and visualize CIF format, syntax and content, while the CCDC visualiser, Mercury, is fully CIF-enabled, and the most recent version will represent atomic displacement parameters, a feature that will also be incorporated into enCIFer.

We are now developing software tools that assist the processing of CIFs to the CSD: (a) We use knowledge-based and bond valence sum data to establish crystal connectivity, and then assign chemical bond types algorithmically. The algorithm has a success rate of 86.4% when validated against a test set of 1104 structures, including a significant proportion of challenging metal-organics. (b) Heuristic analysis of CIFs permits resolution of disordered molecules/ions into their discrete components, using atomic occupancy factors, together with any CIF 'group' and 'assembly' fields that may be available. (c) We are generating chemical diagrams directly from the CIF using a variety of software tools and measures of chemical similarity with existing CSD structures.

Keywords: CIF applications, cambridge structural database, data processing and visualisation

OCM03.26.5

Acta Cryst. (2005). A**61**, C128

PDBML: the XML-based Database and its Applications

Haruki Nakamura[a], Nobutoshi Ito[a], Reiko Yamashita[b], Daron, Standley[b], Arno Paehler[b], Atsuro Yoshihara[b], [a]*School of Biomedical Science, Tokyo Medical and Dental University.* [b]*PDBj, BIRD, Japan Science and Technology Agency.* [c]*Institute for Protein Research, Osaka University.* E-mail: harukin@protein.osaka-u.ac.jp

A canonical XML for PDB, PDBML [1], has been developed by PDBj (Protein Data Bank Japan) [2] and RCSB. Its structure is defined in XML schema (http://deposit.pdb.org/pdbML/pdbx.xsd) and all the contents in the PDBML files are now well validated.

We have built a native XML database with this new formant and made it available to public through a web service called xPSSS (http://www.pdbj.org/xpsss). In addition to simple keyword search full XPath searches with the SOAP interface are also implemented for complicated searches and large-scale analyses. The contents of the database are also enhanced; additional data such as the biological and biochemical functions and the experimental details extracted from the literatures and other databases are included.

A few applications have also been developed around the database; (i) a molecular graphics viewer, jV, which can parse the PDBML data [3], (ii) electron density maps for evaluation of the structure quality, (iii) the sequence and structure neighbors defined by maximizing the number of equivalent residues (NER) rather than minimizing conventional RMSD [4].

[1] Westbrook J., et al., *Bioinformatics*, 2005, *in press*. [2] Berman H., et al., *Nat. Struct. Biol.*, 2003, **10**, 980. [3] Kinoshita K., Nakamura H., *Bioinformatics*, 2004, **20**, 1329. [4] Standley D., et al., *Proteins*, 2004, **57**, 381.

Keywords: databases, database manipulation, data representation

OCM03.26.6

Acta Cryst. (2005). A**61**, C128

Use of mmCIF in the Publication of Macromolecular Crystallography Communications

Brian McMahon[a], Peter R. Strickland[a], Howard M. Einspahr[b], J. Mitchell Guss[c], Louise E. Jones[a], [a]*IUCr, 5 Abbey Square, Chester CH1 2HU, UK.* [b]*PO Box 6395, Lawrenceville, NJ 08648-0395, USA.* [c]*School of Molecular and Microbial Biosciences, University of Sydney, NSW 2006, Australia.* E-mail: bm@iucr.org

Acta Crystallographica Section F: Structural Biology and Crystallization Communications was launched by the IUCr in 2005 as a rapid-publication electronic-only journal for communications on the crystallization and structure determination of biological macro-molecules. Structure reports are expected to come initially from structural genomics and protein–ligand studies. For such reports, the IUCr is collaborating with the Protein Data Bank (PDB) to facilitate the deposition and publication procedures by extracting as much information as possible from program output files and capturing additional information in a standard exchange mechanism. The mechanism that is used is the macromolecular crystallographic information file (mmCIF). Authors may send the mmCIF generated during the PDB deposition process directly to the journal office, where automated scripts use it for preparation of validation reports for scrutiny by the referees, and for generation of summary tables of information, suitable for embedding within the article, about the structure determination and refinement. Details of these procedures will be given. Further work will be undertaken to refine the mmCIF dictionary and facilitate its use for publication purposes.

Keywords: IUCr journals, data definition, mmCIF

OCM03.26.7

Acta Cryst. (2005). A**61**, C128

mmCIF and Dictionary Driven Software with the MSD Database Production Pipeline

Kim Henrick, Sameer Valenkar, Harry Boutselakis, Ayzaz Hussain, John Ionides, Adamandia Kapopoulou, Peter Keller, Richard Newman, Jorge Pineda, Antonio Suarez, Jawahar Swaminathan, John Tate, *European Bioinformatics Institute, EMBL Outstation Hinxton, Wellcome Trust Genome Campus Hinxton Cambridge UK.* E-mail: henrick@ebi.ac.uk

The Macromolecular Structure Database (MSD) group, http://www.ebi.ac.uk/msd, based at the European Bioinformatics Institute (EBI) (an outstation of the European Molecular Biology Laboratory EMBL) is a member of the wwPDB (http://www.wwpdb.org) and provides one of the PDB deposition sites. The MSD uses AutoDep4 (http://www.ebi.ac.uk/msd-srv/autodep4/) for PDB depositions and carries out not only the annotation tasks required to produce PDB entries it also provides a stable and clean repository of macromolecular structure data services that allow users to access, search and retrieve structural data. In addition the MSD handles the deposition and archive for Maps from cryo-electron microscopy through the separate deposition interface EMDep, http://www.ebi.ac.uk/msd-srv/emdep/. mmCIF and XML are used throughout the data processing pipeline including the mapping of PDB entries to an in-house extended mmCIF for loading intio the Oracle databases and in the numerous processes that are run on the deposition data to enrich the PDB data with extensive derive information. The MSD export the PDB entries to the RCSB in the standard pdbx dictionary format. This talk will outline part of the processing pipeline used.

Keywords: mmCIF, relational databases, processing pipeline

OCM03.26.8

Acta Cryst. (2005). A**61**, C128

mmCIF Applications at the RCSB Protein Data Bank

Zukang Feng, Helen M. Berman, Huanwang Yang, John D. Westbrook, *RCSB Protein Data Bank, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey.* E-mail: jwest@rcsb.rutgers.edu

The RCSB Protein Data Bank (www.pdb.org) has developed a variety of software tools to manage mmCIF data. These applications include validating parsers, format translators, database loaders, and data extraction tools. The latter suite of extraction tools, pdb_extract[1], automates the collection of key data items from most popular X-ray and NMR structure determination applications. These data are assembled into an mmCIF data file ready for PDB deposition that can be submitted using the AutoDep Input Tool (ADIT). The use of these tools in automating the PDB deposition and validation process will be described.

The RCSB PDB is funded by NSF, NIGMS, Office of Science DOE, NLM, NCI, NCRR, NIBIB, and NINDS.

[1] Yang H., Guranovic V., Dutta S., Feng Z., Berman H.M., Westbrook J.D., *Acta Cryst*., 2004, **D60**, 1833.

Keywords: mmCIF, ontologies, protein structure representation