

Towards Data Management for PX Structure Determination Within CCP4

Peter J Briggs and Wanjuan (Wendy) Yang

Computational Science and Engineering Department, CCLRC Daresbury Laboratory, Warrington WA4 4AD, UK

Introduction

BIOXHIT [1] is an Integrated Project funded within the 6th Framework Programme of the European Commission, and is coordinating scientists at European synchrotrons along with leading software developers with the aim of consolidating and automating the process of macromolecular structure determination using X-ray protein crystallography (PX), from crystallisation to deposition.

A key part of the project is the development of automated structure determination software “pipelines” that cover the post-data collection stages of the 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.

This poster reports work that CCP4 [2] is undertaking within the BIOXHIT project to develop a data management system that address the needs both of automated software pipelines, and manual structure determinations.

Project Aims and Contributors

The Collaborative Computational Project No 4 (CCP4) is a UK-based software initiative which provides a suite of programs for macromolecular structure determination by PX. Currently CCP4 offers basic data management within its graphical user interface system CCP4i [3], which records information such as date, status input parameters and files associated with each run of a particular task, and through technologies such as Data Harvesting [5].

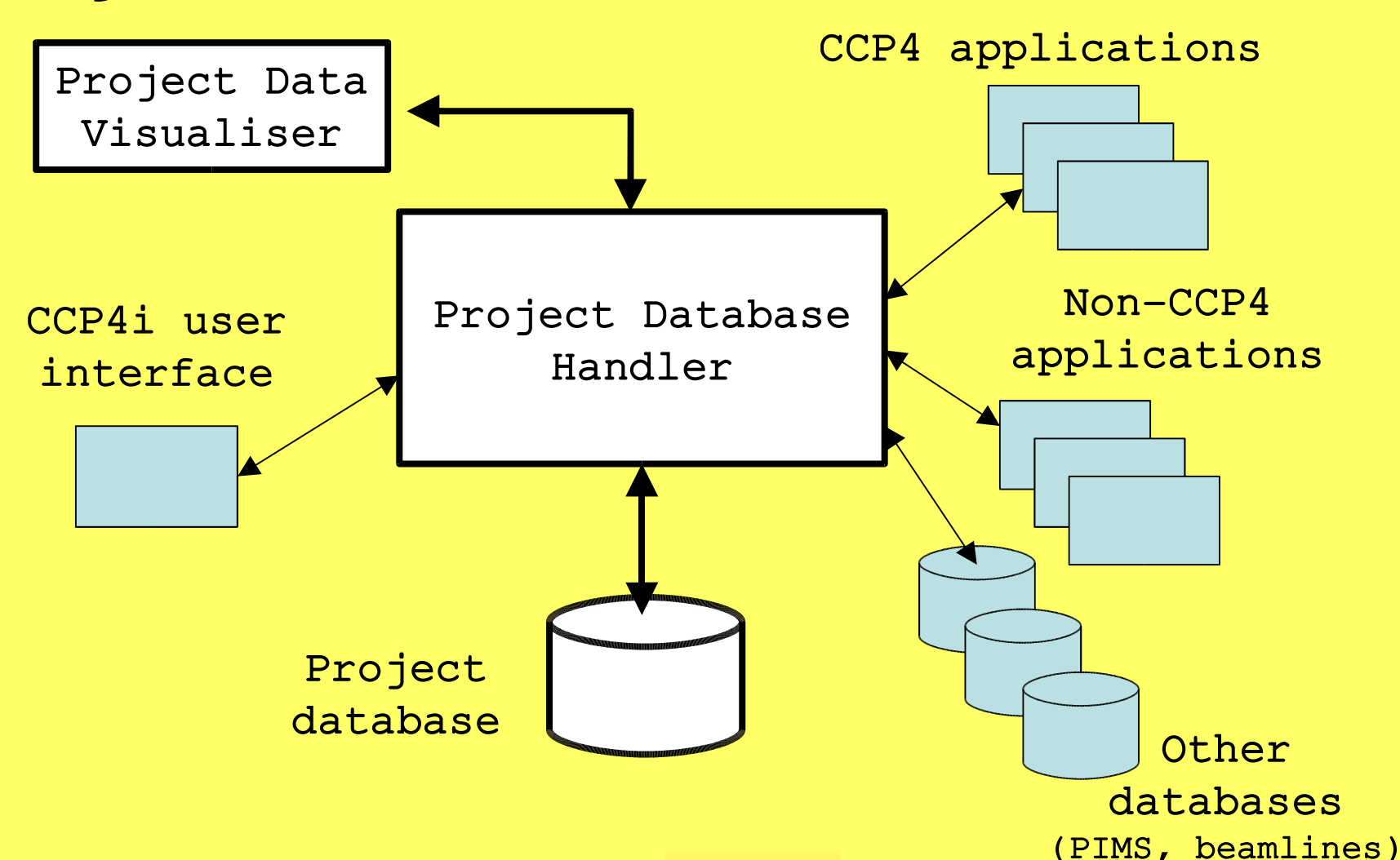
The proposed data management system builds on and extends this existing functionality, aiming to provide a rich database which is easily accessible to a variety of different systems, plus a set of tools to visualise the project history and other aspects of the data. The components are being developed with contributions from the developers of the CCP4 Automation (HAPPy) [6] and XIA [7] Projects; discussions have also taken place with the PiMS project [8] and beamline scientists at the new UK synchrotron DIAMOND [9].

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.

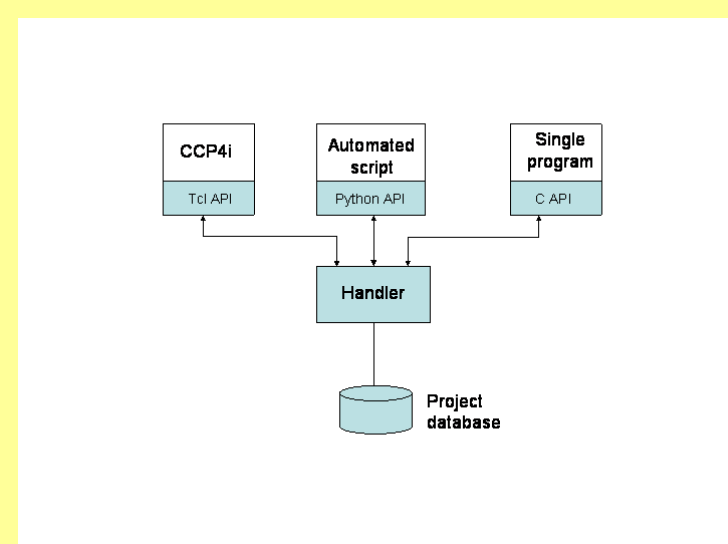


Key considerations

- 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 heterogeneous 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 library, which is implemented in different programming languages (left).

Communications between the handler and the API are encoded in XML.

The handler is written in Python and currently supports two embedded databases (CCP4i and SQLite). A version of CCP4i is under development which uses the handler via a Tcl client API; a Python client API will be developed to support other programs such as CCP4mg [9] and Coot [10].

Current Status

The current focus is on integrating the handler into CCP4i using the existing database backend, and on extending this to other software within CCP4 such as CCP4mg and Coot. After this the focus will shift to developing the visualisation tools and the database schema, to incorporate into automated pipelines like HAPPy and XIA.

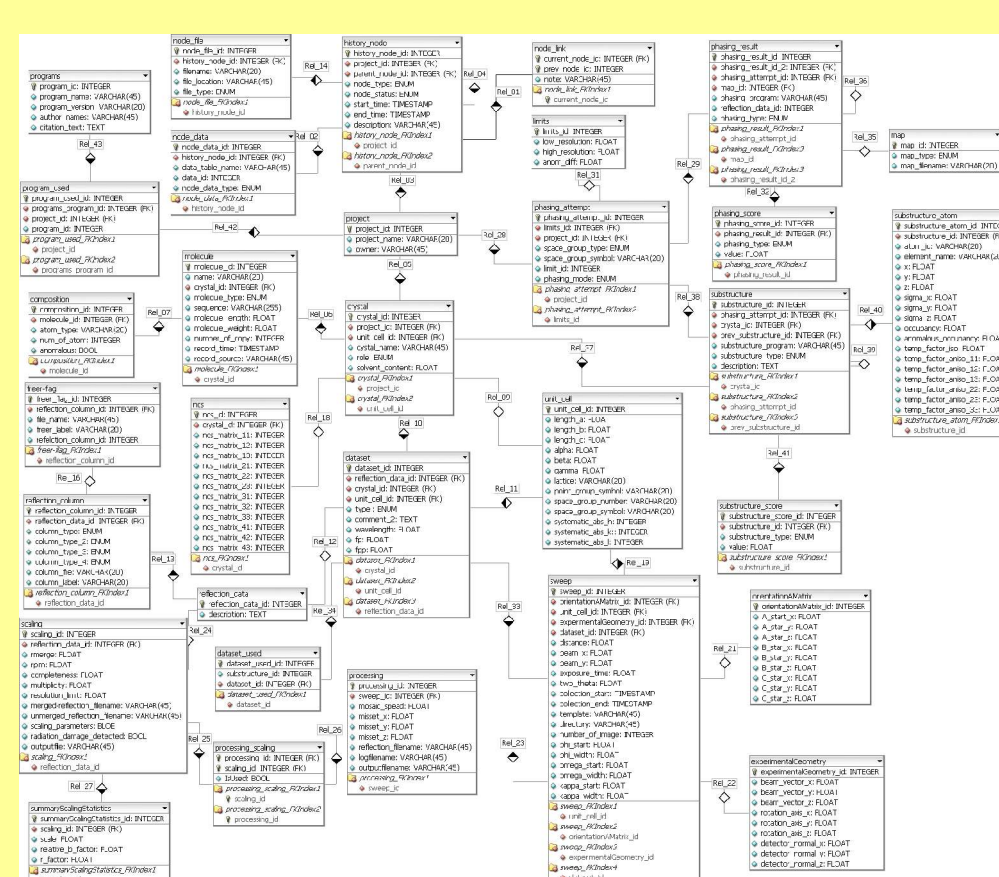
For more information about the project see <http://www.ccp4.ac.uk/projects/bioxhit.html>

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).

Currently there are two database implementations: one supporting the existing simple CCP4i database, and another using SQLite to implement an extended database with three conceptual 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.
- **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.

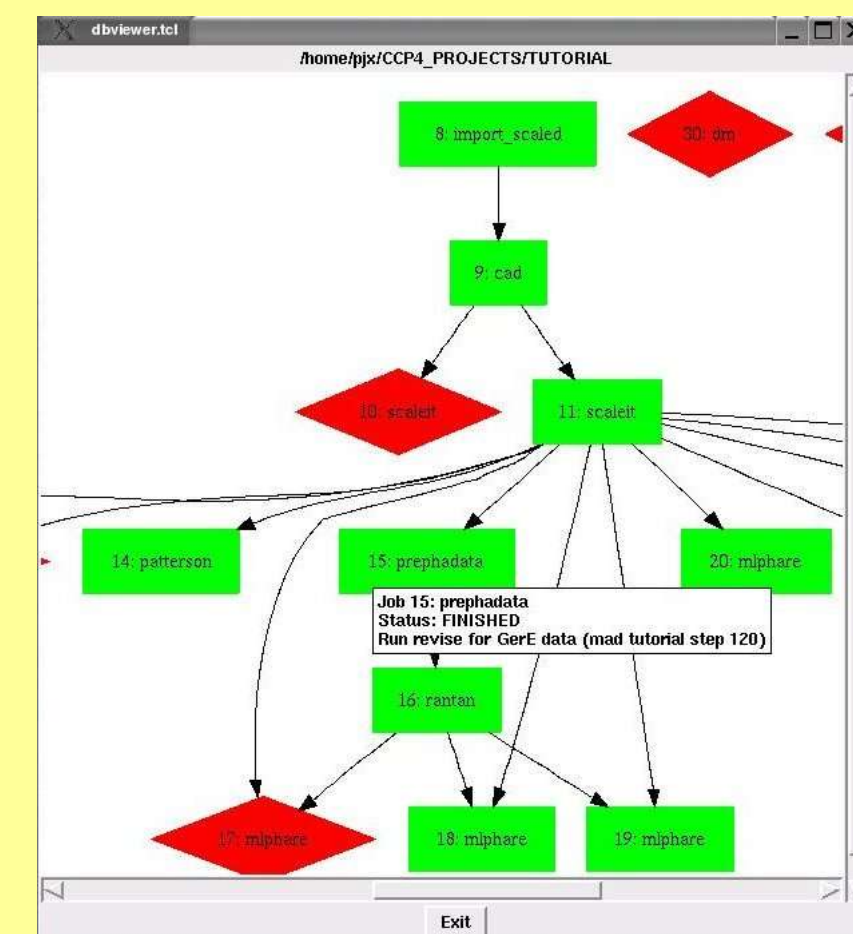


The knowledge base and tracking databases are currently being developed as SQL schema using DBDesigner (left), with the aim of making a first version available before the end of the year.

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.

Prototype tools based on the Graphviz [11] package (right) have been used to explore project history within the existing CCP4i project database. More sophisticated visualisation tools are envisaged for the extended database later on in the project.



References

- [1] BIOXHIT “Biocrystallography (X) on a Highly Integrated Technology Platform for Structural Genomics: <http://www.bioxhit.org/>
- [2] CCP4 <http://www.ccp4.ac.uk>
- [3] CCP4i: Potterton et al, *Acta Cryst* **D59** 1131-1137 (2003)
- [4] Data Harvesting: Winn, *CCP4 Newsletter* **37** (October 1999)
- [5] HAPPy: <http://www.ccp4.ac.uk/HAPPy>
- [6] XIA: <http://www.ccp4.ac.uk/xia>
- [7] PiMS: Protein Information Management System <http://www.pims-lims.org/>
- [8] DIAMOND: <http://www.diamond.ac.uk/>
- [9] CCP4mg: CCP4 molecular graphics <http://www.ysbl.york.ac.uk/~ccp4mg/>
- [10] Coot: semi-automated model completion and validation <http://www.ysbl.york.ac.uk/~emsley/cool/>
- [11] Graphviz: graph visualisation <http://www.graphviz.org/>

Acknowledgements

CCP4 is funded by the BBSRC; PB is funded by CCLRC from CCP4 industrial income, and from the BIOXHIT project; WY is funded from the BIOXHIT project. BIOXHIT is funded by the European Commission via its 6th Framework Programme, under the thematic area “Life Sciences, genomics and biotechnology for health”, contract number LHS-G-CT-2003-503420.

