

Summary of BIOXHIT Cross-section Developers Meeting on Data Tracking

Daresbury Laboratory, UK, 14-15th September 2006

1 Introduction: aims of the meeting

Following the review of the BioXHIT section activities after the 2nd Annual Meeting, a number of cross section working groups were proposed in order to encourage more coordination between the standards and data management activities in section 5 and the activities in other sections of the project.

This small workshop was one such cross section working group meeting, and focused on integration of some of the tasks in section 5 (particularly those concerning data exchange and data storage) with some of the tasks in section 4 (concerned broadly with software for structure determination). The aim of the meeting was to identify practical areas for collaboration between some of the developers working in Sections 4 and 5, specifically:

- The standard data items described in XML that need to be transferred between software components within software pipelines
- The data items that should be stored in the tracking database in task 5.2 (which may be the same as those being transferred)
- Items that are needed for deposition (again, these are most likely a subset of those above), and
- Making the project tracking database available to developers of software pipelines (essentially, the provision of an API)

As there was some overlap with the objectives and activities of the CCP4 XML working party the CCP4 Automation Project, for the purposes of this meeting, considered to fall into the activities of partner 10 (CCP4). The overall aim of the meeting was to identify targets that could be achieved practically in these areas within year 4 of the BioXHIT project.

The programme for the meeting can be found at the URL:

http://www.ccp4.ac.uk/peter/workshop/BIOXHIT14-15Sept2006/BIOXHIT_Developer_WG.html.

This report briefly summarises the presentations, discussions and outcomes from the meeting.

2 Participants

The meeting participants and the BioXHIT Partner whom they represent are given in the following table:

Participant	BioXHIT Partner	Abbreviation
Peter Briggs (chair)	CCLRC-CCP4 (10)	PJB
Wanjuan Yang	CCLRC-CCP4 (10)	WY
Avi Naim	EMBL-Hinxton (1C)	AN

Santosh Panjekar (2 nd day only)	EMBL-Hamburg (1A)	SP
Alun Ashton	DIAMOND (17)	AWA
Graeme Winter	CCLRC-SRS (3)	GW
Martyn Winn	CCLRC-CCP4 (10)	MDW
Charles Ballard	CCLRC-CCP4 (10)	CCB
Ronan Keegan (breakout only)	CCLRC-CCP4 (10)	RMK

3 Summary of the meeting

The meeting took place over two days. The intention was that each participant should give a brief overview of their activities and interests, and that based on these presentations a number of breakout groups would be set up. Each breakout group would form around a particular set of issues common to a subset of participants, and would focus on possible integration activities between the partners in that group.

In practice the initial presentations stimulated sufficient discussion at the time that the breakout groups were effectively relegated to the end of the meeting on the second day.

3.1 Avi Naim: BioXHIT Standard XML Tags/Deposition Issues

AN's primary concern was that the EBI contribution to BioXHIT is due to end at the end of 2006, one year before the scheduled end of the project as a whole. Therefore he was keen to focus on what could be achieved practically before the end of the year. His aim was to have some data flowing as a demonstration of the standards that have been developed, and that the best chance of achieving this is to focus on deposition into the PDB.

His interests in this area concerned collecting data from the DNA system to populate the PDB REMARK 200, and to discuss collecting data from the EMBL Autorickshaw software pipeline. The "DNA and Deposition" breakout group was proposed to discuss some of these issues.

MDW mentioned that a number of CCP4 programs already provide much of the data required for deposition via the data harvesting mechanism. REFMAC5 was cited as a specific example, and some heated debate ensued regarding the stability of the REFMAC5 output for deposition. It was forcefully suggested by CCB that the deposition sites needed to notify CCP4 when unexpected changes occurred. However AN said that he would prefer an "external" solution from within the CCP4i tracking or from the individual pipelines.

AN also mentioned the draft document for how to build the proposed XML for BioXHIT (essentially this is a framework), but felt that SP's presence was required to discuss this more fully.

3.2 Peter Briggs: CCP4i Tracking Database

PJB gave an overview of progress with the BioXHIT task 5.2 CCP4i project tracking database. The system comprises three components: the database handler/server process, the database, and the visualisation application. The database component was most relevant for this meeting, and was further intended to comprise three parts: a tracking/history aspect, a “knowledge base” aspect (storing “generic” crystallographic data common to all applications), and an “operational” aspect (enabling storage if necessary of application-specific data). PJB attempted to clarify that the aim was to make a data model for storage only – not a general PX software data model.

Work had been done through 2006 by WY to build a large SQL database that encompassed the tracking and knowledge base parts, but there had been a number of problems with this approach. At the same time a more modest tracking database built on the existing CCP4i def file database had been implemented.

In response to AN’s presentation it was suggested that the existing CCP4i database should contain sufficient information for the deposition process, and that in principle it should be possible to extend the system to write out agreed BioXHIT tags for deposition by the end of the year.

AN commented that CCP4i holds explicit lists of tasks that have been run but not necessarily the precise CCP4 programs within those tasks. However it is the underlying programs that are needed in deposition, therefore some way of extracting this additional data from the CCP4i task log files would be required. The “Deposition and Tracking” breakout group was proposed to discuss the incorporation of exchange tags into the tracking database.

3.3 Martyn Winn: CCP4 XML Working Party

MDW gave a brief report of progress with the XML Working Party since the last meeting of BIOXHIT partners (in November 2005, also at Daresbury). He presented a reasonably long list of CCP4 applications that have been converted to output XML of some form, but noted that the drive in almost all cases was to facilitate the operation of some higher level system (e.g. DNA, CCP4i, CRANK and BALBES). He noted that as a result of this, there are a number of different styles of XML being used in the CCP4 applications that have already been XML-ised, and therefore it would be useful to discuss how these might be standardised in future.

AN asked what commitment the program authors had made to maintaining the XML output in future; MDW responded that there was no long term commitment to do this, which raised the issue of how to ensure that it would be maintained. It was not clear how this could be ensured. GW suggested that based on his own experiences with e.g. Phil Evans and POINTLESS, that it would need active participation from and interaction between pipeline developers and the program authors.

AN then asked whether CCP4 would consider maintaining some form of processor application that could take an existing XML output and then gather any missing items by processing raw program log files, to produce one large XML file. MDW felt that while this could be done for a specific purpose but that it would be impractical to do this in a general fashion. It was also noted that this doesn't really solve the problem of maintaining the XML, it simply shifts the problem from the programs to the processor.

A "Standardised XML Tags" breakout group was convened to discuss some of these issues.

3.4 Graeme Winter: DNA, e-HTPX and XIA2

GW covered the relevant aspects of three different projects that he is involved with:

- e-HTPX: this has an exchange model covering the structure determination process from wet lab to synchrotron. It is expressed as an XML schema, with the intention of matching up to the data in the IUCr CIF dictionary.
- DNA: this has an exchange/communication model which is predominantly for internal use; ultimately the outputs of DNA are images and a logfile (it is unclear whether there are also DNA tables in the ISPyB database). For the purposes of deposition he felt that it would be useful for DNA to "explain" why it collected the data that it, and he would also like to have a framework to be able to describe the data processing.

AN commented that he would also like to be able to collect the data required for PDB REMARK 200 that is generated by DNA.

- XIA2: this is a "ground-up" rewrite of XIA and is based on a data model from the MTZ header hierarchy. XIA is intended to be an expert system for data processing, data reduction and structure solution.

Key features include tracking data flow and versioning of every datum used. Also it uses keywords rather than XML or CIF to output information (as this is more easily understood by humans), although it could also output XML at the end.

AN asked who would use XIA2 – GW responded that it would be included in CCP4 release 6.1 and would also tie into the CCP4i tracking database (there have already been a number of discussions about this).

GW also suggested that since XIA2 has a suite of wrappers for programs which translate logfiles into XML, this could form the basis of a library to be incorporated into CCP4.

3.5 Charles Ballard: CCP4 Automation

CCB talked briefly about the status of the CCP4 Automation project, which uses XML formatted input which is otherwise similar to XIA2. He also talked

about the automation Python library, which consists of a set of wrappers and drivers.

3.6 Santosh Panjkar: Autorickshaw

SP gave a detailed description of the Autorickshaw pipeline being used at EMBL Hamburg. The primary aim at present is to produce an interpretable electron density map and partial structure in minimal time in order to confirm the success of the diffraction experiment while the user is still at the beamline.

The approach is to “use everything that is available” (i.e. all possible programs); try to mimic what an experienced crystallographer would do; try to do everything as fast as possible i.e. be just as good as necessary; and try to minimise user input (in practice this is of the order of: method selection, number of residues per subunit, expected number of heavy atoms, number of subunits, PDB file if doing MR, and the spacegroup).

Various protocols are available which result in there being 36 pathways through the system per phasing method, so a next step would be to try and characterise the pathways according to the attributes of the input data. A number of critical decision points have been identified, for example: what resolution cut-off should be applied for substructure determination and phasing? When should SHELXD be terminated?

The output summary from Autorickshaw gives key data extracted from each stage (for example, the best SHELXD trial). It includes a human-readable summary of the process that Autorickshaw went through. There is also a final output tarball that contains the MTZ, PDB and map files.

MDW suggested that harvesting files could also be extracted from Autorickshaw and SP agreed this should be possible.

SP is interested in adding data to a database and using “proper” XML tagging in order to improve decision-making. His plan is that for each program:

- Produce XML output
- Combine this into a master XML file
- Put this into a database
- Use the database for data mining

At the moment Autorickshaw uses a single small-ish XML file however SP would like to expand this. MDW & SP were actioned to discuss XML-ising certain CCP4 programs in a breakout session.

SP also talked about the deposition of test datasets at Hamburg for the purposes of testing software (a “data depository”) – this creates an XML file with data items (submitted and calculated data). GW suggested that there should be a central repository for synchrotron data.

- **Action:** AN requested that SP add the beamline information to the output of Autorickshaw

A “Autorickshaw and MR_BUMP” breakout group was convened to discuss the issues of integrating these two applications.

3.7 Alun Ashton: Status and Requirements for DIAMOND

AWA gave a brief overview of the status of facilities planned for DIAMOND:

- Users will have a single login for all facilities and for the SRB (storage resource broker, a Grid-enabled data repository). He noted that there are data protection issues with this.
- Users will have access to DIAMOND computing resources for a few days before and after their visit
- DIAMOND will use NEXUS to store data in the SRB system (aside: NEXUS is a neutron and synchrotron data format, see e.g. http://www.nexus.anl.gov/nexus_intro.html)

AN suggested that the data stored in the NEXUS header could feed into REMARK 200. (Only metadata will be stored indefinitely in the SRB system, the raw data is presumably considered too large to store for a long time period.)

Some discussion of REMARK 200 followed. GW identified some data items that are not in the diffraction image headers but which are needed for REMARK 200. Some are held in ISPyB, others could come from harvest files plus the tracking system. (The status and degree of ubiquity of ISPyB seemed to be ambiguous – essentially there is no uniform database system).

AWA also arranged a brief discussion with one of the e-science database/data portal developers (Shoab Sufi), who gave an overview of the system being designed for data storage at ISIS and DIAMOND – see for example <http://www.e-science.clrc.ac.uk/web/groups/Data-Management/>.

4 Breakout group discussions

As a result of the presentations and discussions a number of breakout groups formed to discuss the issues raised in more detail. These interactions are summarised below.

4.1 DNA and Deposition Breakout Group (AN, GW, AWA)

This group formed to discuss how to get the PDB REMARK 200 data from DNA. AWA noted that the underlying programs [used in DNA] can already produce harvesting files, and that the other REMARK 200 data items are really synchrotron-specific and may not be known to DNA – essentially AWA does not want DNA to be responsible for producing the “raw” REMARK 200 data.

As a compromise two actions were suggested:

- **Action:** GW will ensure that XIA will produce the harvesting data required for DNA

- **Action:** AWA will investigate harvesting in DNA as a “proof of concept”.

4.2 Deposition and Tracking Breakout Group (AN, PJB, WY)

The aim is to produce a summary of the final process, plus a set of intermediate files. This will be done by producing a standard format (XML) describing a sequence of events, plus the resulting files for each step. The data will be taken from the tracking database, with a method provided for supplementing the data or providing missing items (e.g. manual entry). A minimal approach should be taken to produce the initial schema, as this can be grown later.

- **Action:** PJB/WY and GW to produce a list of data items (name-value pairs) that are produced from CCP4i. AN will turn this into a set of XML tags. PJB will then implement the code to produce this.

Once the XML is created a viewer could be created e.g. using XSLT plus a web browser.

- **Action:** AN to provide the REMARK 200 data items and look at ways to get the information to populate them e.g. from ISPyB or image headers.

4.3 Autorickshaw and MR_BUMP Breakout Group (MDW, SP, RMK)

MR_BUMP will slot into Autorickshaw, although MR_BUMP will require a slight change of emphasis to fit into the Autorickshaw philosophy of speed versus exhaustively following all possibilities. Autorickshaw could also use trial models from MR_BUMP as input.

- **Action:** SP to investigate using MR_BUMP in Autorickshaw; MDW/RMK to investigate producing XML output from MR_BUMP to facilitate integration with Autorickshaw
- **Action:** MDW to look at adding XML output to specific CCP4 programs as requested by SP

4.4 Standardised XML Tags Breakout Group (MDW, AN, GW)

This group looked at the XML tags already incorporated into the CCP4 program MATTHEWS_COEFF, to see what would need to be done to fit this into the BioXHIT schema (mark-up of input and output data and errors), and used this as an example of what would be required generally.

Optionally, references and errors could be tagged – there was a suggestion that the Fortran programs could output the citations and errors (2nd axis on AN’s proposed schema) but a wrapper would have to deal with the 1st axis (corresponding to the life-cycle).

- **Action:** AN to send MDW an updated schema which MDW will then use to update 9 CCP4 programs (plus another 5 requested to have mark-up from SP)
- **Action:** AN to send CCB and PJB the schema, in order to mark-up other programs (e.g. MTZ2CIF).

5 Additional outcomes and actions

The actions from the meeting are highlighted in the text above. Progress is currently being made on the XML schema for deposition from the tracking database.

In addition an initial version of a Python module for extracting program names from CCP4 and CCP4i log files (“smartie.py”) has been created by PJB, in order to enable a more complete description of the project history to be generated at deposition time.

Peter Briggs 30th November 2006