly but well enough for beamline staff and users to find it useful. There are two indicators showing that the system is used:

1. The person who has his email address entered in the <local_info><dna_contact_email> field of the DNA configuration file gets notified every time there is either a successful "Characterise Crystal" sequence accomplished or there is an error. Olof presented all the emails he has got from the ID 14 stations, and amongst occasional error messages there are many successful "Characterise Crystal" sequences being reported.

2. The DNA system can now copy the reference images to a central archive if the users agree (on the GUI the user can deactivate the archiving mechanism). Over the weekend over 50 pairs of reference images where collected from the ID 14 stations. The reference images retrieved on the beamlines will be stored in a collection of reference images that should be used for testing and further developing the DNA system (see Section 5.1).

Graeme demonstrated the DNA system running on his Linux laptop. What was new since the last full DNA meeting in May was (apart from a help button and the reference archiving enable check box) the integration, see next Section.

4. Online integration, status and perspectives – MOSFLM and XDS

4.1 MOSFLM

Graeme showed in his demo also the current implementation of the image integration scheme. On the GUI there are two buttons, one for "Collect data" only and one for "Collect and integrate data". If one selects the latter the DNA system collects six reference images before the data collection, three at the phi starting angle and three 90 degrees apart. After collecting these images MOSFLM refines the cell. Once the cell refinement is finished the data collection starts. For the moment the DNA system does not integrate the data as it is collected; data is integrated only upon completion of the data collection.

More work is needed to make the integration useful. In particular, the data needs to be scaled in order to assess as quickly as possible the right space group and give feedback to the user about data quality. The space group determination work was assigned priority 1 (see Section 1.2).

4.2 XDSAUTO

Pierre Legrand has been working on a Python wrapper for XDS called XDSAUTO (see the included slides below.)

XDSAUTO

- Simple program written in python (1500 lines)
- Needs only standard python and the XDS package
- Based on a XDS interface that should be easy to integrate in DNA

XDS

- XDS Advantages:
  - Very robust FORTRAN program
  - Uses 3D profiles (masks for summation and σ estimation)
  - Adapted for a very large number of detectors.
  - Well suited for automation
  - Multi processor version for up to 32 cpu's on SMP machine (based on OpenMP)
  - Run on IRIX, Tru64, OSF1, Linux, MacOSX
- XDS drawbacks:
  - Slow
  - has an extremely limited GUI
  - Not optimized for cluster…

XDSAUTO

Reads very simple parameter text file:

```
beamline = ESRF_ID14EH4
frame_first = 1
frame_last = 99
phi_init = 150.0
delta_phi = 1.
prefix = iplow_1_
suffix = img
img_dir = /home/legrand/images/i4
beam_x = 94.641
beam_y = 95.877
anomalous = yes
wavelength = 0.97960
distance = 139.99
angle_det = 0.00
resolution_range = 40 0.0
spg = p63
```

Beamlines characteristics are described by python object.

XDSAUTO

- Run in 6 stages
  - I Peak Search
  - II Auto-indexing
  - III Profile Optimization
  - IV Testing Laue Symmetries
  - V Integration (multithread)
  - VI Exports to other programs
- XDSAUTO can be restarted at any stage

Example of output 1/6

```
Reading Data collection parameters in file                    auto.par

       beamline                ESRF_ID14EH4
       img_template            ../img_dir/mn4_1_???.img
       frame_first             1
       frame_last              181
       delta_phi               0.5
       anomalous               yes
       wavelength              0.939283
       distance                180.01
       beam_x                  94.478
       beam_y                  96.133
       resolution_range        50 2.1


 INFO:  First and last frame number found                    1 - 180
 INFO:  Maximum number of mosix node used                          4
 INFO:  Truncating the highest resolution to                    2.10
```

Stage I - Peak Search

Using images	1 to 2 and 179 to 180
Mean Gain value	0.23
Average background counts	116.6
Number of diffraction spots found	558

Stage II - Auto-indexing

Selected reduced cell axes	74.7 74.6 86.3
Selected reduced cell angles	64.4 89.9 60.0
Fraction of indexed spots	(498/558) 89.2%
Standard deviation of spot position (pixels)	0.53
Standard deviation of spindel position (degrees)	0.05

Selected list of possible Bravais lattices

Bravais Lattice	Symm	fit	cell a	b	c	alp	bet	gam	ORDER	
R hexagonal	(hR)	1.8	75	75	224	90	90	120	3	***
C monoclinic	(mC)	0.8	129	75	86	90	120	90	2	**
C monoclinic	(mC)	2.8	129	75	86	90	120	90	2	**
C monoclinic	(mC)	1.2	129	75	86	90	120	90	2	**
C monoclinic	(mC)	2.0	129	75	86	90	120	90	2	**
P triclinic	(aP)	1.0	75	75	86	90	116	120	1	*

ORDER is defined as the ratio: reduce cell volum/asym. unit volum.

Stage III - Integrate and Optimize Profiles

Selected reduced cell axes	74.7 74.6 86.3
Selected reduced cell angles	64.4 89.9 60.0
Runing integration with Space Group	P1
Anomalous set	OFF
Number of optimization steps	2
Step 1/2 Integrate image range	1 to 8
Initial fitted beam divergence / reflecting range	0.422 / 1.032
New refined beam divergence / reflecting range	0.388 / 0.827
Step 2/2 Integrate image range	1 to 12
New refined beam divergence / reflecting range	0.370 / 0.707

Stage IV - Testing Laue Symmetries

Lat	fit	cell						Symm	Compar	Compl (%)	Rsym (%)	Rmeas (%)	I/sigI	Misf	ORDER
		a	b	c	al	be	ga								
hR	1.8	75	75	224	90	90	120	R3	686	17.3	4.0	5.7	7.4	0	3
								R32	1630	29.1	4.1	5.7	7.9	0	6
mC	0.8	129	75	86	90	120	90	C2	382	11.7	3.7	5.2	9.5	0	2
mC	2.8	129	75	86	90	120	90	C2	382	11.7	3.7	5.2	9.5	0	2
mC	1.2	129	75	86	90	120	90	C2	292	11.8	3.1	4.4	8.7	0	2
mC	2.0	129	75	86	90	120	90	C2	382	11.7	3.7	5.2	9.5	0	2
aP	1.0	75	75	86	90	116	120	P1	10	6.2	4.1	5.8	12.8	0	1

ORDER is defined as the ratio: reduce cell volum/asym. unit volum.

Mean and RMS for Rmeas = 5.31 0.48
Weighted Mean and RMS for Rmeas = 5.45 0.38

Chosen Lattice by order of preference:

1)	74.65	74.65	224.40	90.00	90.00	120.00	R32	155	29.1	5.7
2)	74.65	74.65	224.40	90.00	90.00	120.00	R3	146	17.3	5.7
3)	129.30	74.60	86.40	90.00	120.00	90.00	C2	5	11.8	4.4
4)	129.10	74.70	86.30	90.00	119.90	90.00	C2	5	11.7	5.2

Stage V - Integration

Selected cell axes	74.7 74.7 224.4
Selected cell angles	90.0 90.0 120.0
Selected Bravais lattice	hR
Runing integration with Space Group	R32
	(number 155)
Using frames number	1 to 180
Anomalous set	ON
Using refined beam divergence / reflecting range	0.370 / 0.707

==> Threads XDS started for integration of images 1 to 45
==> Threads XDS started for integration of images 46 to 90
==> Threads XDS started for integration of images 91 to 135
==> Threads XDS started for integration of images 136 to 180

<== Threads XDS finished for integration of images 1 to 45
<== Threads XDS finished for integration of images 46 to 90
<== Threads XDS finished for integration of images 91 to 135
<== Threads XDS finished for integration of images 136 to 180

Integration time: 189.2 seconds

Scaling Statistics

Resolution	100.00 - 2.10	(2.18 - 2.10)
Completeness	99.8%	(100.0%)
I/sigma(I)	13.9	(3.9)
Rmeas	5.2%	(37.2%)
Rsym	4.2%	(29.8%)
Compared	76888	(8110)
Measured	77432	
Unique	27149	
Multiplicity	2.9	
Rejected misfits	14	
Anomalous contrib.	1.1	
Wilson scaling (B/Corr)	49.5	0.99
Estimated Res_max	2.10	

STOP

5. Setting up a database or collection of reference images – ranking of crystals

5.1 Database or collection of reference images

We agreed that it would be very useful to have a collection or database of reference images. The idea is that these images could be used for testing and further developing the DNA system. As reported earlier (see Section 3), the DNA system can now copy reference images to a central location. Since the size of these images can be of the order of Gigabytes and since many of them come from users' experiments, it was felt that it is safer to distribute the images on DVD disks. Olof was appointed to organize the collection or database of reference images.

5.2 Ranking of crystals

Harry contributed the following résumé of his presentation:

Two categories of ranking will ultimately need to be considered:

1. Ranking crystals of different types (different projects)
2. Ranking crystals of the same type (crystal screening)

At present, only consider the second, as the first will be more difficult to implement.

Factors that will contribute to the ranking

1) Success of autoindexing (crucial).
Crystals which cannot be indexed according to the defined criteria (rms error in spot positions, fraction of spots successfully indexed) will be assigned a zero score.

2) Strength of diffraction as defined by $I/\sigma(I)$
Need to consider how this is defined. Three possibilities are:

(i) Overall (ii) At "desired" resolution (iii) At "lowest acceptable" resolution

where the user has specified the "desired" and "lowest acceptable" resolutions. Probably need a combination of (ii) and (iii), because none of the crystals may diffract to the "desired" resolution.

This is by far the most important criterion.

3) Presence of ice (or salt) rings or spots

Ice rings can be detected in the image by looking at the background at known resolution shells. Salt rings (where the resolution will generally not be known) can be detected using Graeme's code which recognises circles of spots.

Ice or salt spots (with no rings present) are more difficult to detect. One possibility is to examine the intensity distribution in resolution shells for outliers. This will work if the ice/salt spots are much stronger than the protein spots.

If spots or rings are present, this could result in a penalty factor proportional to (1-f) where f is the fraction of data that is lost due to exclusion of resolution shells around rings. The penalty will then be less for very high resolution data than for low (as the presence of ice rings may not be considered to be such a problem if collecting atomic resolution data as when collecting lowish (2.5-3Å) resolution data.

4) Presence of more than one lattice
Use number of spots rejected (not indexed) at autoindexing stage to assess the presence of additional spots. Score is the fraction of spots successfully indexed.

5) Good spot shape
This will usually influence I/σ(I) as well. Could also use correlation between standard profiles in different regions of detector ? This will probably also affect the rms error in spot positions (bigger spots will give bigger errors). By using the weighted rms error (rather than the actual error) this should at least partly remove the correlation with the strength of diffraction (ie I/σ(I)). Score is a combination of these factors (mean correlation coefficient and weighted rms error).

6) Mosaic spread
Will also affect I/σ(I) as well, so may not need a separate score.
Note: In extreme cases, may lead to rejection of weak spots at lune edges due to spot overlap, artificially *increasing* I/σ(I) for remaining spots.

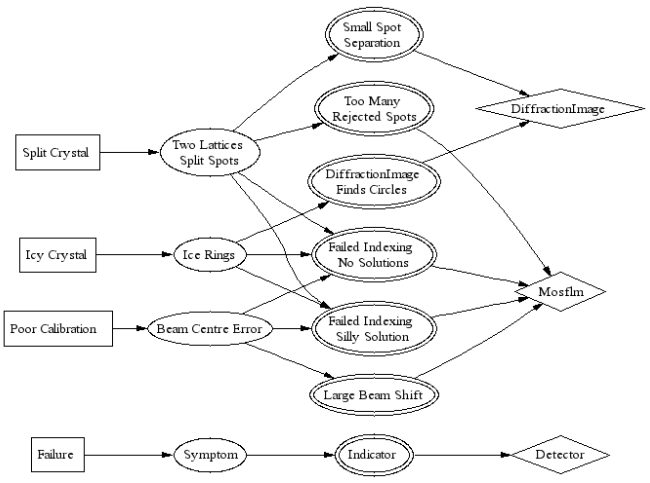
6. Causal graphs – how to represent knowledge

Graeme presented the following slides:

Causal Graphs

- A “causal graph” is a way of expressing cause and effect, so that causes can be determined from effects.
- The “effects” are split into symptoms and indicators – a symptom bay be a deduction from a number of indicators.
- There are four layers, Failure, Symptom, Indicator and Detector.

An Example



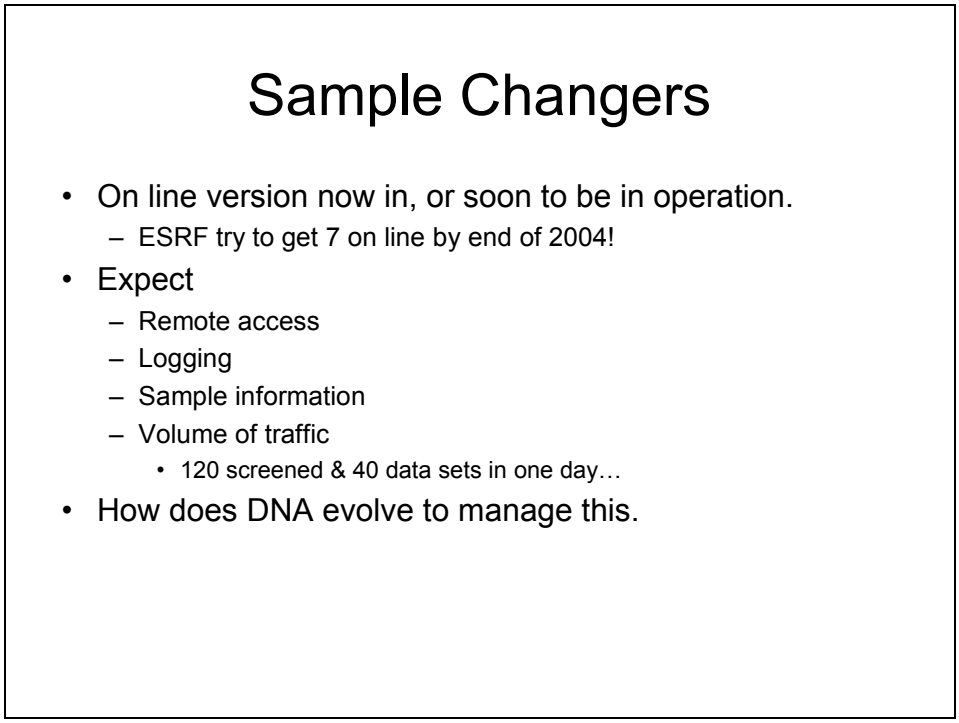
- This causal graph indicated a few of the things which can go wrong.
- This kind of information can be naturally encoded in an expert system shell, for example clips….

```
(defrule work
(and (it is a weekday)
     (I am not on holiday))
 => (assert (go to work))
)
```

- Therefore a conclusion can be drawn from a number of facts. This is a great deal simpler to manage than an enormous number of if, then, else statements!

7. Sample changers and LIMS

Sean presented the following slide as an introduction to a discussion about sample changers:



The subsequent discussion was very lively and hence difficult to resume, however many of the discussion points can be found in the following sections.

8. Scoring schemes for multiple samples

Andrew Leslie contributed the following résumé of his presentation to these minutes:

Can combine individual scores using "Z-score" (as used in SOLVE). Define the score, $Z(X)_i$, for criterion X for the ith crystal as:

$Z(X)_i = (X_i - \langle X \rangle)/\text{sigma}(X)$

Where X_i is (say) $I/\sigma(I)$ for the ith crystal and $\langle X \rangle$ and sigma(X) are its mean and standard deviations over all crystals.

The advantage of using Z scores is that it should allow a straightforward combination of several individual scores (without additional weighting terms). It is not clear how well this will work in practice. Individual scores would therefore be combined with individual weights to give the overall score for the crystal:

$Z_i = A*F*Z(I/\sigma(I))_i + B*Z(\text{rejected spots})_i + C*Z(\text{spot shape})_i + D*Z(\text{mosaic spread})_i$

Where F is the factor to allow for ice/salt rings/spots and A,B,C,D are weighting factors to be determined.

Note only one $I/\sigma(I)$ term is included, but in practice this could different terms (with their own weighting ?) for data at different resolutions (eg desired and minimum acceptable).

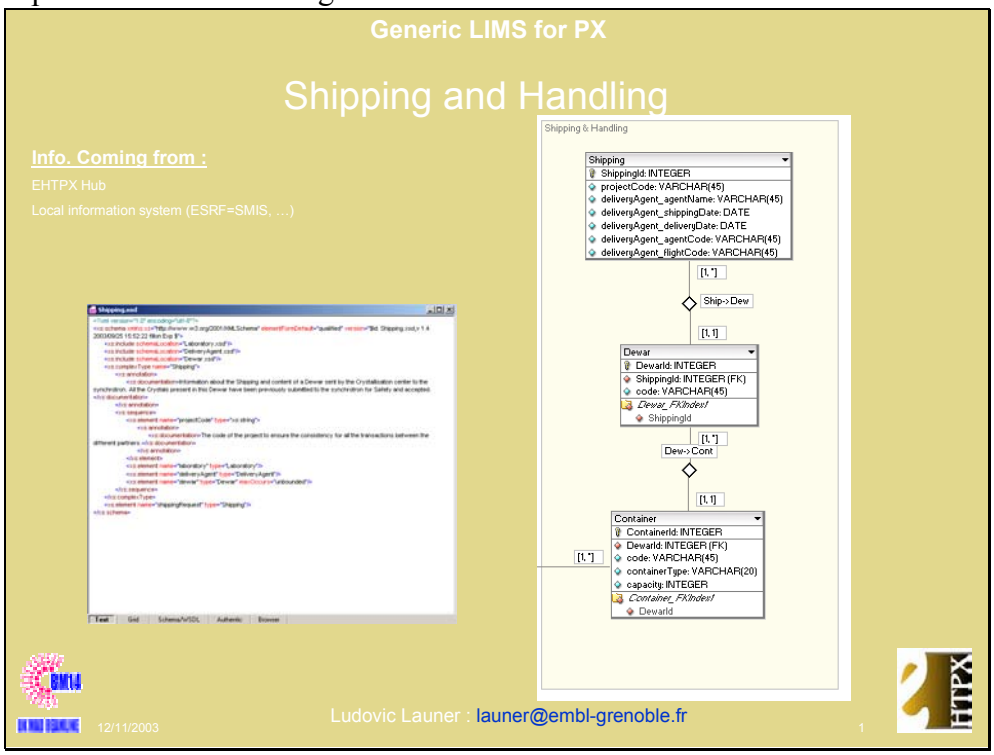
We need a good basis set of images, scored by "experts" to determine how well this will work and to choose values for fitted parameters A,B,C,D,F.

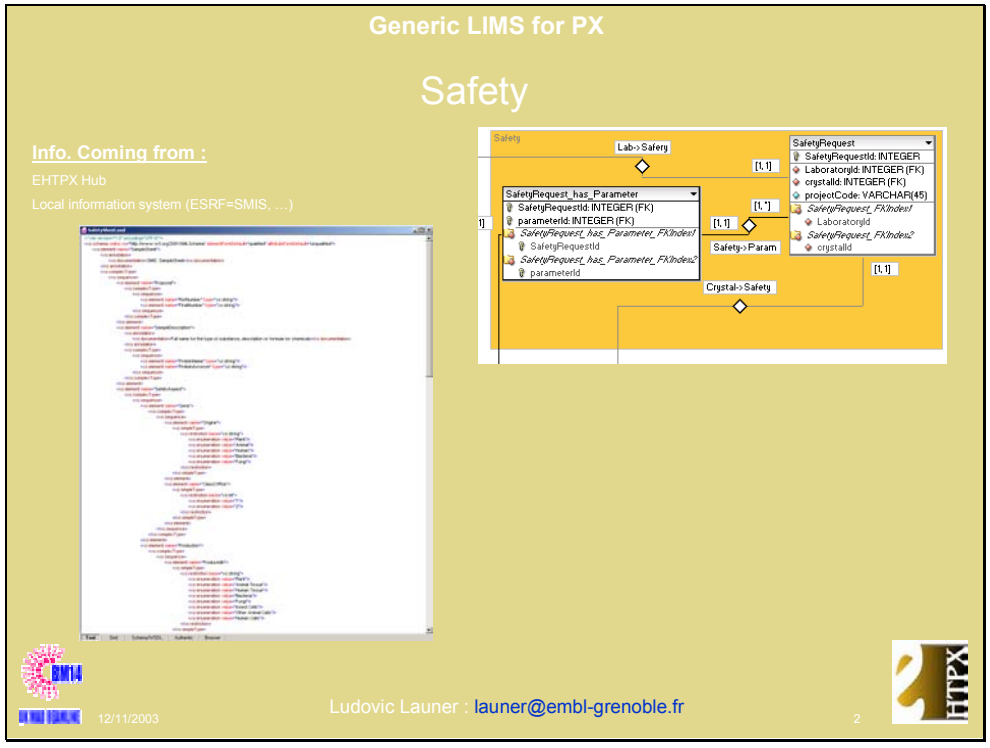
Problems

$I/\sigma(I)$ varies in different parts of reciprocal space (for example, it can show quite a variation from one image to the next).We will not (in general) be comparing statistics for identical reflections from the different crystals, so we need a measure of when the difference in $I/\sigma(I)$ for different crystals is actually significant. A better measure might actually be obtained by fitting the observed variation with resolution to a Wilson plot, as done in BEST.

9. Automation – what information will come along with a Fedexed sample

Ludovic presented the following slides:



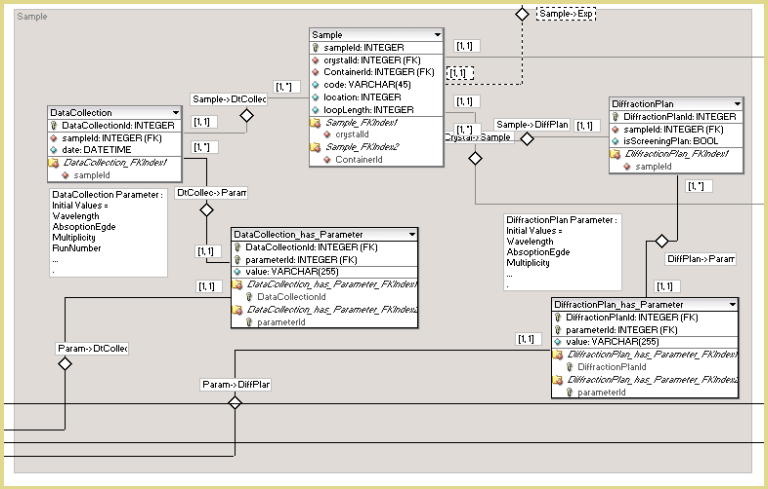


Sample

Info. Coming from :

EHTPX Hub

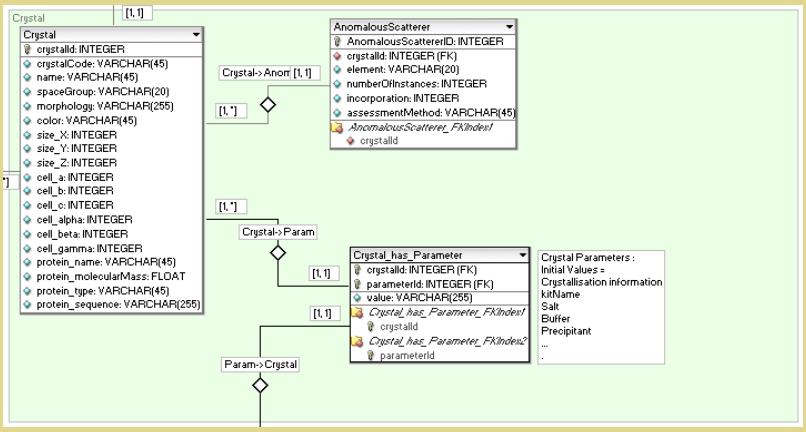
Beamline control module



Crystal

Info. Coming from :

EHTPX Hub



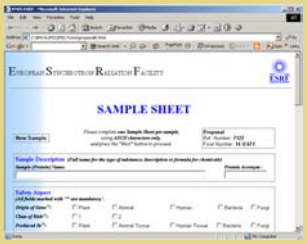
Generic LIMS for PX

The path to High Throughput : Automation

Case study : The Sample (safety) sheet

Present situation : Manually fill in a web form

- Information might have to be typed in several times (BM14, ERSF, …)
- Does not allow a “high throughput approach”

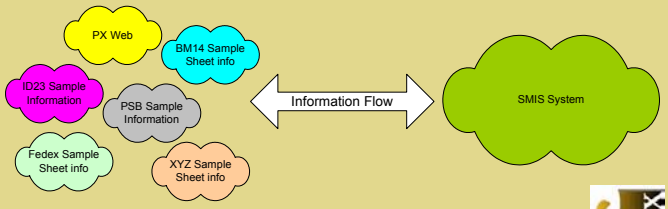


Solution : Automatic submission = bridge to the existing system

Required information not typed in but “given” automatically through a program

Possible implementation :

- WebServices
- Socket communication
- …



Ludovic Launer : launer@embl-grenoble.fr

Generic LIMS for PX

An introduction to WebServices

A collection of XML-based technologies developed to address issues of:

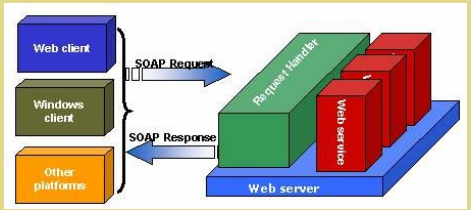
– Service discovery - Business processes

– Interoperability - Data exchange

http://www.webservices.org

Sample Sheet WebService up and running.. Ready to forward information to SMIS :

```
<?xml version="1.0" encoding="UTF-8"?>
<SampleSheet xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="./SafetySheet.xsd">
<Proposal>
    <RefNumber>345</RefNumber>
    <FinalNumber>U345</FinalNumber>
</Proposal>
    <SampleDescription>
    <ProteinName>SFG</ProteinName>
    <ProteinAcronym>SFG23b</ProteinAcronym>
</SampleDescription>
……..
</SampleSheet>
```



Ludovic Launer : launer@embl-grenoble.fr

More information available at :

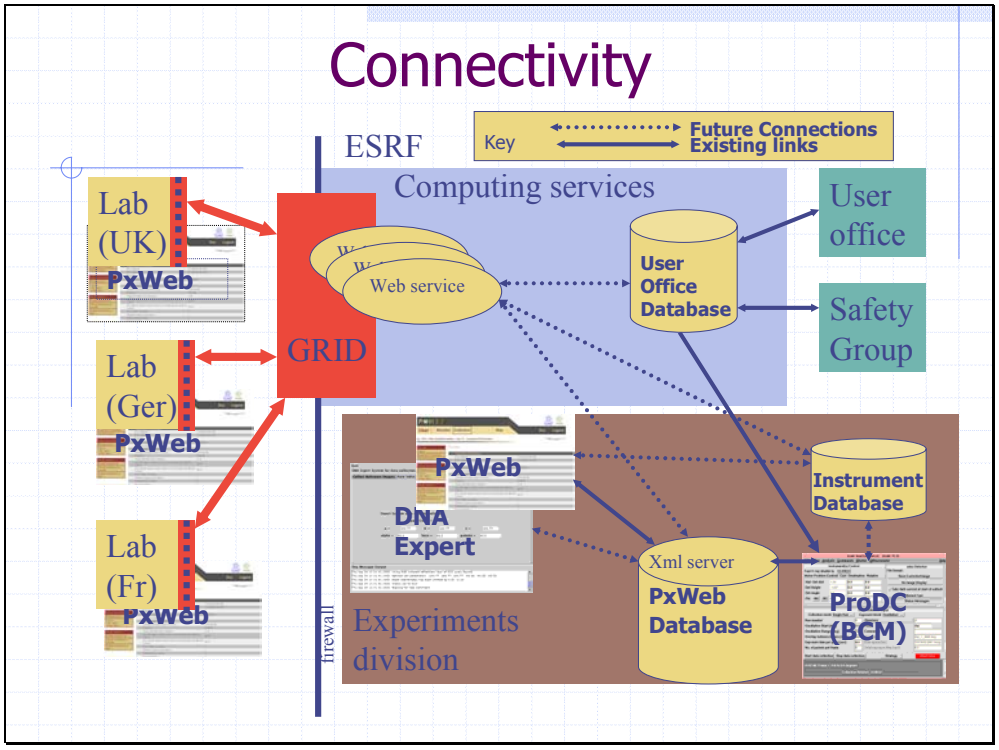
http://www.ebi.ac.uk/msd-srv/docs/ehtpx/fillon/

http://www.bm14.ac.uk then browse to eScience Section

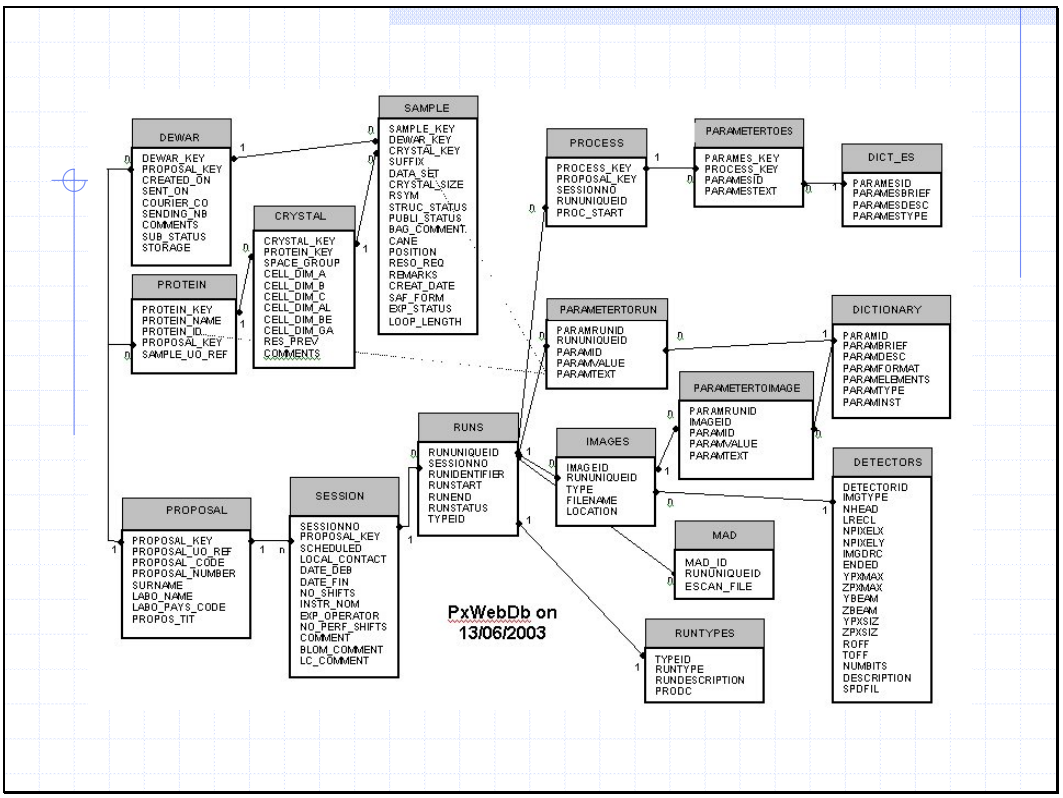
Ludovic Launer : launer@embl-grenoble.fr

10. LIMS

Darren presented the slides below as an overview of the ESRF beamline LIMS called PXWEB. As for the sample changer, the discussion was very lively, we agreed however to pursue the initiative taken during the last DNA developers meeting (See http://www.dna.ac.uk/minutes/230903/minutes.pdf, Section 1).



This is a diagram showing potential and existing connections between outside laboratories, inside the ESRF infrastructure and various programs that need to use the database.
Dotted arrows are potential connections, solid arrows are existing connections.



11. Relationship with other projects

Colin contributed with the following résumé:

Several funded projects (e.g. SPINE, e-HTPX, and soon BioXHIT) have specific deliverables to link x-ray data collection and analysis. In addition several institutions (MRC Cambridge, SRS, ESRF) are putting in their own resources to do this. The Max-Inf funding for the travel for the DNA project will cease in just over a years time but the meeting felt the present collaboration (which is owned by no one institution) should continue. Andrew Leslie, Sean McSweeney and Colin Nave should discuss with Gerard Bricogne (the relevant workpackage leader for BioXHIT) how the DNA project could fit in with BioXHIT.

12. Next meetings

It was decided that the next full meeting will take place in Cambridge in June 2004.

Olof proposed to have the next developers meeting at the EMBL Hamburg in order to boost the collaboration with this institute concerning in general DNA and in particular BEST.