The Macromolecular CIF Dictionary

Paula Fitzgerald¹, Helen Berman², Phil Bourne³, Brian McMahon⁴, Keith Watenpaugh⁵ and John Westbrook²

¹Merck Research Laboratories, PO Box 2000 Ry50-105, Rahway NJ 07065 USA
²Department of Chemistry, Rutgers University, PO Box 939, Piscataway NJ 08855 USA
³San Diego Supercomputer Center, PO Box 85608, San Diego CA 92186 USA
⁴The International Union of Crystallography, 5 Abby Squarer, Chester CH11 2HU UK
⁵Pharmacia Upjohn, 7255-209-1, 301 Henrietta Street, Kalamazoo MI 49007 USA

Background

'CIF' (Crystallographic Information File) is a subset of STAR (Self-defining Test Archive and Retrieval format [1]). The CIF format is suitable for archiving, in any order, all types of text and numerical data. The goals of

CIF are a data structure that is general, upwardly compatible, flexible, and which facilitates electronic publication.

CIF was developed by the IUCr Working Party on Crystallographic Information in an effort sponsored by the IUCr Commission on Crystallographic Data and the IUCr Commission on Journals. The results of this effort were seen in a dictionary of data items sufficient for archiving the small molecule crystallographic experiment and its results[2]. This dictionary was adopted by the IUCr at its 1990 Congress in Bordeaux. CIF is now the format in which structure papers are submitted to Acta Crystallographica C; software has been developed to automatically typeset a paper from a CIF.

In 1990, the IUCr formed a working group that would expand this dictionary by including data items relevant to the macromolecular crystallographic experiment. This working group was chaired by Paula Fitzgerald (Merck) and included Enrique Abola (Protein Data Bank), Helen Berman (Rutgers), Phil Bourne (then at Columbia) Eleanor Dodson (York), Art Olson (Scripps), Wolfgang Steigemann (Martinsreid), Lynn Ten Eyck (SDSC), and Keith Watenpaugh (then Upjohn).

The original short term goal of the working group was to fulfill the mandate set by the IUCr: to define CIF data names that needed to be included in the CIF dictionary in order to adequately describe the macromolecular crystallographic experiment and its results. Long term goals were also established: to provide sufficient data names so that the experimental section of a structure paper could be written automatically and to facilitate the development of tools so that computer programs could easily interface with mmCIF. During the course of the development of the mmCIF dictionary, however, these goals were greatly expanded, and the resulting dictionary can now be thought of as a flat-file representation of a fully-relational database schema describing the complete macromolecular cryst-allographic experiment and its results.

The mmCIF Workshops

In order to describe the progress of this project and to solicit community feedback, several informal and formal meetings were held. The first meeting, hosted by Eleanor Dodson, convened in April 1993 at the University of York. The attendees included the mmCIF working group, structural biologists and computer scientists. A major focus of the discussion was whether the formal structure of the dictionary that was implemented using Dictionary Definition Language (DDL 1.0) was adequate to deal with the complexity of the structural data items. Criticisms included the idea that the data typing was not strong enough and that there were no formal links among the data items. A new working group was formed to try to address these issues. The second Workshop was hosted by Phil Bourne in Tarrytown, NY, in October 1993. The topics at that meeting focused on the development of software tools and the DDL. In October 1994, a workshop hosted by Shoshana Wodak at the Free University of Brussels resulted in the development of a new DDL that addressed the various problems that had been identified. Following the Brussels meetings, the mmCIF dictionary (including a complete image of the CIF core dictionary) was recast in DDL 2.1.

Community Review

The mmCIF dictionary has continued to grow and be refined during the several years of its development, originally based on input from the working group, and subsequently based on input garnered at the three CIF workshops. By mid-1995, a version of the mmCIF dictionary that was considered complete in most regards was in hand, and that dictionary was presented to the community at large for review at the 1995 ACA Meeting in Montreal.

The review was (and still is) managed via a Web page and a mailing list. The Web page (http://ndbserver-rutgers.edu/mmcif) contains copies of the dictionary (as plain text and as an HTML Web-searchable version), as well as background material, examples of mmCIF files, and archives of the discussions on the mmCIF mailing list. The Web page also contains information on the DDL, and access to a number of mmCIF software tools.

The mailing list is used for posting comments from the community, suggestions for changes, errata and such. To subscribe to the mailing list, send a one-line message containing the text "subscribe mmciflist Your Name" to

requests@ndbserver.rutgers.edu. To post to the mailing list, send messages to mmciflist@ndbserver.rutgers.edu.

The review process was an active one, with a large number of people taking a close look at the dictionary and making very useful comments, corrections, and suggestions for additional data items. The New Jersey contingent of the working party met regularly, discussed responses to each of the issues that were raised on the mailing list, and made changes to the dictionary based on the results of those discussions. Updated versions of the dictionary were then posted on the Web page.

By late winter of 1996, we felt that the dictionary had assumed its final form, and we posted announcements about the mmCIF dictionary and its availability to a number of widely-read crystallographic newsgroups. These announcements have generated a small number of rather minor corrections and additions to the dictionary.

Final Approval

Following the IUCr meeting in Seattle, Version 1.0 of the dictionary will be released. There are still a number of wording changes that will need to be made to mmCIF dictionary definitions to bring them into

alignment with the newly revised version of the CIF core dictionary, but we DO NOT ANTICIPATE ANY FURTHER REVISIONS OF SUBSTANCE to the mmCIF dictionary. In particular, the ATOM SITE records, the heart and soul of the dictionary, will not be modified. We thus encourage users of all types, including software

developers, to begin working with the dictionary. We anticipate that as people begin to really use the mmCIF data structure, they will find further data items that they would like to see included, but only those data items that constitute obvious omissions to the current schema will be added to Version 1.0; true expansions of the data structure will be deferred to the eventual Version 2.0.

Acknowledgments

The development of the mmCIF dictionary and the associated DDL 2.2.1 has been an enormous task, and any list of contributors to the effort will certainly be incomplete. Still, we have so appreciated the people that have taken the time to think carefully and constructively about all of this, and we would like to recognize their efforts. But we must begin by recognizing Syd Hall, David Brown and Frank Allen, who began the entire CIF effort and who recruited us to do the extensions for macromolecular structure.

The background given above lists people who were members of the original working party, but the number of people who contributed to the original design of the mmCIF data structure is in fact much larger. We would like to thank Steve Bryant (NCBI), Vivian Stojanoff (PDB), Jean Richelle (Brussels), Eldon Ulrich (Madison), and Brian Toby (NIST).

There are also the people who realized the shortcomings of the original DDL and worked hard to convince us that a more rigorous underpinning for the dictionary would been needed. Their suggestions (and pointed criticisms) resulted in the development and implementation of DDL 2.1. Out thanks go out to Michael Scharf (EMBL), Peter Grey (Aberdeen), Peter Murray-Rust (Glaxo), Dave Stampf (PDB), and Jan Zelinka (York).

Writing the dictionary and developing the new DDL were just the starting points for evaluation and critique, and this effort has been greatly aided by the input from COMCIFs, the IUCr committee with oversight over this process (Brian McMahon, Coordinating Secretary). But the real process of review, after the dictionary was released to the public for comment in August 1995, has involved a much larger cast. We cannot say enough about the valuable input we have gotten from Fran Bernstein (PDB), Herb Bernstein (BNL), Dale Tronrud (Oregon), and Peter Keller (Daresbury).

Our efforts has been greatly enabled by the staff of the Nucleic Acid Database at Rutgers University, who have dealt with many of the technical issues of implementation of mmCIF with real data. So we would also like to thank Anke Gelbin. Shu-Hsin Hsieh. and Christine Zardecki.

Without the three CIF workshops, this effort would never have taken the shape and focus it now has, and we are eternally gratefully to the organizers of those workshops - Eleanor Dodson, Phil Bourne, and Shoshana Wodak - and to the sponsors who provided the funding - ESF, EU, NSF, and DOE.

References

[1] S.R. Hall (1991) J. of Chemical Information and Computer Science, 31, 326-333.

[2] S.R. Hall, F.H. Allen and I.D. Brown (1991) Acta Cryst., A47, 655-685.