The macromolecular dictionary was built using DDL2 and was presented as a draft to the community at the American Crystallographic Association meeting in Montreal in July 1995. The draft was subsequently posted on a website and community comment solicited *via* an e-mail discussion list. This provoked lively discussions, leading to continuous correction and updating of the dictionary over an extended period of time. Software for parsing the dictionary and managing mmCIF data sets was developed and was also presented on the website.

In January 1997, the mmCIF dictionary was completed and submitted to COMCIFS for review. In June 1997, version 1.0 was approved by COMCIFS and released (Bourne *et al.*, 1997; Fitzgerald *et al.*, 1996). A workshop was held at Rutgers University in October 1997, at which tutorials were presented to demonstrate the use of the various tools that had been developed.

More than 100 new definitions have been incorporated in version 2 of the mmCIF dictionary, presented in Chapter 4.5.

1.1.10. The Crystallographic Binary File

In 1995, Andy Hammersley approached the IUCr with a proposal to develop an exchange and archival mechanism for image data using the CIF formalism. There was an increasing need to record and exchange two-dimensional images from a growing collection of area detectors and image plates from several manufacturers, each using different and proprietary data-storage formats. The project was encouraged by the IUCr and was developed under a variety of working names. In October 1997, a workshop at Brookhaven National Laboratory, New York, was convened to discuss progress and to coordinate the development of software suitable for this new format.

At the workshop, it became apparent that there was broad consensus to adopt and further develop the working set of data names that had been devised by the working group on this project; it was further decided that the relationships between these data names were best handled by the DDL2 formalism used for mmCIF. While image plates were not the sole preserve of macromolecular crystallography, it was felt that there would be maximum synergy with that community, and that the proposed imgCIF dictionary was most naturally viewed as an extension or companion to the mmCIF dictionary. The adoption of DDL2 would not preclude the generation of DDL1 analogue files if they were found to be necessary in certain applications, but to date such a need has not arisen.

However, on one point the workshop was insistent. For efficient handling of large volumes of image data on the necessary timescales within a large synchrotron research facility, the raw data must be in binary format. It was argued that although ASCII encoding could preserve the information content of a data stream in a fully CIF-compliant format, the consequent overheads in increased file size and data-processing time were unacceptable in environments with a very high throughput of such data sets, even given the high performance of modern computers. Consequently, the Crystallographic Binary File (CBF) was born as an extension to CIF (Chapter 2.3). The CBF may contain binary data, and therefore cannot be considered a CIF (Chapter 2.2). However, except for the representation of the image data, the file retains all the other features of CIF. Information about the experimental apparatus, duration, environmental conditions and operating parameters, together with descriptions of the pixel characteristics of individual frames, are all provided in ASCII character fields tagged by data names that are themselves fully defined in a DDL2-based dictionary (Chapter 3.7).

At first sight, the distinction between CIF and the crystallographic binary file may therefore seem trivial. However, in allowing binary data, the CBF requires greater care in defining the structure and packing by octets of the data, and loses some of the portability of CIF. It also precludes the use of many simple text-based file-manipulation tools. On the other hand, the importing of a stable and well developed tagged information format means that developers do not need to write novel parsers and compatibility with CIFs is easily attained. Indeed, the original idea of a fully compliant image CIF has been retained. By ASCII-encoding the image data in a crystallographic binary file, a fully compliant CIF, known as imgCIF, may be simply generated. One could consider an imgCIF as an archival version and a CBF as a working version of the same information set.

1.1.11. Other extension dictionaries

Since the introduction of the major CIF dictionaries, several other compendia of data names suitable for describing different applications or disciplines within crystallography have been developed. Four of these are described in the current volume.

The powder CIF dictionary (pdCIF; Chapter 3.3) is a supplement to the core dictionary addressing the needs of powder diffractionists. The structural model derived from powder work is familiar to single-crystal small-molecule or inorganic scientists. However, the powder CIF effort had the additional goals of documenting and archiving experimental data. It was always intended that powder CIF be used for communication of completed studies and for data exchange between laboratories. This is frequently done at shared diffraction facilities such as neutron and synchrotron sources. The powder dictionary was written with data from conventional X-ray diffractometers and from synchrotron, continuous-wavelength neutron, time-of-flight neutron and energy-dispersive X-ray instruments in mind.

The modulated-structures dictionary (msCIF; Chapter 3.4) is also considered as a supplement to the core dictionary and is designed to permit the description of incommensurately modulated crystal structures. The project was sponsored by the IUCr Commission on Aperiodic Crystals and was developed in parallel with a standard for the reporting of such structures in the literature (Chapuis *et al.*, 1997).

A small dictionary of terms for reporting accurate electron densities in crystals (rhoCIF; Chapter 3.5) has recently been published as a further supplement to the core dictionary. It has been developed under the sponsorship of the IUCr Commission on Charge, Spin and Momentum Densities.

The symmetry dictionary (symCIF; Chapter 3.8) was developed under the direct sponsorship of COMCIFS with the objective of producing a rigorous set of definitions suitable for the description of crystallographic symmetry. Following its publication, several data names from the symCIF dictionary were incorporated in the latest version of the core dictionary to replace the original informal definitions relating to symmetry. The symCIF dictionary contains most of the data names that would be needed to tabulate space-group-symmetry relationships in the manner of *International Tables for Crystallography* Volume A (2002). It is intended to expand the dictionary to include group–subgroup relations in a later version.

Other dictionaries are also under development, often under the supervision of one of the Commissions of the International Union of Crystallography.

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.