1.1. How to navigate this volume

BY JAMES R. HESTER AND BRIAN MCMAHON

1.1.1. Introduction

This volume of *International Tables for Crystallography* gives a comprehensive account of the Crystallographic Information Framework (CIF) and its applications and related standards, and caters for a wide range of interested readers. To help in finding the information of most use to different classes of reader, we present the following suggested reading strategies. In all cases, a good starting point will be the overview Chapter 2.1. The 'Executive summary' introducing that chapter (Section 2.1.1) is a concise account of essential information about CIF that provides a convenient *aide-memoire* for established users. The remainder of Chapter 2.1 provides a more detailed account of the CIF standards that may be helpful to the reader with little previous knowledge of the subject.

1.1.2. The general reader

The introductory Chapter 1.2 establishes the need for data exchange standards, and relates this to best practice for scientific data management. The history of how the International Union of Crystallography (IUCr) commissioned the CIF project to establish such standards is described in Chapter 8.1.

Chapter 4.1, on the management and use of CIF dictionaries, explains the significance of the data definitions that establish an interoperable ontology. If the general reader is interested in details of the concrete format that is used within crystallography both for data files and dictionaries, Section 2.1.3 is probably sufficient. If further detail on the syntax is desired, it may be found in Chapter 2.2.

Parts 6 and 7 discuss how CIF is used to facilitate publishing of structure report papers and deposition of structures in curated crystallographic databases, respectively.

1.1.3. The structural scientist

The typical researcher who views CIF as a necessary tool for submitting articles or depositing structures in databases might also wish to begin with the introductory Chapter 1.2 to gain a suitable perspective of the benefits of fully-featured data definitions and exchange standards.

Most researchers need not concern themselves unduly with the details of the CIF format, since this is usually catered for by the refinement software they are using, or by post-refinement end-user applications (Chapter 5.6) for editing or visualizing the CIFs they have produced. If they do need to modify CIFs by hand, the specifications in Chapter 2.2 may be consulted.

They may also have limited interest in the dictionary definitions, since again most refinement packages will select the appropriate data names to tag the data that they output. However, if users need to assign values to data items without the benefit of prompts from interactive programs, they may consult individual definitions in the dictionary listings that form Part 4 of this volume; the index of data names at the end of the volume may also be a useful starting point. Where they do need a better understanding of the significance of data names, especially in relation to other data names that might appear in a data file, the explanatory chapters for each dictionary in Part 3 will provide that information.

Part 6 discusses how CIF is used in the publication of chemical (Chapter 6.1) and macromolecular (Chapter 6.2) structures, and in articles describing raw data sets (Chapter 6.3). These chapters are particularly useful for understanding the validation criteria that may be applied to different types of structures submitted for publication.

The handling of structures deposited in chemical (Chapter 7.1) and macromolecular (Chapter 7.2) structural databases will also be of interest.

1.1.4. The software developer

Programmers wishing to write software that handles crystallographic data in CIF format should begin with the discussion of general principles in Chapter 5.1. They should also read the relevant format specification in Chapters 2.2 (CIF versions 1.1 and 2.0) and 2.3 (imgCIF/CBF).

1.1.4.1. Refinement or data processing packages

Many developers of crystallographic software are focused on data processing and analysis problems, and are interested in CIF only as an input/output channel for passing data to and from an application. They may find that one of the standard libraries discussed in Chapter 5.3 will provide all the support needed for this purpose. Otherwise, they may find it helpful to read the description of the CIF application programming interface (Chapter 5.2), either to use directly the reference implementation (although this is not optimized for performance), or to design their own API on similar principles.

Writing native code to output CIF data is relatively straightforward, since the programmer may choose the order in which data may be written, and has few layout constraints. Attention to the details of the format specification should ensure that syntactically correct CIFs are formed. The programmer must of course take care to ensure that the correct data names are used, and that their associated values conform to the restrictions detailed in the relevant dictionary. The dictionary listings in Part 4 should provide enough information for this, though the matching commentary chapter in Part 3 should also be read to ensure correct usage. If the programmer wishes to reduce the burden of conformance to dictionary attribute constraints, it is possible to write routines to validate directly against the dictionaries. In this case, an understanding of the DDL in which the relevant dictionary is written can be gained from Chapter 2.4. To implement any dictionary methods written in the dictionary relational language (dREL), one should also read Chapter 2.5.

If it is wished to archive in a CIF file some data that are not characterized by existing dictionary definitions, new data names may be created, provided they are differentiated from the existing

Affiliations: JAMES R. HESTER, Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC, NSW 2232, Australia; BRIAN MCMAHON, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK.

definitions by use of a `[local]` or otherwise registered prefix, as discussed in Section 2.1.4.2.

Handling input CIF data is more complex, because of the free-form layout and ordering of data in an input file. If a suitable library is not used, the programmer may wish to use a filter program to preprocess the input into a normalized presentation that is easier to write native code to handle. Some programs to do this are discussed in Chapter 5.4.

1.1.4.2. General utility programs

Some developers may wish to write generic programs to reorder, validate or otherwise transform CIFs (*e.g* to convert to a different format). Such developers will need a detailed understanding of the relevant format specifications (Chapters 2.2, 2.3) and might also benefit from studying how edge cases are handled, for example by the CIF API (Chapter 5.2). If the goal is to transform CIF data into another syntactic representation, Chapter 2.6 might also provide some useful ideas.

Developers of applications that validate against dictionaries, or transform data based on relationships in the dictionaries (*e.g* to separate or merge data values and their standard uncertainties, or to convert matrices into lists of their individual components) should understand the DDL specifications in Chapter 2.4. They may also be interested in implementing relational methods expressed in the dREL language (Chapter 2.5).

1.1.4.3. System developer

In this context, a system developer is one who aims to provide an end-to-end workflow, with tools that can take full advantage of all existing CIF features. Such an ambitious programmer should read all the specification chapters of Part 2, and the introductory discussions of general principles in programming (Chapter 5.1), dictionary construction (Chapter 3.1) and dictionary maintenance (Chapter 4.1).

Plans to develop re-usable application programming interfaces or software libraries should be informed by consideration of the CIF API (Chapter 5.2) or existing libraries (Chapter 5.3). It may also be useful to read the other chapters in Part 5 to identify existing programs that may be incorporated into the developing pipeline, ported to a more convenient programming language, or that may provide ideas on end-user function and usability.

1.1.5. The database manager

Databases may be required to ingest CIF data from a variety of sources, and so must be able to handle input conformant with any format specification they support (Chapters 2.2, 2.3). They may wish to incorporate existing filter programs into their workflow (Chapter 5.4), but they may also need to write a complete system to handle ingest, validation and database loading. In such a case, they may wish to build their own application programming interface, perhaps informed by the approach of the reference CIF API (Chapter 5.2) or by some of the design decisions of existing CIF libraries (Chapter 5.3).

Validation programs may be built to check all the constraints and relations expressed in dictionaries through the DDL (Chapter 2.4) and dREL (Chapter 2.5) languages. For correct interpretation of the ingested data, the definitions presented in all the relevant dictionaries (Part 4) must be studied. Data for a complex structure or system may be spread across several data blocks within the same file, and Chapters 3.3 and 4.3 provide guidance on how this might be presented.

Approaches to validation and, indeed, overall design and control of workflows of existing databases are discussed in depth in the chapters of Part 7.

1.1.6. The non-crystallographer

Here we mean the software developer charged with implementing a pathway between crystallographic data and some other scientific domain with its own data representation standards. The introductory Chapter 1.2 will be useful in contextualizing the need for interoperable data management protocols, and in emphasizing the need to work within the FAIR principles (Wilkinson *et al.*, 2016).

A detailed understanding of the relevant CIF format specifications is needed (Chapters 2.2, 2.3), and some study of existing alternative syntaxes may be useful (Chapter 2.6).

The most significant task, however, is a detailed understanding of the semantic content associated with each data name. Here the developer must understand the formalism in which the CIF dictionaries are constructed (DDL, Chapter 2.4), and must carefully study the definition and attributes of any data items that will participate in the inter-domain traffic. Chapter 4.1 provides an overview of the entire CIF ontology, while the individual dictionaries contributing to this are presented in the remaining chapters of Part 4. Although the developer may not be interested in the scientific relevance of individual data items, it may still be useful to read the commentaries on each dictionary in Part 3 in order to understand restrictions on how they may be used (as well as amplifying some of the dictionary definitions for the benefit of the non-specialist).

The discussion of NeXus/HDF5 in Chapter 2.7 may be a useful case study of the interaction of crystallographic information with a more general multi-discipline standards and software environment.

1.1.7. The dictionary developer

The dictionary developer should have a reasonable understanding of the dictionary file format (Chapter 2.2) and dictionary definition languages (DDL, Chapter 2.4; dREL, Chapter 2.5) that define the dictionary formalism. When contributing to COMCIFS-managed dictionaries, familiarity with the layout style rules (Section 2.4.2.6) is also desirable.

Chapter 3.1 provides a detailed account of the general principles of dictionary writing, applicable both to local dictionaries and those intended for wider distribution under the aegis of the IUCr or wwPDB. Chapter 4.1 provides more details of integrating dictionaries into the ecosystem of canonical dictionaries managed for the IUCr by COMCIFS. This chapter also provides an overview of existing dictionaries, and how they compose an overall crystallographic ontology.

Chapter 3.3 and its companion dictionary Chapter 4.3 are important for understanding complex structures or systems that are described across several distinct data blocks within the same file.

Developers extending an existing dictionary will of course familiarize themselves with the existing dictionary content in Part 4, as well as the detailed commentary on each to be found in Part 3. Developers of new dictionaries may find useful the sections describing the design decisions behind each existing dictionary, also discussed in the chapters of Part 3.

References

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J. *et al.* (2016). *The FAIR Guiding Principles for scientific data management and stewardship*. *Sci. Data*, **3**, 160018. https://doi.org/10.1038/sdata.2016.18.