

The Crystallographic Semantic Web

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

The semantic Web [Berners-Lee, <http://www.w3./2001/sw/>] is a vision of a global knowledge network where machines understand and reason from web-based resources. We provide working demonstrations that this technology is ideally suited to create an Open crystallographic knowledge base. The complete information chain - experiment, publication, storage, dissemination, searching and re-use can be completely managed by machines. CIFs can be converted to XML, annotated, and redistributed as WebServices and indexed by conventional search engines. New structures can be announced through crystallographically enhanced RSS feeds.

Our World Wide Molecular Matrix (WWMM, <http://wwmm.ch.cam.ac.uk>) uses free-text XML indexing in a repository (eXist) to collect hundreds of thousands of compounds. IUPAC InChI and chemical substructure searching (e.g. <http://openbabel.sf.net>) to provide an instant, Open and freely redistributable crystallographic knowledge base. Data are abstracted from publishers who allow Open access to CIFs and authors can publish CIFs directly into XML (<http://www.xml-cml.org>). Institutional and national collections ((<http://eprints.soton.ac.uk/1633/>) with appropriate metadata allow conventional search engines to index the data and effectively create a complete Open database of crystal structures.

Keywords: XML, CIF, semantic web