Development of an Informatics Tool for Crystallography Laboratory Administrators

Leah Sandvoss * Dennis P. Groth
Indiana University School of Informatics,
Bloomington, IN 47408, USA.
lsandvos@indiana.edu, dgroth@indiana.edu

Abstract

With increased demand for storage of scientific data comes a corresponding demand for efficient retrieval mechanisms necessary for analytical and reporting purposes. As is often the case, the design of databases to support scientific data is (rightly so) driven by the science. Unfortunately, this "science-centric" view of data management does not make the subsequent reporting of information easy. The natural solution to such a problem is, of course, a data warehouse approach to scientific data management.

The focus of this project is the design of a data warehousing architecture for scientific data. Our goal is to separate the science-specific aspects of the information from the reporting and analysis requirements. The problem domain we are currently investigating is crystallographic structure data, coupled with metadata concerning the structure. This paper describes ITCLA - an Informatics Tool for Crystallography Laboratory Administrators. We report on the current state of ITCLA, as well as our future plans.

1 Introduction

A challenge faced by the database research community is the development of techniques appropriate for scientific disciplines. The tension between genericity of support and specificity of application remains as a hurdle for database developers.[5] With increasing success, however, the use of databases for managing scientific data has gained some acceptance. Crystallography databases, for example, first emerged in the late 1960's and contain many hundreds of thousands of structures. The most well-known and well-used crystallography databases to date are the Cambridge Structural Database (CSD) [4] for small organic compounds, the Inorganic Crystal Structure Database (ICSD) [1] for inorganic compounds, and the Protein Data Bank (PDB) for macromolecules [12].

As was the case with the adoption of database technology by the business community for archival purposes in support of operations, scientific data is increasingly in demand for analytical purposes. It is clear that creating a data warehouse for scientific data is a sound strategy.[3, 8]

The aim of this project is the development of an Informatics Tool for Crystallography Laboratory Administrators (ITCLA). This tool will allow the user, which is the crystallography lab administrator, to analyze the data in the current collection of crystal structures, known as the Indiana University Molecular Structure Center (IUMSC) database, part of the Reciprocal Net collaboratory (http://www.reciprocalnet.org). The current design of the database is focused on accurately capturing the information as it is generated in the course of the crystallography experiments. Consequently, the main vehicle for retrieving information about the crystal structures is a simple search mechanism that retrieves a single sample at a time. The current system lacks an ability to get a high-level view of search results. The most appropriate search mechanism should adhere to the "overview first, then details on demand" maxim of Shneiderman.[9]

The objective for ITCLA is to serve as a tool for crystallographers to search for and retrieve information from the IUMSC database. The design of a data warehouse to support this activity is our first step, with the recognition that the users of the system are not expert query writers. Our intention is to store summarized data in the data warehouse, with access driven by the metadata. Output for data from the system will be processed via an XSLT (eXtensible Style Language Transformations) [11] for presentation in a number of forms, including HTML (HyperText Markup language) [13] and delimited text.

The remainder of this paper is structured as follows.

*Work supported by NSF Grant 0121699

Section 2 provides a brief introduction to our problem domain. Section 3 describes our current design and details some example queries to be supported. Lastly, we conclude with future directions in Section 5.

2 Crystallography Metadata

Many parameters are recorded and necessary to define a crystal structure well. One crucial set of descriptors are the unit cell parameters, which describe the simplest of possible repeating units. These are known as a, b, c values, which define the length of the edges of the unit cell, and the α, β, γ values, which define the angles created by these edges.[6] Within these unit cells, the structure can be classified by a shape descriptor called the "space group."[2] The number of molecules composing the unit cell is also important, and is referred to as the Z-value. The resolution of the structure is a good indicator of expected structure quality: the higher the resolution, the better the quality. Another, although less reliable, indicator of quality is the R-factor, which is the ratio between the observed and calculated intensities. Other parameters that are often noted are the volume and density of the crystal, as well as the temperature at which the compound was irradiated.

All of the above parameters are recorded in the IUMSC database. Each crystallographer fills in an online form in order to capture all the information pertaining to solving the structure. In addition to the crystallography data, the dates of sample submission, report completion, and final sample results are automatically generated and entered into an appropriate field in the database. Other essential information recorded are the names of the people related to the crystal sample: the provider (laboratory), the submitter (student, postdoctoral fellow, or professor), and the crystallographer examining the crystal. The billing information may also be recorded in the same database; for example, the account number, amount paid, and the billing date are often inserted.

3 Current Design

The details for the current design of ITCLA are illustrated in Figure 1. The sample metadata are obtained from the IUMSC database, while the structured data is stored in the data warehouse. A Perl-CGI script calls the appropriate queries based on the input to the web site form, which may change based on the data returned.

The use of an offline data warehouse is appropriate for ITCLA because of its need for efficient querying and reporting. (MySQL, the open-source database engine that supports IUMSC's online transactions, lacks particularly flexible query support.) Creating relations that contain more accessible data will greatly facilitate the return of information. This data warehouse is housed in a separate database from the online database, and is routinely regenerated from the latter by executing approximately fifteen relational joins.

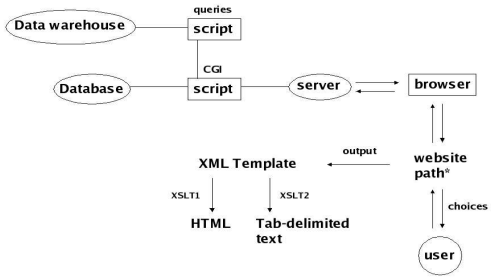

Figure 1. Proposed data flow for ITCLA.

Because the data in the IUMSC database was recently restructured to match other Reciprocal Net sites, some of the new fields pertain only to data that has been added after this change. Thus, it is important that a sample's status be tracked explicitly throughout its lifetime. Possible sample status values already identified are known as retracted, complete, incomplete, nonscs (not a single crystal), nogo (would not crystallize or a poor crystal), and withdrawn. Additional status types pertain to whether the sample is public, i.e. have been published, and include: complete_public, incomplete_public, non_scs_public and retracted_public.

As is generally the case with the construction of a data warehouse, the tables in the ITCLA warehouse are subject based. Examples of the types of tables that will be included are listings of information on the people submitting samples, crystallographers, the instruments used, sample status, information on the sample report, account information, and crystallography data.

Against this data warehouse a number of queries can be executed. Examples are calculating billings by provider per year; average expenditure of a provider per year; average number of samples submitted per provider; average number of samples solved per provider; number of completed samples per specific diffractometer, etc. In addition, aggregate functions for the crystal data are computed: the minimum, maximum or average Z-value, R-factor, cell parameters, temperature and volume. All information is related to a specific sample, identified by a unique sampleId, with results sorted according to the numeric value of the ID. Figure 2 shows a diagram of choices the user can make for forming a list or to manipulate data.

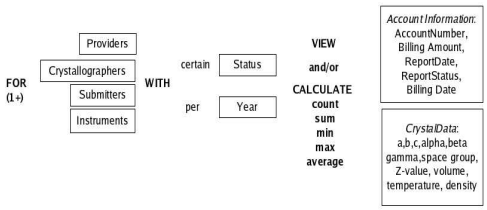

Figure 2. Choices for data listing and manipulation. Each tuple in database is per crystal structure attempted.

4 Output

Initially, the data returned will be in a report-type format, unlike many search capabilities currently available for chemistry. For the most part, existing search capabilities for chemical data simply list all possible samples pertaining to the search request, allowing the user to click and view only one sample at a time. As illustrated in Figure 1, after the user has navigated through the data driven web page path and chosen to view the report generated, they can then decide upon the format for the report. The choices will be HTML or tab-delimited text. Initially, the pertinent data will be placed in an XML [5, 7, 10, 11] template, then appropriate XSLTs will be written to transform the data from XML format to tab delimited text or to HTML.

ITCLA will provide the capability of retrieving many results and viewing all the requested data for that information, then providing the decision of how to manipulate or view those results. This report format will allow administrators swift viewing of the output, as well as the ability to paste or import the results into a statistical software package for further analysis.

5 Conclusion

This project is unique for several reasons. First, as just described, the data is returned in an operative format. Next, the data being manipulated is crystallographic data; less specifically, chemical data. In many academic chemistry departments, it is uncommon that experimental data for chemistry be stored in an organized electronic format as sophisticated as an SQL database. This new academic arena of developing better and more efficient methods for storing, displaying, filtering, and mining chemical data has been deemed chemical informatics, and in this case it is better defined as crystallographic informatics.

Other novel points of the project include the method of extracting the information, which is largely data driven, a design that is especially helpful in this data-intensive science. Finally, the capabilities provided by ITCLA will save a significant amount of time for the administrative user, who otherwise would have performed all of these tasks manually, allowing him/her to focus on more complicated statistical calculations.

Acknowledgements

Special thanks merited to John C. Huffman, head scientist of the IUMSC, for providing this project opportunity, as well as to John C. Bollinger and Eric F. Koperda for leading the Reciprocal Net development effort.

References

[1] ICSD, Inorganic Crystal Structure Database, FIZ KARLSRUHE Information Services, Fiz KarlRuhe, Germany.

[2] *International Tables for Crsytallography*, volume A. Kluwer Academic Publishers, 1996.

[3] K. Aberer and K. Hemm. A methodology for building a data warehouse in a scientific environment. In *Conference on Cooperative Information Systems*, pages 90–101, 1996.

[4] F. Allen and O. Kennard. CCDC, cambridge crystallographic data centre [cambridge structural database]. In *Chem Des. Autom. News*, volume 8, page 31, 1993.

[5] P. Buneman. What is special about scientific data? In N. M. Patrikalakis, editor, *Proceedings of the NSF Invitational Workshop on Distributed Information, Computation, and Process Management for Scientific and Engineering Environments (DICPM)*, pages 45–46, 1998.

[6] e. L. Giacovazzo. *Fundamentals of Crystallography*. Oxford Science Publications, 1993.

[7] P. Murray-Rust and H. Rzepa. Scientific publications in XML - towards a global knowledge base www.ch.ic.ac.uk/rzepa/codata/, September 2002.

[8] T. B. Pedersen and C. S. Jensen. Research issues in clinical data warehousing. In *Statistical and Scientific Database Management*, pages 43–52, 1998.

[9] B. Shneiderman. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison Wesley Longman, Inc., 3rd edition, 1998.

[10] The Apache Software Foundation. DTD/XML schema links xml.apache.org/cocoon/link/dtd-schema.html, December 2002.

[11] T. Usdin and T. Graham. XML: Not a silver bullet, but a great pipe wrench. In *StandardView*, volume 6. September 1998.

[12] H. B. J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. In *Nucleic Acids Research*, volume 28, pages 235–242, 2000.

[13] World Wide Web Consortium. Hypertext markup language (HTML) Massachusetts Institute of Technology, Institut National de Recherche en Informatique, Keio University. www.w3.org/Consortium/Legal, 2003.