Mention should also be made of the use of STAR Files by the BioMagResBank group at the University of Wisconsin to record NMR structures. This work (Ulrich *et al.*, 1998) endeavours to be complementary to the mmCIF descriptions of structures in the Protein Data Bank.

1.1.12. The broader context: CIF and XML

In the light of more recent data-exchange developments, it will be surprising to newcomers to CIF that more use is not made in crystallography of the extensible markup language XML (W3C, 2001). However, the development of CIF predates XML, and the CIF format can be easily translated to and from suitable XML representations. Most current crystallographic software imports and exports data in CIF format and the use of XML only becomes important in applications that cross the boundary of crystallography and involve interoperability with other scientific domains.

At one time, the antecedent of XML, standard generalized markup language (SGML; ISO, 1986), was considered as a candidate for a crystallographic exchange mechanism. SGML is a highly flexible and extensible system for specifying markup languages, but is extremely general. In the late 1980s, successful SGML implementations stretched the capacity of affordable computers and little accompanying software was available. SGML was at that time a suitable data and document tagging mechanism for large-scale publishers, but was far from appropriate for smaller-scale applications. XML was introduced during the 1990s as a specific SGML markup with a concrete syntax and simplifications that resulted in a lightweight, manageable language, for which robust parsers, editors and other programs could be written and implemented on desktop computers. The consequence has been a very rapid adoption of XML across many disciplines. Parallels may be drawn with the decision to implement CIF as a subset of the more general STAR File.

XML provides the ability to mark up a document or data set with embedded tags. Such tags may indicate a particular typographic representation. More usefully, however, they can reflect the nature or purpose of the information to which they refer. The design of such useful and well structured content tagging (in XML and other formalisms) is referred to in terms of constructing a subject or domain *ontology*. This term is rather poorly defined, but broadly covers the construction for a specific topic area of a controlled vocabulary of terms, the elaboration of relationships between those terms, and rules or constraints governing the use of the terms.

In XML and SGML, document-type definitions (DTDs) and schemas exist as external specifications of the markup tags permitted in a document, their relationships and any optional attributes they might possess. If carefully designed, these have the potential to act as ontologies. A simplification that XML offers over SGML is the ability to construct documents that do not need to conform to a particular schema. This makes it rather easier to develop software for generating and transmitting XML files. However, if diverse applications are to make use of the information content in an XML file, there must be some general way to exchange information about the meaning of the embedded markup, and in practice DTDs or schemas are essential for interoperability between software applications from different sources.

Recent initiatives in chemistry, under the aegis of the International Union of Pure and Applied Chemistry (IUPAC), suggest an active interest in the development of a machine-parsable ontology for chemistry. Building on an existing XML representation of chemical information known as Chemical Markup Language (CML; Murray-Rust & Rzepa, 1999, 2001), IUPAC project groups are mapping out areas of the science for which suitable DTDs and schemas may be constructed that tag relevant chemical content.

In this context, the CIF dictionaries described and annotated at length in this volume provide a sound basis for a machine-parsable ontology for crystallographic data. Given the orderly classification and relationships between tags in a CIF data set, format transformations using XML tags are not at all difficult to achieve. Chapters 5.3 and 5.5 describe tools for converting between mmCIF and XML formats for interchange within the biological structure community.

In the future, we expect to see the growth of XML environments where the full content of the CIF dictionaries is imported for the purpose of tagging crystallographic data embedded in more general documents and data sets. There have already been examples of ontology development using CIF dictionaries; for example, the Object Management Group has developed software classes for middleware in biological computing applications that are modelled on the mmCIF dictionary (Greer, 2000). The use of interactive STAR ontologies is described by Spadaccini *et al.* (2000).

It is possible that as interdisciplinary knowledge-management software systems develop, the exchange of crystallographic data will occur through XML files or other formats. Neverthless, CIF will remain an efficient mechanism for developing detailed data models within crystallography. We can confidently say that, whatever the actual transport format, the intellectual content of the dictionaries in this volume and their subsequent extensions and revisions will continue to underpin the definition and exchange of crystallographic data.

References

Allen, F. H. (2002). *The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Cryst.* B**58**, 380–388.

Anklesaria, F., McCahill, M., Lindner, P., Johnson, D., Torrey, D. & Alberti, B. (1993). *The Internet gopher protocol (a distributed document search and retrieval protocol).* RFC 1436. Network Working Group. http://www.ietf.org/rfc/rfc1436.txt.

ANSI/NISO (1995). *Information retrieval (Z39.50): Application service definition and protocol specification.* Z39.50-1995. http://lcweb.loc.gov/z3950/agency/1995doce.html.

Barnard, J. M. (1990). *Draft specification for revised version of the Standard Molecular Data (SMD) format. J. Chem. Inf. Comput. Sci.* **30**, 81–96.

Berners-Lee, T. (1989). *Information management: a proposal.* Internal report. Geneva: CERN. http://www.w3.org/History/1989/proposal-msw.html.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.

Bourne, P., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D. & Fitzgerald, P. M. D. (1997). *Macromolecular Crystallographic Information File. Methods Enzymol.* **277**, 571–590.

Brown, I. D. (1988). *Standard Crystallographic File Structure-87. Acta Cryst.* A**44**, 232.

Busing, W. R., Martin, K. O. & Levy, H. A. (1962). *ORFLS.* Report ORNL-TM-305. Oak Ridge National Laboratory, Tennessee, USA.

Chapuis, G., Farkas-Jahnke, M., Pérez-Mato, J. M., Senechal, M., Steurer, W., Janot, C., Pandey, D. & Yamamoto, A. (1997). *Checklist for the description of incommensurate modulated crystal structures. Report of the International Union of Crystallography Commission on Aperiodic Crystals. Acta Cryst.* A**53**, 95–100.

Fitzgerald, P. M. D., Berman, H., Bourne, P., McMahon, B., Watenpaugh, K. & Westbrook, J. (1996). *The mmCIF dictionary: community review and final approval. Acta Cryst.* A**52** (Suppl.), C575.

Freer, S. T. & Stewart, J. (1979). *Computer programming for protein crystallographic applications, University of California at San Diego, 28-29 November 1978. J. Appl. Cryst.* **12**, 426–427.