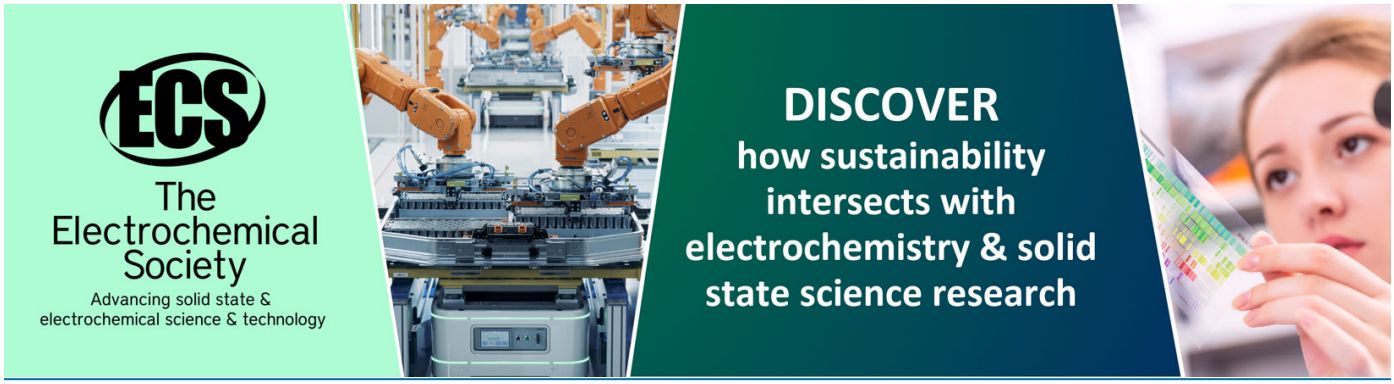
Data Management System at the Photon Factory Macromolecular Crystallography Beamline



Journal of Physics: Conference Series **425** (2013) 012017 doi:10.1088/1742-6596/425/1/012017

Data Management System at the Photon Factory Macromolecular Crystallography Beamline

Y Yamada[1], N Matsugaki, L M G Chavas, M Hiraki, N Igarashi and S Wakatsuki

Photon Factory, Institute of Materials Structure Sciences, High Energy Accelerator Research Organization, 1-1 Oho, Tsukuba, Ibaraki, Japan

E-mail: yusuke.yamada@kek.jp

Abstract. Macromolecular crystallography is a very powerful tool to investigate three-dimensional structures of macromolecules at the atomic level, and is widely spread among structural biology researchers. Due to recent upgrades of the macromolecular crystallography beamlines at the Photon Factory, beamline throughput has improved, allowing more experiments to be conducted during a user's beam time. Although the number of beamlines has increased, so has the number of beam time applications. Consequently, both the experimental data from users' experiments and data derived from beamline operations have dramatically increased, causing difficulties in organizing these diverse and large amounts of data for the beamline operation staff and users. To overcome this problem, we have developed a data management system by introducing commercial middleware, which consists of a controller, database, and web servers. We have prepared several database projects using this system. Each project is dedicated to a certain aspect such as experimental results, beam time applications, beam time schedule, or beamline operation reports. Then we designed a scheme to link all the database projects.

1. Introduction

Macromolecular crystallography is one of the largest communities at the Photon Factory. There are five beamlines, and more than 150 active beam time applications. The typical user has beam time once a month and makes several visits within a year. Because the throughput of the beamline improves each year, more experiments can be performed within the allotted beam time.

The five beam operation staff members support beamline operations 24 hours a day. Each year, beamline operations have advanced; besides the traditional experiment setup, users can conduct a fully automated experiment or chose to use either the mail-in or remote access service. Consequently, the experimental data from a single beam time and the data derived from beamline operations have drastically increased, making it difficult for users and the operation staff to organize the discursive data. To overcome this problem, we have developed a data management system, which can integrate all information derived from beamline operations, including experimental data, beam time applications, beam time schedule, and beamline operation reports.

2. Implementation of Data Management System

[1] To whom any correspondence should be addressed.

To construct a data management system, we employed commercial middleware, RCM (R&D Chain Management) System, developed by Quatre-i-Science Inc., Japan. In this middleware, we have developed and integrated several database projects for beamline experiments, beamline operations, beam time, and user information.

2.1. RCM System

The RCM System was developed to manage a large volume of data generated in the R&D process. An example application is the data analysis system for the Subaru Telescope camera [1]. The RCM System consists of three main components: Control server, Database server, and Web server (Fig. 1). All data management procedures are described in 'templates', which are chains of small-sized tasks in the RCM; examples of templates include registering data into the database, uploading files from a file server, executing a command on an external server, and sending emails. The RCM Control server sequentially executes tasks defined in a template as 'workflow'. The RCM Database server employs an XML-based database management system, which is advantageous because the management system provides flexibility. The RCM Web server is the interface of the RCM System for users and developers, and is equipped with a customizable user interface to execute templates and view workflow results. The interface can also be replaced by a Web page coded in general HTML or another scripting language for a more specialized purpose.

In the RCM System, a database and a set of templates for a particular purpose are treated as a 'project'. Adding new projects can extend the system.

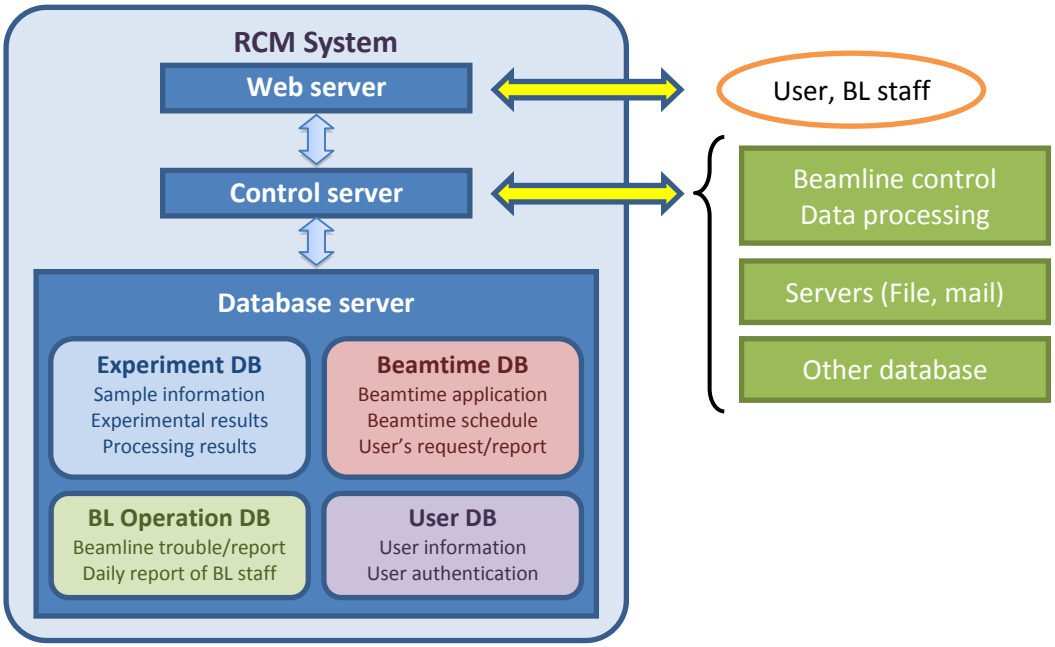


Figure 1. Schematic overview of the data management system.

2.2. Experiment database project

PReMo (PF Remote Monitoring system) is a platform to view experimental conditions and results performed at PF MX beamlines. This platform can process data automatically or remotely. PReMo was originally developed in 2006 by combining two beamline control clients and a house-made web interface [2]. However, this original system had difficulties with large volume databases and extending new functions. Consequently, we decided to replace the backend of PReMo to the RCM system in 2010.

In this project, the workflow continuously monitors the beamline control system. When an experiment is performed at a beamline, this workflow registers the experimental conditions and results

into the database, and then submits a data processing job to a workstation cluster. Upon completion of the data processing job, the workflow registers the results into the database. To search and visualize the results, there is a set of templates that a user can execute from their Web browser to view experiments performed at the beamlines. For automated data processing, currently a user can choose the following software: Labelit [3], Denzo/Scalepack [4], XDS [5], or Xia2 [6], and this preference is stored in the user database described later. Additionally, there is a template to submit sample information within containers of the sample exchange system. Upon executing this template, a user can register sample information into the database and beamline control software. In fully automated data collection experiments and mail-in services, this function is routinely used.

Figure 2a shows an example of the database content using the XML database format, which can be easily exported to other databases such as Information Platform of Targeted Protein Research Program, Japan.

a)

```
<expInfo>
 <userid>yamada</userid>
 <beamline>17a</beamline>
 <experiment>snapshot</experiment>
 <executiontime>2011-12-15 08:15:19</executiontime>
 <cameraheight>0.0</cameraheight>
 <cameralength>65.2</cameralength>
 <wavelength>2.0000</wavelength>
   (…snip…)
 <file compress="gzip">
     (…snip…)
 </file>
 <process>
  <status>Processed</status>
  <software>labelit</software>
  <total_spots>1259</total_spots>
  <ice_rings>0</ice_rings>
  <resolution1>1.89</resolution1>
    (…snip…)
  <file compress="yes">
     (…snip…)
  </file>
 </process>
</expInfo>
```

b)

```
<application>
 <proposalNumber>2010G565</proposalNumber>
 <title>STRUCTURE-BASED DRUG DESIGN FOR
NEURAMINIDASES FROM HUMAN AND VIRUSES</title>
 <label_j>Leonard CHAVAS</label_j>
 <label_e>Leonard CHAVAS</label_e>
 <spokesperson>Leonard CHAVAS</spokesperson>
 <emailAddress>leonard.chavas@kek.jp</emailAddress>
 <affiliation>High Energy Accelerator Research Organization
-Assistant Professor</affiliation>
 <status>active</status>
 <validFrom>2010-10-01</validFrom>
 <validTo>2011-03-31</validTo>
 <userTagid>33351483</userTagid>
</application>
<beamtime>
 <registeredDate>2012-07-24 14:18:11</registeredDate>
 <beamline>AR-NW12A</beamline>
 <startTime>2012-04-17 18:30:00</startTime>
 <endTime>2012-04-18 08:30:00</endTime>
 <proposalNumber>2010G565</proposalNumber>
 <status>active</status>
 <type>general</type>
</beamtime>
```

Figure 2. XML database structure: Examples from (a) the PReMo database and (b) the beam time database.

2.3. Beamline operation database project

Beamline operation database includes reports on maintenance and issues during beamline operations as well as the business reports of all staff members. To effectively share information, when a staff member submits a report via a Web browser, the report is registered in the database and is sent to other staff as an E-mail. The registered report can contain several flags, one of which marks the report as "unresolved". By using this flag, all beamline staff are aware that there is a beamline operation problem and can easily share information.

2.4. Beam time database project

In this project, information about beam time applications and the beam time schedule are registered into the database. A template, which is coded to send a reminder E-mail to the spokesperson for allotted beam time on certain days, is executed daily. Besides beam time information, the reminder provides additional useful information such as the user account to login to the computers at the

beamline. Furthermore, the reminder E-mail contains a link to a Web page with a questionnaire about the user's upcoming beam time, which inquires about vital information such as "What wavelength will you use?", "Will you use the sample exchange robot?" and "Will you access the beamline remotely?" The answers are stored into the database, and the beamline operation staff can prepare the beamline as per each user's request prior to arrival. Figure 2b shows an example of the XML database content in this project.

2.5. User database project

The user database project contains information regarding user account and preferences for options related to beamline experiment and data processing. By modifying these options, a user can customize their experimental environment.

2.6. Integration of database projects

Because all database projects are implemented in a single management system, these databases can be easily integrated. For users, experiments are linked with their schedule, and the results can be easily organized. For the beamline staff, operation reports and user's experiments are interlinked, so they can easily investigate reported issues.

3. Conclusion and future perspectives

By introducing the RCM system for data management at the PF macromolecular crystallography beamlines, we successfully developed several database projects and implemented them in a short period. The system is flexible; new projects can be added and integrated with other database projects while continuing to operate existing projects.

In the near future, a new database project for the large-scale crystallization system, which was developed by Structural Biology Research Center at our facility, will be launched. Integrating the crystallization and diffraction experiment databases should construct an effective pipeline for macromolecular crystallography.

Acknowledgements

The authors would like to thank Mr. Ken-ichi Kawasaki of Nihon Axis Co., Ltd, for his help in coding the RCM workflow templates and Web pages. We also thank all the staff at the PF macromolecular crystallography beamline for their suggestions to improve the system. We would like to thank Dr. Manobu Tanaka at KEK for his useful suggestion for the initial development of our data management system. The RCM system was developed by Quatre-i-Science Inc., Japan. This work was partly supported by Grants-in Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), and the Targeted Proteins Research Program (C-1) of the MEXT.

References

[1] Furusawa H., Okura Y., Mineo S., Takata T., Nakata F., Tanaka M., Katayama N., Itoh R., Yasuda N., Miyazaki S., Komiyama Y., Utsumi Y., Uchida T. and Aihara H. 2011 *Publ. Astron. Soc. Japan* **63** S585
[2] Yamada Y., pHonda N., Matsugaki N., Igarashi N., Hiraki M. and Wakatsuki S. 2008 *J. Synchrotron Rad.* **15** 296.
[3] Sauter N.K., Grosse-Kunstleve R.W. and Adams P.D. 2004 *J. Appl. Cryst*. **37** 399
[4] Otwinowski Z. and Minor W. 1997 *Methods in Enzymology* **276** 307
[5] Kabsch W. 1993 *J. Appl. Cryst.* **26** 795
[6] Winter G. 2010 *J. Appl. Cryst.* **43** 186