The BIOXHIT Project

www.bioxhit.org

Coordinated by EMBL Hamburg

2nd BIOXHIT Annual Meeting
18th - 19th January 2006
ESRF, Grenoble, France

Progress in WP 5.2: BioXHIT Data Management for PX Structure Determination Software Pipelines

Peter Briggs and Wanjuan (Wendy) Yang, CCLRC Daresbury Laboratory, Warrington WA4 4AD, UK

Introduction

A key part of the integrated technology platform being delivered by the BIOXHIT project is the development of automated structure determination software pipelines that cover the post-data collection stages of structure solution by protein X-ray crystallography (PX). These pipelines need to accurately record and track the data that they produce, both for their own operation and for final deposition of the determined structures.

BIOXHIT Workpackage 5.2 is principally concerned with developing a system for performing this data management in order to address the needs of software pipelines. This poster reports on the progress that has been made by CCP4 (BIOXHIT Partner 10) towards this end in the last year.

Project Contributors

Workpackage 5.2 is co-ordinated by the **Collaborative Computational Project No4** (CCP4), which provides a software suite for macromolecular structure determination by X-ray crystallography. Funding for the work described here has been provided by the European Commission as part of the BIOXHIT project.

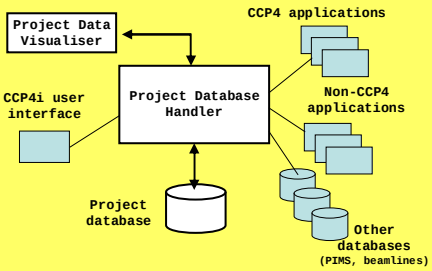
The handler and database components are being developed with contributions from the developers of the CCP4 Automation (HAPPy), XIA/e-HTPX and CRANK Projects; discussions have also taken place with the PIMS project and the MX beamlines at DIAMOND. The scope of collaboration will be broadened as part of the development of the version 1 database schema described below.

Project tracking system for the structure solution software pipeline

Components of the system

- Project database handler
- Database for Project Data & Tracking
- Visualisation tools

These components and their relationships are shown schematically in the figure (right), and are described in more detail in the sections below.

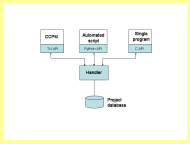


Key considerations

- Implement a system for both manual and automated structure determination
- Allow multiple database back-ends
- Gather as much information from client programs as possible automatically
- Open architecture accommodating heterogenous software components

Project Database Handler

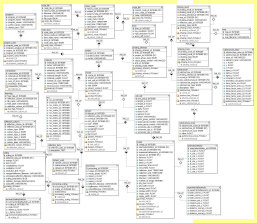
The Project Database Handler is a brokering application that mediates interactions between the project database and the external applications and databases (local or remote). It acts as a single point of access to the data for external applications and hides the implementation of the database from them.



Applications talk to the handler via a "client API" which will be implemented in different programming languages (see left). Communications between the handler and the API are encoded in XML.

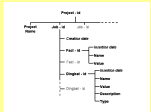
Version 1 of the handler was developed in Python and acted as a testbed used with the XIA and HAPPy pipelines. The code can be obtained from ftp://ftp.ccp4.ac.uk/pjx/bioxhit/

Work is also ongoing to integrate the handler into CCP4i (the CCP4 graphical user interface).



Database for Project Data & Tracking

A database is being designed and implemented which will be capable of storing both project data (the information used by each step in a pipeline) and project history (the steps taken and the provenance and evolution of information as the project progresses).



Version 1 of the handler was developed using a "minimal database" (see right) implemented in MySQL, to explore the requirements of the full system.

As a result of this work a more detailed database architecture is envisaged, with three components:

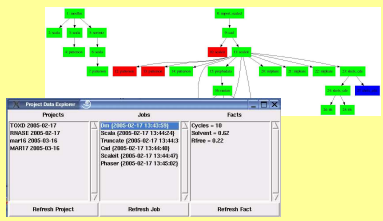
- **Knowledge base:** consisting of the common crystallographic data items used in the software pipeline that are shared between different applications. This will link to external databases (e.g. PIMS and beamlines) as well as providing data for deposition, and will be consistent with the standards in WP5.1.
- **Operational database:** containing application-specific data and representations (for example parameter files or Python objects) that are not intended to be shared between applications.
- **Tracking database:** storing the history of the data generation in the knowledge and operational databases.

Work is now ongoing to develop a MySQL schema using DBDesigner 4 for the knowledge base and tracking databases (see left).

Visualisation Tools

These tools will provide interfaces to the database, to display the project data in selective views and thus focus on particular aspects of the data-flow or logical flow – for example, as work flow diagrams.

Prototype tools (below) have been developed based on the existing CCP4i project database, and a "project data explorer" based on the prototype minimal database schema.



Next steps

Over the next 12 months the aims are to:

- Improve the Python handler's functionality and robustness and integrate it into CCP4i
- Specify and implement the full knowledge base and project tracking databases with input from other BIOXHIT Partners
- Develop and release the version 1 visualiser application.

See http://www.ccp4.ac.uk/projects/bioxhit.html for more information