McMahon, *Acta Cryst.* **2008**, A*64*, 38-51. [4] J.D. Westbrook, H. Yang, Z. Feng, H.M. Berman, *International Tables for Crystallography* **2005**, *G*5.5, 539-543. [5] L. Lyon, *Consultancy Report* **200,** Bath, UK: UKOLN. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

Keywords: CIF, electronic publishing, information management

MS.89.2

Acta Cryst. (2011) A**67,** C194

Data management for photon and neutron sources

Brian Matthews, *e-Science Centre, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Didcot, OX11 0QX, (U.K.).* E-mail: brian.matthews@stfc.ac.uk

Photon and Neutron sources, such as the UK's Diamond Light Source and ISIS Spallation Neutron Source are large-scale facilities providing high resolution data for crystallography and other materials analysis techniques. Traditionally, the raw data generated from such facilities has been managed by the instrument and user scientists themselves. However, the current generations of such facilities can undertake a large number of experiments, and generate hugely increased volumes of data. As a consequence, the traditional approach has become unsustainable and a more automated approach to data management has had to be developed.

In this talk, I shall outline the data management infrastructure developed within STFC to manage raw data. This infrastructure takes an integrated approach to aggregate, store and catalogue data generated at ISIS and Diamond. In particular, I shall describe ICAT, a suite of tools which catalogues data as it is generated by beam lines, and provides access to that raw data to its user community, allowing them to search and retrieve their data, within the facilities themselves or within their home institution. This is provided using a service application programming interface so that a variety of different search and analysis tools can be interfaced to search and access the data, and also register and catalogue derived data.

The management of raw data is part of a wider scientific process, starting from proposals for research through to the publication of results. We shall further discuss how the ICAT and similar tools can be extended to support this wider process by allowing data to be federated across a number of different data sources and also linking the raw data to analysed and published data so that the provenance of data can be tracked; this is being considered in the project Integrated Infrastructure in Structural Sciences (I2S2). This allows data to be formally cited and reused, and results to be validated. We relate this work to the publication process being developed by the International Union of Crystallography, tracing the relationship between raw data generated from beam lines, and the CIF files lodged during the publication process.

This integrated data infrastructure is being taken forward by the European Photon and Neutron Data Infrastructure initiative (PaNData), a consortium of European photon and neutron sources serving an expanding user community of tens of thousands of scientists across Europe. The experiments in these facilities are of increasing complexity, they are increasingly done by international research groups and many of them will be done in more than one laboratory. The resulting data needs to be accessible over the Internet and remain on-line until the results are published and in many cases much longer to allow re-processing and to allow for the preservation of knowledge. PaNData is developing common data formats, data and software catalogues within the framework of a common data policy.

Keywords: data management, information management, large-scale facilities

MS.89.3

Acta Cryst. (2011) A**67,** C194

Crystaleye: Publication and re-use of open semantic crystallographic data

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

Berners-Lee's vision of the Semantic Web (SW) is now a reality in many scientific fields (including macromolecular structures and much bioscience). The SW is based on Linked Open Data (LOD) where each component of information is Openly available with a published unique identifier scheme. The LOD are linked together through RDF triples where the semantics are provided by published ontologies or dictionaries. This creates a graph (or "cloud") of data on web sites and in triple stores that can be explored by the current generation of SW tools.

In our Crystaleye system we have applied this approach to "small molecule" crystal structures (organic, inorganic and organometallic) by extracting Open Data from published CIFs, mainly on publishers' websites. The extraction is performed daily by our "Pubcrawler" system and any new CIFs are added and processed. Each CIF, split into individual data blocks if necessary, is automatically processed into semantic form (using Chemical Markup Language (CML) and RDF). During this process many validity checks are applied, in particular to extract and check the chemistry. After any reported disorder is processed the chemical connection table (CT) is created and checked against any reported formula and chemical names. The CT and compositional formula are then re-usable as primary indexes and search terms. All reported data in the CIF are translated to RDF, stored in our Chempound (chem#) repository where they can be searched through a SPARQL endpoint.

Where the full-text of the article is Open (as in Acta Cryst. E) we extract information from the text such as methods of preparation and crystallization as well as citations. This enhances the data in the CIF and creates a potentially valuable node in the LOD cloud. By comparing the deduced CT with the images and names in Acta E papers we show that the automatic generation of CTs has a precision/recall > 99%

Crystaleye (http://wwmm.ch.cam.ac.uk/crystaleye) provides a natural browsing interface to the crystal structure which includes interactive exploration and search. All bond lengths are indexed and can be searched by element types. Readers can link back to the original splah page and article if it is published on the web.

All data and software is fully Open (i.e re-usable for any purpose without further permission). Crystaleye, whose maintenance cost is near-zero, shows that it is possible to create a global knowledge base of crystallography simply by publishing CIFs to the open web and letting machines do the rest. The technology is also applicable to theses (which are currently under-used) and for departments to expose their unpublished data. Unfortunately restrictions imposed by some publishers and some data aggregators mean that the current coverage of Crystaleye is only partially complete. Besides the technology the presentation will address aspects of Openness in crystallography and low-cost approaches to sustainability.

Keywords: semantic, linkeddata, open

MS.89.4

Acta Cryst. (2011) A**67,** C194-C195

The wwPDB Working Format: A Simplified Application of CIF Technology

John Westbrook,[a] Helen M. Berman,[a] Jasmine Young,[a] Gerard J. Kleywegt,[b] [a]*RCSB PDB, Rutgers University, Piscataway, NJ (USA).*