LHSG-CT-2003-503420

BioXHIT

A European integrated project to provide a highly effective technology platform for Structural Genomics.

Life Sciences, Genomics and Biotechnology for Health

WP5: De5.3.5 Report on the feasibility of an integrated connection between VCS and CCP4 Project Tracking Database

Due date of deliverable:	**31.12.2006**
Actual submission date:	**22.12.2006**

Start date of project: **1.1.2004** **Duration: 60 months**

Organisation name of lead contractor for this deliverable: EMBL-EBI
European Molecular Biology Laboratory - European Bioinformatics Institute,
Hinxton, Cambridge, UK. **Author** Peter Briggs

Section 5: Data Bases and Networking (managed by Partner 1C)
The tasks in Section 5 are "horizontal" in nature and cut across the tasks described in Sections 1 to 4. The main focus of this Section is on the management, transfer and access of data and of software components that are the results of the developments in those earlier sections. This is to be achieved through the setting of BIOXHIT standards for software and for data sharing, which will be used in implementations both in this Section and in others. Section 5 also contains an element of remote access to data and services provided by BIOXHIT. Remote access will also be facilitated by the setting and observance of standard protocols, for example by the use of emerging GRID technologies.

WP 5.3: Remote Access and Networking. Coordinator Roberto Pugliese (Partner 7), contributing Partners 1A, 1C, 3, 7, 10, 21.

This workpackage aims to increase the effectiveness and efficiency of experiments at the synchrotron beamlines by providing remote access and collaboratory tools.

De5.3.5 *Report on the feasibility of an integrated connection between VCS and the CCP4 Project Tracking Database*

Report on the feasibility of an integrated connection between VCS and the CCP4 Project Tracking Database: CCP4 perspective

CCP4-CCLRC (Partner 10) contribution to BioXHIT deliverable D 5.3.5

1 Introduction

This report outlines and assesses the possible connections between the Virtual Collaboratory System (VCS) developed by BioXHIT partner 7 (Elettra) and the project tracking database system (dbCCP4i) being developed by partner 10 (CCP4-CCLRC).

The content of the report is based on discussions between representatives of the two partners (Roberto Pugliese and Enrico Mariotti of Elettra, and Peter Briggs and Wanjuan Yang of CCP4) during a meeting held at Daresbury Laboratory UK on 12th May 2006, and followed up by email correspondence.

2 Background to VCS and dbCCP4i

2.1 What is VCS?

The Virtual Collaboratory System (VCS) is the basis of the Elettra Virtual Collaboratory (EVC) being used at the Elettra Synchrotron Light Laboratory in Trieste Italy. EVC is an example of a "collaborative virtual environment" – a software tool to support human-human and human-machine communication and collaboration. EVC is specifically intended for researchers collaborating on X-ray experiments at the Elettra synchrotron Light Laboratory, and allows individuals in geographically distinct locations to more easily share data, hardware and software resources as part of a collaborative research project.

The website for VCS is at https://ulisse.elettra.trieste.it/evc/home.do.

2.2 What is the CCP4 Project Tracking Database System?

The Collaborative Computational Project Number 4 (CCP4) is a UK-based software initiative that provides a suite of computer programs to facilitate determination of the three dimensional atomic structure of proteins and biological macromolecules from X-ray diffraction data. Part of the software suite is a graphical user interface system called CCP4i, which offers an integrated way of running many of the CCP4 programs. A key component of this system is a simple project tracking database that records details of program runs and associated data automatically as the researcher uses each program. This database is being developed into a separate resource (currently called dbCCP4i) that will be accessible to multiple software systems as well as CCP4i.

The CCP4 BioXHIT website is at http://www.ccp4.ac.uk/projects/bioxhit.html.

3 Relevant technical details of the systems

3.1 Details of VCS

VCS is built around the concepts of "stations" (each of which represent a resource or group of resources) and "projects" (which is an activity associated with a specific station).

Software that is web service enabled can be run through VCS via a web service interface. Software that does not have web service capabilities can still be run as a "legacy application" – essentially, similar to running the software on a remote computer using secure-shell forwarding. This is a sub-optimal method of running graphics-intensive software as it requires high bandwidth internet connections.

In the future it is planned that users of VCS will be able to organise all their projects on different stations into a single "superproject", to make managing their collaborations easier.

3.2 Details of dbCCP4i

The CCP4i project tracking database system dbCCP4i currently being developed within BioXHIT consists of a small server application (the "database handler") that controls access to data stored in the database. Communications between the handler and other software requiring access to the data ("client applications") are via pseudo-XML messages being passed through sockets. Application programming interfaces implemented in different programming languages ("client API libraries") are provided to simplify the database communications for developers of client applications.

The dbCCP4i system also incorporates the concept of a "project", however this is different from the VCS definition of the same term. In CCP4i, a project is a "container" for data files (experimental data and data derived from running various crystallographic applications). In practice a file system directory or folder is used to hold all the files relating to a single project, and each project has a single tracking database. It is left up to each researcher to decide what constitutes a CCP4i project (for example, it could be the determination of a complete structure from initial data to final coordinates, or just a small part of that process e.g. the molecular replacement component of the determination).

4 Possible connections between VCS and the CCP4 Tracking Database

4.1 Running CCP4i as a legacy application

Roberto Pugliese has already demonstrated running CCP4i via VCS, in "legacy application" mode. So this form of integration is trivial to achieve, however it is suboptimal for a number of reasons (for example it requires high bandwidth internet connections which may not always be available for remote users of VCS).

4.2 Running dbCCP4i on a VCS station with client applications on remote systems

As the database handler communicates with client applications via sockets, in principle it is possible for a client application to run on a different computer system from the one where the database handler is running.

In this scenario, a dbCCP4i server process would run on a VCS station and would allow client applications running on remote computers to access the database – for example a "visualiser" client (a program that displays the database contents to a user in different graphical views) could be run on a researcher's desktop computer in her lab, but would communicate with the dbCCP4i process on the VCS station in order to acquire the raw data to be displayed.

There are two key issues that have implications for this "half-way house" implementation:

- **Security issues:** this includes considerations such as authentication and authorisation, and the security of the communications e.g. encryption. The current implementation of dbCCP4i is not suitable for networked operation since there is no user authentication or authorisation, and no implementation of secure communications.
- **Data location:** in the current implementation, the tracking database would have to reside on the same VCS station that hosted the dbCCP4i server process. While the database stores references to resources such as data files, there is no provision for these resources

to be obtained directly from the server process. It is therefore unclear what the practical usage implications would be for running programs on a local machine whilst attempting to store tracking and other data on a remote system.

Implementation of a web services interface to dbCCP4i might overcome some of the security issues, and would make it much simpler to integrate into VCS. However this extension is not considered for practical implementation within the lifetime of the current project.

4.3 Decoupling of CCP4i graphical and non-graphical components

As developed the current CCP4i system does not have a clean separation of its graphical and non-graphical components. Crudely speaking, the graphical components are those parts of the system responsible for rendering user interfaces and the non-graphical components are responsible for the rest (running scripts and programs and managing the project tracking data).

Decoupling means re-implementing CCP4i in a way that the graphical elements are cleanly separated from the non-graphical parts. It would then be possible to run the graphical interface on a local machine which would communicate with “backend” processes (such as the underlying crystallographic software and the database server) that might be resident on remote machines (for example, a VCS station). It would essentially involve applying the technology used for implementing dbCCP4i to other parts of the CCP4i system (including the implementation of web service interfaces, if this avenue is pursued). Once the separation had been achieved it would be possible to create a distributed version of CCP4 and to generate multiple interfaces for CCP4i, including a web interface (an approach which might be best suited to integration with VCS).

This proposal is an extension of 4.2 above. It would involve a substantial change to the architecture of CCP4i and significant effort in rewriting and testing code. As such it is probably beyond the scope of the BioXHIT project.

5. Discussion and Conclusions

At a very basic level integration between CCP4i and VCS is possible, by running CCP4i and the CCP4 software as a “legacy application” within VCS, (option 4.1) and this has already been demonstrated to work. Whilst this approach is suboptimal, it does offer a baseline level of integration.

The next option is to provide some kind of web services interface specifically to the CCP4 project tracking database component dbCCP4i being developed within BioXHIT (option 4.2). Such an interface would allow more natural integration with VCS; however there are some issues which would need to be resolved. These issues include the lack of security mechanisms in the current implementation of dbCCP4i, and issues to do with data files potentially residing on different systems from the one holding the tracking database and from those where programs are run.

Some of the security issues may be addressed as part of an effort to provide a web services interface, however this has not been explored in any depth as yet. The issues of data file location raise more fundamental questions of how such a system (one in which only the project tracking database part ran as a web service on a VCS station) would be used, and whether it would provide genuine benefits to users of VCS. However it might provide an interesting pilot study for the final option 4.3. It might therefore be useful to consider these issues in more detail in a future project.

This option is the most ambitious, as it would involve a re-implementation of the CCP4i system in order to separate graphical and non-graphical components along similar lines to that already been done for the project tracking component (alternatively it should be a specification for any replacement system commissioned by CCP4). Achieving this is beyond the scope of BioXHIT, but could form the basis of a longer-term project.

Peter Briggs 30th November 2006