s9.m29.o5 **Data Validation 10 Years On: The Use of Interactive Ontologie**s. Sydney R Hall[a] and Nick Spadaccini[b], [a]*School of Biomedical & Chemical Sciences;* [b]*School of Computer Science & Software Engineering, University of Western Australia, Nedlands 6009, Australia. E-mail: syd@crystal.uwa.edu.au*

Keywords: **Data-validation; Data-ontology; Methodology**

Data dictionaries are widely used by databases and journals to support the validation of submitted data. These dictionaries contain precise computer-readable information about data items used in particular disciplines. In crystallography, CIF data dictionaries exist for data used for structure analysis, macromolecular structures, powder diffraction, symmetry, incommensurate structures, and precision density studies. These are described in detail in the soon-to-be-published *International Tables Volume G* [1]. The primary function of these dictionaries is the precise identification and characterisation of frequently-used data items. This characterisation, which includes the definition of attributes that specify the dependencies between data items; whether they are numbers or text; and their allowed enumeration, underpins many data validation processes used currently by journals and databases when accepting deposited data.

Data dictionaries, or ontologies, as they are more generally referred to, can also provide detailed relational knowledge about data. This is usually in the form of 'methods' that record the functional relationship of *derivative* data items to *primitive* (i.e. measured or postulated) and other derivative data. In the main, methods are algorithms that allow non-primitive data to be evaluated from other data, and may be applied to classes of data as well as to individual items.

The next generation of ontologies will be capable of direct application to a particular data instantiation i.e. they are interactive and executable. Moreover ontologies will provide method scripts for the dynamic *redefinition* of the attributes (e.g. the enumeration of an item can be changed according to the value of another), *conformance* (important in DDL dictionaries where the instantiated data are data ontologies) and validation (i.e. methods for both consistency and quality checks).

This paper will describe an interactive ontology approach based on the *StarDDL* and *dREL* languages [2], and demonstrate how these are applied to particular data instantiations.

[1] *International Tables for Crystallography Volume G* (2004) Eds: S R Hall and B McMahon. Kluwer Academic Press: London.
[2] Spadaccini, N., Hall, S.R. and Castleden, I.R. (2000) *J. Chem. Inform. & Computer Sci.* **40**, 1289-1301.

s9.m30.o1 **Automatic Structure Determination. Is it Just a Dream?** D.J.Watkin & R.I.Cooper, *Chemical Crystallography Laboratory, 9 Parks Rd, OXFORD, OX1 3PD, UK. E-mail: david.watkin@chem.ox.ac.uk*

Keywords: Validation; Automation; Analysis

"Methods have been developed by Ford, Hodgson, Rollett & Stonebridge (unpublished) for automatic solution of crystal structures".[1]

Some of this 30 year-old optimism has been justified by subsequent events. SIR92 and its successors have been remarkably successful at both locating atomic sites and at assigning atomic types even for quite complicated structures - given an accurate estimate of the atomic composition. Even so, every Service Analyst will know that fully automatic determination of crystal structures is still a dream. The 'crystal-in ORTEP-out' black box may work for some structures, but examination of 500 structures completed by the service analyst in Oxford indicated that 30% of 'small' organic and 60% of organometallic structures need human intervention for their completion.

This leads us to ask 'What do humans know that programs don't, and what can humans do that programs cannot?'

The answer to the first question is that humans can develop a real understanding of chemistry and physics, so that they have a completely independent check on the plausibility of a proposed structure. In the event that something goes wrong this knowledge plus imagination enables them to propose alternative solutions.

The answer to the second question is that humans can learn from their own and other peoples experience. Current crystallographic programs can only do (if one is lucky) what their designer intended them to do. Some years ago, in the heyday of Artificial Intelligence, there seemed to be the prospect of programs improving their own reaction to problems, but so far this technology has made little impact in crystallography. If the resources being spent on Google were available to crystallographers, things might be very different.

For the moment we must base our confidence in automatically determined structures on the findings of programs such as PLATON, CHECKCIF and MOGUL. These may spot when things have gone wrong, but it will still take human imagination to put difficult cases right.

[1] (Rollett, J.S. 'Least Squares Procedures', in Crystallographic Computing, Ed Ahmed, 1970).