ISPyB Technical Workshop

SOLEIL September 12-13 2019

Participants:

ALBA: Daniel Sanchez

Diamond: James Hall, Karl Levik, Neil Smith

ESRF: Maxime Chaillet, Solange Delageniere, Alejandro de Maria, Olof Svensson

Diamond/ESRF: Stuart Fisher

Global Phasing: Rasmus Fogh

SOLEIL: Idrissou Chado (part time)

Aim

The ISPyB developers group hold monthly video conferences (VCs) to discuss data model and software development activities. Changes to the database schema are initially proposed and captured on the database modelling github repository https://github.com/ispyb/ispyb-database-modeling. While small data model changes can be agreed during the VCs, larger changes need more detailed discussion. The aim of this Technical Workshop was to discuss and agree refactoring of a number of database tables agreed at previous VCs.

The topics for discussion included:

- Screening Tables (Issue #46)
- Data Collections (Issue #45)
- Sample tables (Protein, Crystal Issue #42)

The agenda was agreed at the start of the meeting to include discussion on auto-processing and a comparison between DLS and ESRF processes.

Screening Tables (Issue #46)

The Screening Tables, name notwithstanding, cover characterisation and strategy determination, including specification of data collection strategy. The tables (and their names) date back to the days of CCD detectors and the DNA project, where it was contemplated to systematically screen crystals to decide which crystals and strategies to use for later data acquisition. The tables have not

been modified since 2012 (except for maybe a couple of column additions), nor has the user interface. EDNA populates these tables but it uses in-memory values and does not read the data back from ISPyB. GDA allows the choice between different characterisation strategies, whereas ESRF (EDNA) produces only a single characterisation strategy in each case. The original purpose of these tables was to calculate alternative strategies for users to choose between, but this facility is not currently used.

From the detailed discussion it was clear that people **do** use strategies with multiple subwedges.

Decisions

It is agreed to follow the refactoring proposal:

- Merge Screening, ScreeningOutput and Strategy tables, removing some fields.
- Add an optional autoprocProgramID column.
- Remove ScreeningRank and ScreeningRankSet, which are no longer needed, taking care to avoid breaking the old ISPyB interface, which is the only program still using these tables.
- Double-check which columns are actually used.
- Keep tables Screening, OutputLattice, Wedge and Subwedge.
- Rename tables to use prefix Strategy instead of Screening.

It is expected that these changes can be done with relatively limited resources.

A detailed proposal should be presented for approval at the Berlin meeting. Global Phasing does not plan to start populating these tables, as its Workflow stores all this information (and more) in its Persistence Layer, so that the expected benefit would be quite limited.

Data Collections (Issue #45)

The basic proposition of unifying DataCollection, XFEFluorescenceSpectrum, and EnergyScan was agreed. It was agreed that a number of columns that are unique to EnergyScan are processing results and thus rightly belong elsewhere, simplifying the work of merging the tables.

There was much discussion on the choice between having a wide table for all data collection types, with many nulls, having a ‘subtype’ table for each technique with the single-technique columns, or some kind of mixture. Either approach would work from a database-technical point of view, and neither alternative was either chosen or excluded.

One advantage of combining the three tables is to make it easier to list all activities, of whatever type, happening on a synchrotron. It was discussed, but not agreed, whether to add a RobotAction to the set of supported ‘collection’ type, in order to accommodate activities such as sample changing, beam calibration etc. that do not collect data. At this point the (topmost?) table would be a kind of ‘BeamlineActivity’ with only timing and scheduling columns.

EM data are currently stored in the DataCollection table, even though EM has quite a few technique-specific columns and little overlap with other techniques. A wider reorganisation, if decided on, might include splitting EM into a separate (sub) table with a foreign key to DataCollection for common data (timestamps etc.).

Decisions

The proposal by KL (#Issue 45) should be implemented, and consideration of further changes should be postponed. Columns should be merged between the three techniques, with names harmonised and clarified, duplicate information (like both energy and wavelength) removed, and units decided and clearly documented.

A detailed proposal should be presented for approval at the Berlin meeting, with Karl Levik (**ACTION**) in charge of this.

Sample Tables

This point triggered a discussion on differences in approaches and priorities between ESRF and Diamond. During this discussion it became clear that DLS and ESRF have significantly different strategic objectives for the use of ISPyB at their respective facilities, and that this raised important issues particularly for any refactoring of the Sample tables. The discussion is summarised in greater detail in section “Diverging Priorities” below.

Notwithstanding the above, there is still interest in exploring whether there is a feasible way to harmonise and expand the Sample tables to cover multiple techniques. The most discussed case is for SAXS, a technique supported at both Diamond and ESRF, which currently has its own separate set of tables. One issue that needs addressing is the distinction between sample components that are a) the main component (protein) under investigation, b) individual molecules as additives, c) more or less well-defined mixtures (such a specific batch of buffer solution or fetal calf serum). Another problem specific to SAXS is tracking the individual subcomponents of a (main) component that is a complex. It was agreed that there is a need for concrete examples and use cases.

Decisions

Stuart and Karl (Alejandro will be keen to provide information and use cases) will gather example cases and study how the MX and SAXS sample descriptions could be mapped (**ACTION**). The team from Diamond will further gather examples of data stored for the various techniques and study the possibility of adapting the sample description to support additional techniques with a minimum of disruptive changes.

ESRF’s viewpoint is that implementing these changes in the official ISPyB software would require a lot of work for no expected gain.

Autoprocessing and pipelines (including Issue #32, phasing tables)

Diamond (Synchweb) has a facility for user-initiated reprocessing, which allows one to reprocess multi-sweep data sets. Synchweb also allows the user to select which image ranges to process. ESRF is developing software that will allow reprocessing (i.e. the re-running of automatic pipelines with different parameters) and offline data analysis (i.e. running any tool on any set of data stored on ISPyB).

The organisation differs between the two sites, whether by DataCollectionGroup (Synchweb), or a list of DataCollections (EXI). For Synchweb results are displayed under only one DataCollection, but you can access the data for multi-sweep processing; whereas for EXI the results are attached to each DataCollection of the list. Diamond, unlike ESRF, stores the results in the same tables as auto-processing results. ESRF does not offer any kind of reprocessing today but it is under development.

No actions were agreed – and the issue is complicated by the use of permanent data storage outside ISPyB by ESRF - but it was agreed by the ESRF that it would be desirable to store data (also) within ISPyB if results fit conveniently into the data model structure of ISPyB.

The original proposal (Issue #32) was to remove the PhasingProgramRun and PhasingProgramAttachment tables, replacing them by AutoProcProgram and AutoProcProgramAttachment. This was agreed. **ACTION**: Stuart

There is agreement in principle on going further with merging tables, including the ProcessingJob table, and to consider changes so that other techniques (EM, SAXS) could also use the more generic processing run tables that would be the result. ESRF. thinks that the current implementation of ProcessingJob might be too MX-centric and might not scale well (example: results are stored in a single DataCollection even if there is a list of DCs). This would require someone to make a future, detailed proposal.

#35 Data collection group into grid info

This had already been done and so requires no further action.

#22 sampleId in DataCollection or DCG

There was a long discussion whether the blSampleID rightly belonged in DataCollectionGroup (its current location) or DataCollection (where it used to be located until a year or two ago, and where DLS seems to have moved it back to). There were coherent arguments that for some techniques there might be a need to group experiments that were collected on different samples, but it was decided to leave the tables unchanged, with the blSampleID in the DataCollectionGroup, and reconsider when a clear need arose in a use case.

Diverging priorities

It was clear from the discussions that there are different aspirations between ESRF and Diamond, the two main developing centres. Differences in aims and viewpoints between the respective approaches arise in several contexts. They can be summarised as follows:

Diamond aims to use ISPyB as the unique Laboratory Information Management System (LIMS) for all beamlines at the facility, and therefore needs to expand ISPyB to cover all data acquisition techniques in use at DLS, as well as a wider range of sample types. This requires expansion and harmonisation of the data model to support additional techniques and refactoring to make the tables and column names appropriate for the wider range of techniques. ESRF, on the other hand, intends to use ISPyB only for MX, SAXS and EM and therefore sees no future utility in these modifications – which means that for ESRF, modification and refactoring beyond these disciplines require substantial work for no expected gain.

The DLS approach is that all processing conducted within the facility (automatic or triggered by users) should be recorded in ISPyB. Processing outputs can be registered as processing attachments (logs, data files, charts etc.) and then viewed by users through interfaces such as SynchWeb. The DLS view is therefore that ISPyB in its current form is suitable for both auto-processing and "offline" processing triggered by users while a session is active. There may be some minor changes required to link processing to data collection groups rather than data collections (as is currently the case).

Unlike Diamond, ESRF does not use ISPyB as the sole data source for (re-)processing but also uses long-term persistent JSON files and object-oriented data bases (Mongo-DB?) as an authoritative storage mechanism for re-processed data. Since ESRF wishes to store more kinds of processing results than currently supported by the ISPyB tables, this creates an alternative and possibly competing data storage mechanism. The work at ESRF is still at an early stage, with rapid changes being made, so it is not yet possible to say how these efforts will eventually pan out.

If pursued in their current directions, there is a risk that these two lines of development might make it impossible for third parties to access (re-)processed data across all participating synchrotrons in a uniform manner.

Since ESRF and Diamond use different user interfaces, front-ends, back-ends and software technology, and only (most of) the actual database structure is shared, there is limited scope for pooling resources, and any major changes to the database will require large amounts of duplicated effort. Unfortunately, any move towards sharing more code would equally require large resources, not to mention some fundamental changes of approach. These problems are serious enough to put the long-term continuation of the collaboration in question – or at least in serious need of much more coordinated forward planning. The developers recommend that the Steering Committee should consider what to do about these issues as a matter of urgency.

Reporting at ISPyB meeting

- All changes agreed here will be presented at the Berlin meeting as detailed proposals with examples by the person who wrote the original issue proposal. They will be added to the Github issues before then to gather comments and objections.
- Actions to gather information or examples should be reported to the developers at the Berlin meeting and discussed.

AOB

- The tables used for Workflow were discussed. Workflow is linked to one or more DataCollectionGroups and consists of a series of steps. The important information consists mainly of pointers to log files, images, and result files generally. Information is stored in JSON, with defined schemas, and rendered in html using DustJS templates. Global Phasing should consider populating these tables to make the GPhL workflow results viewable.
- The data base schemas in use at Diamond and ESRF were merged in 2017, and all changes since then are tracked and agreed.
- Processing is displayed (at ESRF) by DataCollectionGroup. Multi-sweep and multi-crystal data collections could be handled by combining them into a single DataCollectionGroup, under workflow control. The Workflow tables should be able to hold enough information to describe what is going on. This may need some additional work. Could it be used for e.g. Mesh-and-Collect as well? Or even for SSX?
- It was agreed that the meeting had been very fruitful, and that similar developers' meetings should be considered in the future.

Next meeting

Web meeting Monday October 7th, 1400 UK time on whereby.com/ispyb.