CIF2CML - Automatic Processing of Chemical Crystallography in XML/CML

Peter Murray-Rust[a], Simon Tyrrell[a], [a]*Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, CB2 1EW, UK.* E-mail: pm286@cam.ac.uk

The CIF data structure and hierarchy (CIF, dictionaries, DDL) is largely isomorphic with XML (document/DOM, schema, XMLSchema) and XML tools can therefore be configured to process data from CIFs. First our CIF2CML toolkits convert CIF documents to XML. The data are then validated structurally and semantically (against the dictionaries) and further converted to Chemical Markup Language (CML) (http://www.xml-cml.org).

The crystal structures in CML can then be stored, chemically validated and transformed using the JUMBO CML library and other CML-aware tools. Among the steps are (a) checks on chemical composition (b) treatment of disorder (c) application of symmetry (d) assignment of bonds and (e) molecules (f) unique chemical identification (IUPAC InChI) (g) calculation of 2D coordinates (h) storage in XML repository to create a structural knowledge base which can be searched for chemical and geometrical concepts.

The approach is highly modular with many hundred interoperable components, designed for use with WebServices (http://wwmm.ch.cam.ac.uk/gridsphere/gridpshere) and workflows such as Taverna (http://taverna.sf.net) and institutional repositories (http://eprints.soton.ac.uk/1633/) with Open data. We argue that Open source and Open data provide a robust high-throughput crystallographic semantic web whose prototype will be demonstrated.

Keywords: CIF, XML, CML