CCP4i Database Overview

Peter Briggs

Current Data Storage in CCP4i

- CCP4i Projects provide:
  - File storage (i.e. it's a directory)
  - Job database
  - Job data:
    - Parameter files for each job
    - Logfiles from each job
    - Notebook entry (annotation)
  - Amore MR model database
  - Experimental XML files

- CCP4i also stores list of projects and aliases that "belong" to each user

- A project is a directory + a set of "databases"

Mode of operation

- Single user mode
  - Each project owned by a single user
  - Each user runs a single instance of CCP4i
  - CCP4mg also uses CCP4i projects
  - Sharing of data between users is ad hoc
- CCP4i main process
  - Acts as visualiser (job list)
  - Provides an interface to manipulate job database
  - Spawns running jobs as independent processes
    - Running jobs also interact with job database (limited “write-only” operations)

Issues for CCP4i

- Speed of access to data
  - users request data (e.g. lists of files) in real-time
- Sharing projects/data between applications and users
  - multiple processes can write to the same job record
  - do multiple users want to access the same job database?
  - issues of access permissions
- Expansion of tracking information
  - concepts of “subjobs” and “subprojects”

Additional data storage in CCP4i – “project.def”

CCP4i could store common project data accessible to all tasks e.g. project.def file

- Initially populated by hand when project is started
- Updated from output of tasks
  - e.g. from XML files generated by programs
  - also updated by hand
- Tasks could query project.def to automatically populate fields

Possible data items for project.def

- Sequence data
- Molecular weight (theoretical and experimental)
- Experimental details:
    - Type of experiment (MAD, SAD, MIR etc)
    - Crystal identifiers
        - Associated cell information
        - Native or derivative
        - Heavy atom data (type, expected number of sites, coordinates, …)
        - Datasets derived from each crystal:
            - Wavelength, f’ and f’’ …
            - Pointer to MTZ columns with intensities/sf amplitudes
            - Pointer to Scalepack intensities (for SHELX)

Possible data items for project.def (continued)

- Derived quantities e.g.
  - non-crystallographic translation (aka pseudo-translation)
  - results from twinning analysis
  - solvent content
  - number of molecules in asu
- …

Questions are:

- primary: what data are useful for input into CCP4i tasks?
- secondary: what data are useful for input into other applications?