

Preface

BY S. R. HALL AND B. McMAHON

Crystallography is a data-intensive discipline. This is true of the measurement and the analytical aspects of crystallographic studies, and so the orderly acquisition and retention of data is of great importance. There is also a need for computational tools that facilitate efficient data-handling processes. In this data-rich environment, *International Tables for Crystallography* provides supporting sources of reference data and guidelines for analysing, interpreting and modelling the data involved in a wide variety of studies.

This volume in the *International Tables* series, *Definition and exchange of crystallographic data*, deals with the precise definition of data items used in crystallography. It focuses on a particular data representation, the Crystallographic Information File (CIF), developed over the last 15 years for the computerized interchange of data important to structural crystallography, and for submissions to crystallographic publications and databases. The need for the volume stems from the IUCr's adoption in 1990 of CIF as the standard for exchanging crystallographic data across the entire community, and for the submission of research articles to *Acta Crystallographica* and other journals. The data items described in this volume appear as standard tags in CIF. Each data item has an associated definition, with attributes that characterize its allowed values and relationships with other defined items. The definitions provide a comprehensive machine-readable description of the information associated with a structural study, from area-detector counts through to the derived molecular geometry, and allow a full annotation suitable for publication as a primary research article.

The volume is arranged in five parts. Part 1 introduces the need and rationale for a standard interchange format, and the IUCr's commissioning and development of CIF in response to that need. Part 2 presents the formal specifications of the file formats involved in the CIF standard, and the related crystallographic binary file (CBF) and molecular information file (MIF) standards. Details are also given of the dictionary definition language (DDL) used to specify the names of data items and their attributes in a machine-readable way. The definitions and attributes of data items in different areas of crystallography are listed in separate dictionary files. Part 3 describes the structure and content of each of the public data dictionaries currently available. It provides commentaries on the dictionaries and guidance on best practice in their use for information interchange in several fields of crystallography. The dictionaries themselves are structured ASCII files that may be used directly by software parsers and validators. Part 4 presents the contents of the dictionaries in a text format that is easier for humans to read. Part 5 describes software applications and libraries that use CIF and related interchange standards.

The data definitions in this volume can be read and used from the printed page, but their principal use is in a computer-software environment. The volume therefore describes the large number of computer programs and libraries contributed by the crystallographic community for handling data files in the CIF format. Many of these programs use the dictionary files directly to validate and exchange data items. Software by its very nature is under continual development, and individual software implementations become obsolete with disturbing rapidity. Nevertheless, we have felt it important and useful to devote considerable space to the software aspects of CIF data exchange. The programs described in the volume illustrate various basic considerations and

approaches to data exchange and provide a rich gallery of tools suitable for different applications. The volume also includes a CD-ROM containing many of these software packages, as well as machine-readable versions of the data dictionaries themselves, and many links to related web resources.

This volume therefore contains material that is relevant to both crystallographers and information-technology specialists. The data dictionaries that form the core of the volume provide the knowledge base for automated validation of data submitted for publication and for deposition in databases (tasks in which most structure-determination scientists are involved). Part 5 will be of particular interest to the non-specialist user of CIF, in that it contains practical advice on stand-alone application programs, the publication of crystal structure reports in primary research journals and the deposition of biological macromolecular structures in the Protein Data Bank. Software authors will benefit from the discussions on designing and implementing CIF-aware programs and from the detailed descriptions of a number of available libraries.

The data-representation and knowledge-management techniques used routinely in crystallography today were developed well in advance of recent approaches to information dissemination for the World Wide Web through open protocols. Nevertheless, CIF processes fit well with the enormous contemporary effort to build the so-called semantic web, in which information and knowledge are transported and interpreted by computer. The success of such knowledge management depends on well structured data vocabularies, data models and data structures appropriate for different subject areas. The term used to denote such semantically rich formulations is 'ontology'. We believe that the formal description of crystallographic data and information presented in this volume is an exemplary ontology in this sense, and will be a useful model for workers in related fields.

The volume has taken several years to complete. It covers subject matter that is relatively new and still changing, and indeed much of the material has been undergoing revision even as it was being compiled. Thanks to the determined effort by all involved in its production, this volume provides a comprehensive and up-to-date compendium of the current state of crystallographic data definition. This first edition also serves the role of providing the historical and scientific context for future developments in a rapidly evolving field. It is likely that new editions will be needed before long to stay abreast of such developments.

Since it began, the CIF project has involved the participation and collaboration of many colleagues across the breadth of crystallographic practice. In the production of this volume, we wish first to thank all the authors of the individual chapters for their major contributions. Each chapter was independently and anonymously peer-reviewed, and the reviewers' comments were forwarded to the authors for response. This has added to the time and effort involved on everyone's part, but we believe that the improved presentation and exposition of the detailed and often complex material will be of significant benefit to readers and users. We are grateful to the reviewers for their dedication and enthusiasm during this process. We particularly appreciate the help of many colleagues who have contributed to the CIF project generally and who have made valuable contributions to this volume. They are recognized as Associate Editors in the list of contributors. In addition to the individuals mentioned by name in individual chapters, we thank Doug du Boulay, Eldon Ulrich,

PREFACE

Frank Allen, Bill Clegg, Mark Spackman, Victor Streltsov, Bob Sweet, Howard Flack, Peter Murray-Rust and John Bollinger for thought-provoking discussions and contributions. We thank Philip Coppens for initially supporting the proposal for this volume, Theo Hahn and Hartmut Fuess, the previous and current Chairmen of the Commission on *International Tables for Crystallography*, and the other members of the Commission who welcomed this volume to the series. Thanks also go to Timo Vaalsta of the University of Western Australia for technical help with the CD-ROM. Finally, we are grateful to Nicola Ashcroft of the IUCr editorial office for her careful and insightful technical editing of this work.