WorkPackage 5.2: Implementation of Data management and Project Tracking in Structure Solution

Peter Briggs, CCP4

Introduction

• CCP4 (Partner 10) is a UK-based software initiative with core funding from BBSRC plus income from commercial receipts

• CCP4 distributes a software suite for macromolecular structure determination by X-ray crystallography

• Consists of nearly 200 programs plus core libraries and a graphical interface system "**CCP4i**"

Task 5.2.1: Implementation of Data Management and Project Tracking in Structure Solution

Aim:

- *To fill the need for project tracking within the BIOXHIT structure solution software pipeline.*

Partners involved:

- *Partners 1C (EBI), 7 (ELETTRA), 10 (CCP4)*

Why do we need to track data and project history?

Users running manual structure solutions

- benefit from automatic organisation and tracking of data
- can readily locate relevant data when needed
- prevents mistakes
- possible to review progress and determine next steps
- recognise failure points and improve procedures in future

Automated software procedures have similar requirements

- BIOXHIT software pipeline automation (Section 4)
- CCP4 Software Automation Project (starting soon)
- Synchrotron automation efforts e.g. at the SRS Daresbury

Currently:

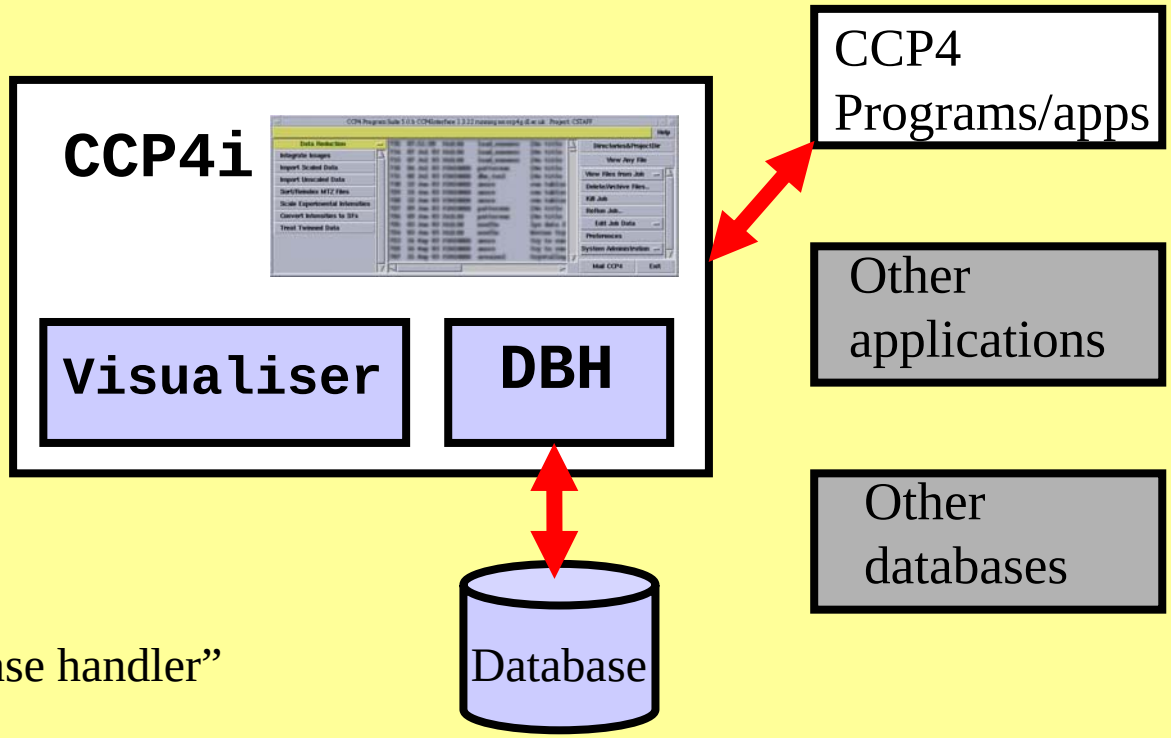
CCP4i provides an interface to manually running programs

- Basic project history database for each “project”
- Visualisation of project history as a simple list of jobs
- Starting point for data management within CCP4

Limitations of the database:

- Only accessible from within **CCP4i** system
- Cannot be accessed by multiple users/processes or remotely
- Scope of data stored is very limited
- Basic flat-file implementation

Current CCP4i model:



DBH="database handler"

Structure determination will most likely not be performed exclusively within a single software package or at a single site

Other applications:

- BIOXHIT Partners
- CCP4 automation
- DNA/e-HTPX spin-offs

Other databases:

- LIMS (e.g. MOLE, HALX)
- Facility databases (at the synchrotron)

Aside: MOLE (Mining Organising Logging Experimental Data)

- LIMS being developed by Alun Ashton at Daresbury
- Based on e-HTPX protein production data model
- See `http://www.mole.ac.uk/`

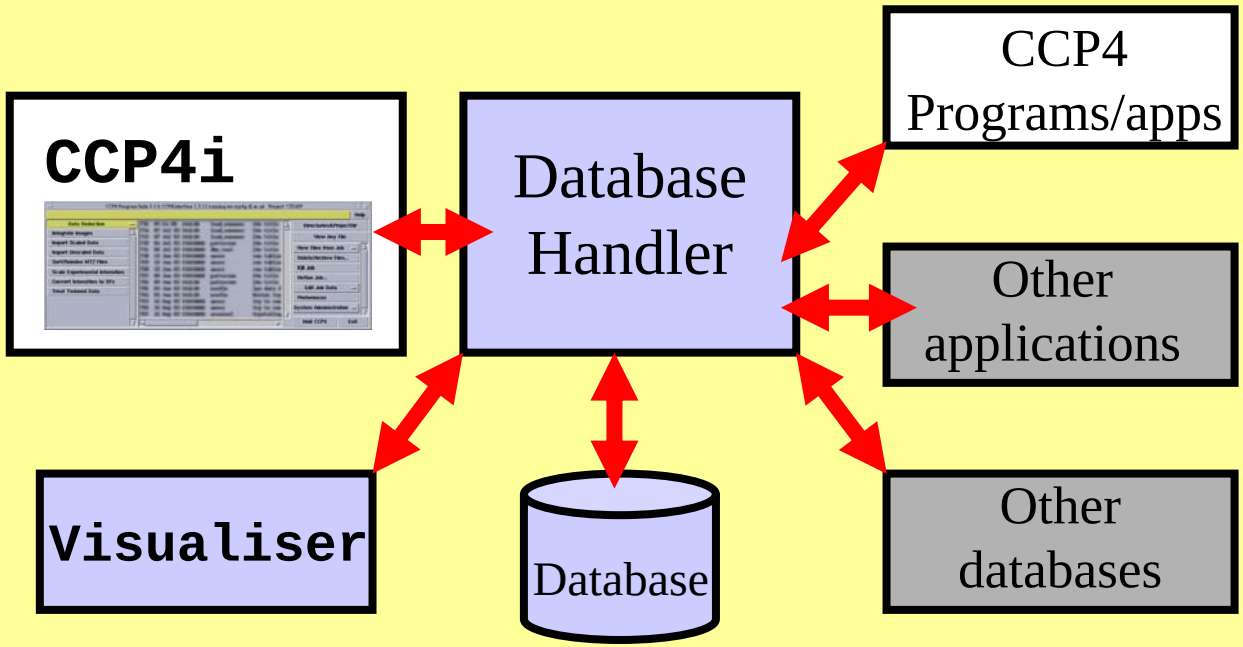
What we would like to be able to do

- Access the database (read & write) from other applications
- Talk to other databases
- Allow remote access from multiple processes

- Store enough data to enable tracking:
  - Project history – see which steps are related
  - Data history – see where data came from
- Provide access to other project-specific data

- Provide more powerful query functionality
- Provide advanced tools to visualise project and data history

New architecture



Database handler

- Server application
    - need to address security and authentication issues
- Mediates interactions between database and other applications
- Interactions via standard data exchange format (XML)
    - use standards agreed within WP 5.1
- Built on top of **`CCP4i`** but independent of it
- Deliverable 5.2.1

Database for Project and Data Tracking

Content: expand scope of data stored

- Store “project-specific” data
- Extend the history record information content to store metadata (explicit connections between steps in procedure, decision points etc)
- Accommodate requirements of other Partners/projects
  - Conform to standards in task 5.1.2 for data models
  - Report on requirements: deliverable 5.2.2

Database for Project and Data Tracking

Implementation:

- Migrate from flat files to a relational database backend
- Consider different possibilities (e.g. mySQL, XML dbs …)
  - Issues: portability, ease of installation, large facility versus single user etc etc …
- Will be consistent with data models developed/adopted by BioXHIT (WP 5.1)

Visualisation Tools

- Interface to the database: provide selective views of data and logical flow which focus on particular aspects of the data

- Could be as simple as colour coding or as complicated as a network diagram

- Different representations facilitate understanding of the structure determination procedure
    - Important aid to reviewing output from automation

- Prototype visualisation tools: milestone Ms 5.2.2

WP Resources

- One full-time staff member working for duration of project
- Input from existing CCP4 staff

Dissemination

- Released through CCP4

Current status

- Developed prototype database handler to explore issues (socket communications, authentication etc)
- Currently recruiting (expect person in post by June 2004)

Summary

• Aim to address the need for project tracking in software pipeline within BIOXHIT

• Database handler application to mediate interactions with database

• Implementation of database for recording and tracking project data and history

• Visualisation tools to display & interact with data

Acknowledgements

- European Commission FP6 (BIOXHIT)
- BBSRC
- CCLRC Daresbury Laboratory

Links

CCP4 home page: **http://www.ccp4.ac.uk**

CCP4-BioXHIT: **http://www.ccp4.ac.uk/projects/bioxhit.html**