Notes from the DNA Meeting

30th September 2002

LMB, Cambridge

Present:

Andrew Leslie (LMB), Darren Spruce (ESRF), Liz Duke (DL), Charles Ballard (CCP4), Harry Powell (LMB), Graeme Winter (LMB), Steve Kinder (DL), Sean McSweeney, Colin Nave (DL), (DL), Takashi Tomizaki (SLS).

1. Report on Progress with the Expert System

Steve Kinder presented a report on the work done so far on the expert system.

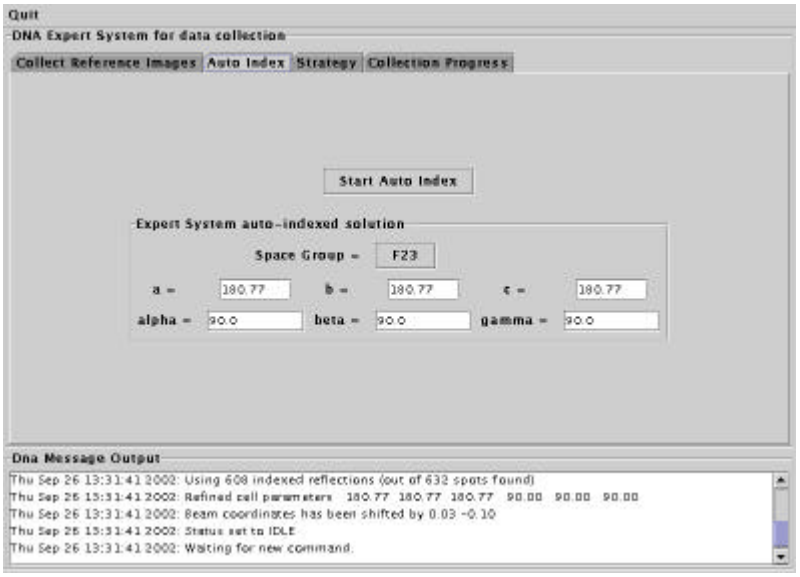
There have been two developers meetings since last full meeting, one in April and the other in July, both at the ESRF. At the first meeting beam time was used to test the system. These tests highlighted the following issues:

- Pointed to a need to increase robustness
- Discussed XML and decided to develop a common GUI

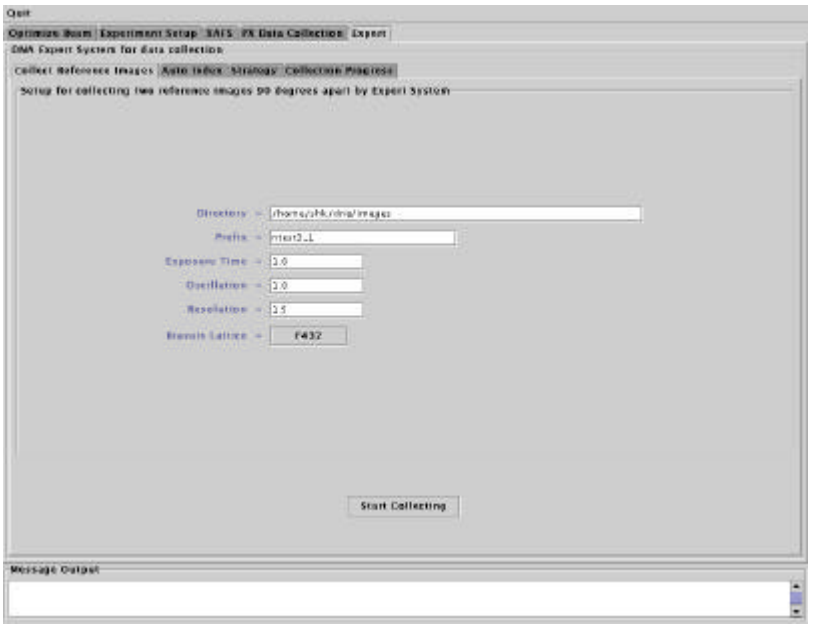
The second meeting was more hands-on, centering on testing and further developing the GUI. This meeting was felt to have been very successful as a lot was achieved in a relatively short period of time. It was felt that the practical nature of meeting worked well.
To summarise the progress of the DNA project:

- Robustness improving and abort now possible
- Level of hard coding reduced
- First version of stand alone GUI developed
  - Collect, Index and Strategy controllable
- XML now defined in XML Schema
  - Coupled with Castor leads to easy parsing by GUI
  - Easier to understand method for formalising XML

Steve showed an image (displayed below) of the ExpertGUI. There are tabbed screens for the various actions required: collect reference images, autoindex, strategy and collection progress.



In PXGEN++ (the data collection GUI developed at Daresbury) there will be a tab in the PXGEN++ window for the expert GUI as shown below:



Steve then outlined the future plans for the short to medium term:

- Continue to improve robustness
- Develop intelligence into Expert
  - Detection of problems with images
  - Integration of reference images
- Continue to develop GUI
  - Feedback results of strategy properly
  - Build automatic chaining of function set to provide Characterise Crystal

Ultimately more testing is required and the deadline of a release would focus minds!
In the longer term there are needs to:

- Develop further intelligence into Expert
- Develop GUI to allow crystal screening, automatic collection and processing
- Allow interactions with an experimental database
- Incorporate other underlying programs (e.g. strategy software)

A discussion followed on the possibility of including Sasha Popov's strategy program which is currently the only program available which allows an estimate of exposure time required to achieve a given $I/\sigma I$. It is currently unclear whether the source code for the program would be made available. It seems that there are differences of opinion at the EMBL over releasing the source code as it was understood that EMBL are exploring the possibility of a collaboration with MAR. It was felt that if the DNA project could demonstrate that the project was being held back due to lack of code rather than just expressing interest then it was possible that the EMBL might be more forthcoming. The outcome of the discussion was that the person responsible for the person incorporating strategy into the expert system would make a formal request to Victor Lamzin at the EMBL for the code for Sasha's strategy program. If that failed to induce the required response then Keith Wilson would be asked to intercede on behalf of the dna project as Keith is active in almost all of the various projects in this area.

In addition with the imminent arrival of additional effort on the project new challenges will need to be faced in co-ordinating the effort which will be from several different sources. Possible solutions to this would be:

- Create a Project Plan
- Create a clear split of responsibility
- Explore the most efficient methods of working whilst allowing for flexibility to accommodate different working patterns and different requirements of local institutions.
- Ensuring good communication is vital for success.

Darren Spruce gave a brief presentation on the situation at the ESRF. He commented that Olof Svensson's opinions on the status of the DNA project mirrored what Steve had said in his presentation.

Prior to the DNA collaboration the ESRF had the desire to link data collection and data processing. To aid this a mysql database of parameters was created. This database would also be linked to the ESRF User Office database. Ultimately also the database would be linked to the EMBL sample changer control software. In addition a link would be created to a database where beamline parameters are stored. This beamline parameter database is currently evolving and it is expected that a snapshot of the beamline would be taken at suitable intervals. This would aid problem diagnosis on the beamlines and also enable a pro-active response to beamline maintenance. Darren Spruce offered to provide documentation of the database to anyone interested in it.

Steve Kinder then gave a demonstration of the expert system GUI. He showed the first window of the tabbed pane (as shown earlier), the collect reference images pane. Here values for the exposure time, oscillation range and resolution can be entered. Commands are then issued so that the images are collected and then autoindexed (via the next tabbed pane) and the results reported back. Raimond Ravelli had provided very useful feedback about how the results should be reported – what information was required etc. However it was felt that this was very much a personal thing with different crystallographers wanting different information. The autoindexing results are reported back in display boxes in the GUI. A strategy calculation (the next tabbed pane) is then done based on the autoindexing results. The strategy results are fed back via a table. Olof is working on the design of the table in the expert system and Steve is working on displaying the table within the GUI. It is ultimately expected that the need for separate tabbed panes for collecting the reference images, autoindexing and strategy determination will be removed as the different sets of commands are chained together. It was also recognised that interaction with other pieces of equipment (eg sample changers) would also be required. It was commented that there now existed a sufficiently large library of images that actual beamtime was not required for testing the software. Graeme Winter offered to burn CD's of the images already available to aid people in the testing of the expert system/GUI. However it was felt that having the deadline of beamtime did concentrate the mind on the goal ahead, which meant that more progress was made. It was felt that currently progress on the project was made in spurts rather continuously. However in each case, work on the dna project must be fitted in around the demands made by other aspects of the job. Communication was again highlighted as very important – it was felt all too easy to end up in the situation where everyone is waiting for everyone else to do something when in fact that task has already been completed.

After the developers meeting in April a list of problems with mosflm was made and questions were asked about the progress made in this area. There was some confusion on whether these problems had been solved. This was felt to be another example where improved communication would be advantageous.

The possibility of having another developers meeting was discussed. It was suggested that having a week long meeting in mid-November at the ESRF would be a good idea. Prior to the

next developers meeting it was proposed that a plan be set up covering which work areas would be tackled and what the objectives of the meeting would be.
The issue of quality indicators was discussed – in particular with respect to whether images have been correctly indexed or not. It was felt that at the start simple rules would be made to classify whether an image had been correctly indexed. For example the rms dev of actual spot position from predicted spot position could be used along with possibly the number of spots used/rejected. It was suggested that Andrew Leslie create a recipe that would allow a score (figure of merit) to be output. It would be useful to then store information on whether this implementation (which would initially be crude) actually worked.
RESULTING ACTIONS:

- **Graeme Winter to send out CD's of test images previously collected.**
- **Andrew Leslie to create a recipe to allow a indexing score to be output from mosflm.**

2. Future plans for the expert system

Colin Nave outlined the plans associated with the BBSRC e-science grant (e-HTPX). Many people from different organisations are involved with the grant including BM14, Randy Read, Daresbury Laboratory, EBI, Oxford, and Kevin Cowtan. The aim of the grant is to ensure that there is coherent flow of data available from the selection of the target through to the final deposition of the co-ordinates and that the data is accessible at all times. Currently aspects of the project include:

- User Interface
- Grid Portal
- Tracking projects
- Automation of data collection
- Parallel processing of data analysis
- Data management
- Outreach to industry (particularly with respect to security)
- Data model

The user interface is being done at York.
The data model is being tackled by the EBI in collaboration with CCPN (the NMR CCP) who have already spent some time looking at data models and who have similar requirements to PX. The data model will be done in UML and will cover a description of the project from target selection through to final structure. It is important to ensure that the data model be easy to extend, easy to maintain and easy to understand and most importantly bear some relation to reality in order to maximise the level to which it is taken up by the community as a whole.
It is expected that there will be links into Bioinformatics work that is being done in Oxford in relation to the use of LIMS systems.
Regarding the data model – it was highlighted that there are many exchanges of information that take place between data acquisition and data processing. However it was not known how much of this would be transferred to the data model. However a meeting is to be arranged for next year where an attempt will be made to standardise on one (possibly European) data model.
There are clear links between the work being done as part of the DNA project and the aims of the e-HTPX work especially if deposition of the final atomic co-ordinates is required. It is quite possible that ultimately the pdb will require an indication to be given of the X-ray dose received in order to place a quality indicator on the structure. It is expected that there will be close interactions between the person based at the EBI and those working on the DNA project.
The RA position on automation of data collection has been accepted by Graeme Winter. He is likely to take up his position in early November.

BM14 has an RA position to work on project tracking and MAD data collection ie tracking data from when the samples arrive to when the data leaves. It is expected that this area will build on what has already been done at the ESRF.
Workplans for each RA on the e-htpx grant are being drawn up to ensure that all know what others are doing and appropriate links between the activities identified. It would be very easy for several RA's to end up doing the same thing or repeating what someone else has done before the e-HTPX grant came into being. Communication was again highlighted to be key for success. It was noted that communication should be both at the PI level and also at the RA level.
In addition to e-HTPX there is also the SPINE project which has several positions available in the area of high throughput work in particular Work Package VI which is the work package associated with SR facilities and is being led by Stephen Cusack. This particular work package related to work on automatic alignment of beamlines and development of an expert system. 6 positions are available with SPINE however only 1 is to do with software development. Again with this project it was felt that clarification was required over who was doing what. It was hoped that the SPINE meeting in January would help to clarify the different roles.
The ESRF have a software appointment associated with Spine WP VI. Sean gave a brief summary of work going on at the ESRF and how the work might fit into existing work at the ESRF. A summary flow chart is attached to the end of this document that summarises what was said.
Again difficulties in co-ordinating all the work between different institutions with their different methods of working and conflicting demands were highlighted. In such a huge project as this it is important to ensure that people don't repeat work that has already been done by someone else. For example a lot of the groundwork associated with the automation of data collection has been tackled already in the development of PXWEB. It is important to ensure that good communication is in place so that "short circuits" do not develop with everyone thinking they are waiting for someone else to complete a certain task. It would also be easy for frustrations to creep in and the for people to feel that it would be quicker to do a certain task themselves rather than wait for someone else to do it as part of their portion of the work package. Concerns were also expressed that some organisations were less open about what they were doing in the area of automation of data collection than others were. This problem will inevitably lead to reinvention of the wheel and subsequent frustrations.
The work package for the automation of data collection has clear overlap with the work of the dna group. It was thought to be worthwhile to expand the dna group to try and co-ordinate effort and ensure that work is not repeated while ensuring that milestones for the individual projects were met.
Two significant issues were raised regarding the DNA project:

- Does DNA get any more effort as a result of these initiatives being funded?
- How do we make sure that what is done meets the milestones set by the various initiatives.

The work completed by those working on the dna project formed a good basis for other people to meet the objectives set by their initiatives. For the automation of data collection part of the e-htpx project, Colin Nave said he envisaged the RA working as part of the dna group and providing extra effort to extend the project further. It was agreed that a single person should take responsibility for co-ordinating the different aspects of the dna project.

- Expert System GUI: Steve Kinder
- Expert System: Olof Svensson
- Mosflm (plus server): Harry Powell and Andrew Leslie (though it is anticipated that Graeme Winter will continue to contribute in this area at least in the short term).
- PX Database: Darren Spruce

However it is the responsibility of each SR source represented to ensure that all aspects are working on their facility even though they may only have overall responsibility for one area. It was felt to be key to ensure that the mosflm server was working reliably on all sites (Cambridge, Daresbury, ESRF and SLS) in order for any progress to be made.

3. Plans for future tests of the system

Further tests of the system would include both on-beamline and off-beamline tests. It was felt important that tests would cover aspects of the expert system plus GUI and also work would be done to identify any instabilities within Mosflm. Sean commented that ESRF had time immediately available on Wednesday (2nd October) and also during the single bunch period in mid October. However it was pointed out that data were already available which could be used for testing and Graeme would email out CD's of the data to those who wanted to receive one. Sean commented that, if required, it would be possible to obtain a day or a shift per run of 6 weeks at the ESRF. However commitment to this beamtime would be required. It was pointed out, that at the end of the day, one of the most important requirements for the system is that it be robust. It would be a PR disaster if the system proved to be unusable purely because it was not sufficiently robust to withstand normal wear and tear on a beamline. It was pointed out that all of the people involved in the project at present are juggling this project along with their other many duties.
It was suggested that there should be tests on the beamline every two months. In between these tests, additional tests could be done using images previously collected. During the tests on the beamline all problems identified should be listed and at the end of the time a list be circulated to all on the DNA mailing list. Those involved in each of the aspects of the project where a problem has been identified should then respond with a timescale for fixing the problem. Based on the timescales the next test would be scheduled, aiming to have roughly 2 months between tests. It was suggested that Sean schedule a day/shift of beamtime for some time between mid-November and early December to follow on from the work generated by the beamtime on Wednesday 2nd October. Time would also be made available for testing the system on the SRS in December.
Takashi discussed how progress could be made at the SLS. In the first instance work would be done to integrate the mosflm server and the expert system in the existing data collection GUI. This was felt to be of significant benefit to the project as a whole as it would provide information on the portability of the work that has already been carried out. Takashi also pointed out that with the SLS being in the early stages of operation there may well be the possibility of more flexible access to beamtime for testing.

4. Identification of targets for development.

Further developments required were discussed and the following items were identified:

- Feedback of strategy (parsing the information)
- The "collect" command must actually do something.
- A mechanism to estimate the exposure time needs to be developed.
- Andrew needs to develop a method of estimating a figure of merit for the indexing solutions returned.
- Tests must be made on the way the system handles errors
- A method of chaining together the sequence of events, which are currently done individually, must be developed.

- The parameters that are to be saved to a database need to be identified. Information which is not normally seen but might be required at a later date eg for learning and improving the system needs to be identified.
- It was also suggested that Olof ensure that the XML documentation was up to date.

5. Alternative suggestions for the DNA acronym.

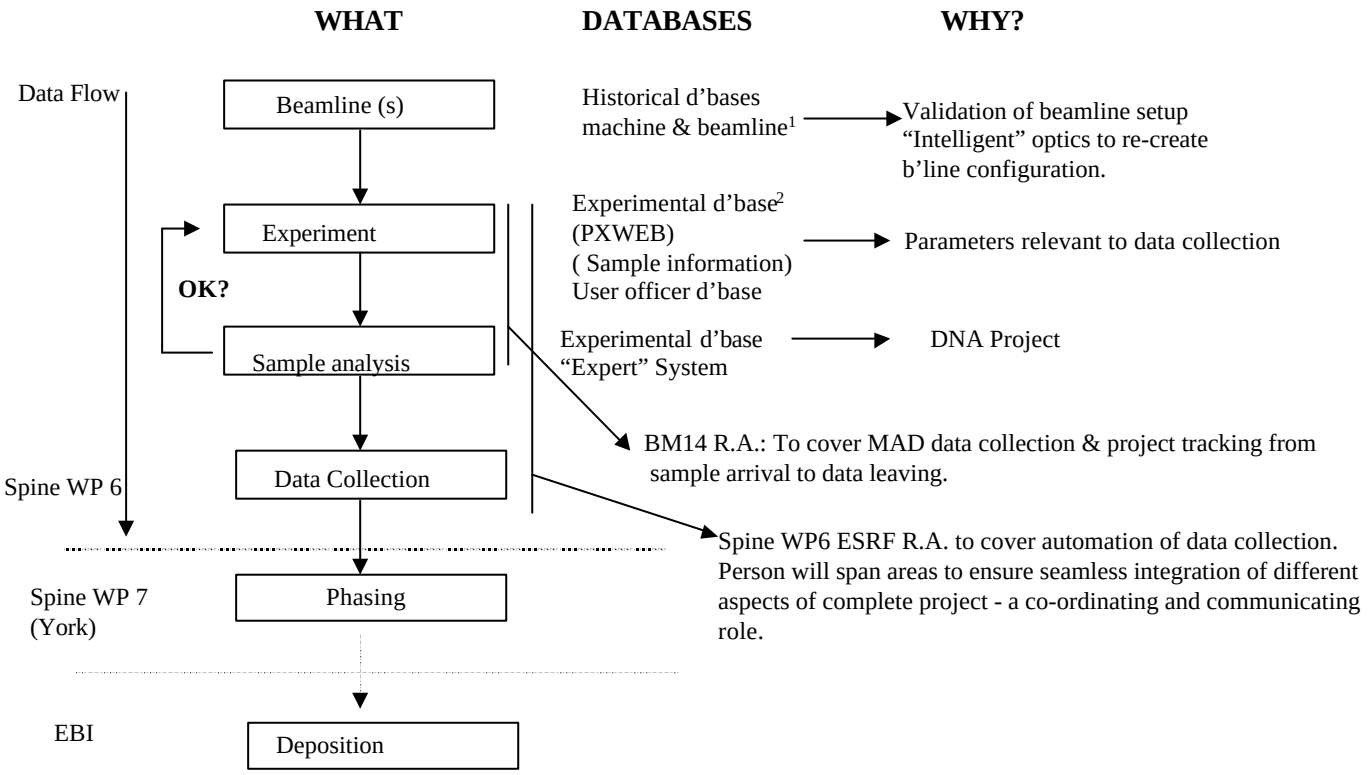
A discussion took place about alternative meanings for "DNA". Several alternatives were put forward such as "distributed network analysis" and "data nearly automatically". However in the end it was concluded to simply remove "dna's not autostruct" from the website. Steve Kinder was asked to do this.

ACTION: Steve Kinder to remove "dna's not autostruct" from the website

Future Meetings

The developers will meet up when there is beamtime for testing the system.

The next full meeting is planned to take place at the end of February at the ESRF. IT was highlighted that the ESRF user Meeting is 12/13th February so possibly a meeting could take place around this time if people were intending to attend the user meeting.



1: This is the work of Darren Spruce and Jeam Michel Chaize (amongst others)

2: This is the work of Olof Svensson and Solange Delageniere (amongst others)