



**Quarterly Progress Report from  
Dowling College, Oakdale, NY, USA  
to the International Union of Crystallography, Chester, England  
concerning new extensions to the capabilities of CIF for IUCr journals**

1 February 2008

The Executive Secretary of the IUCr  
Mr. M. H. Dacombe  
International Union of Crystallography  
2 Abbey Square  
CHESTER CH1 2HU  
England

This is the second quarterly progress report on the IUCr funded project at Dowling College to support the evolving needs of the community for new and upgraded CIF software to facilitate publication in IUCr journals. Additional information will be available on the project website:

<http://arcib.dowling.edu/cifiucr>

a few days after this report.

### **I. Project Summary**

Dowling College is providing to the IUCr the services of Professor Herbert J. Bernstein as project director (PI/PD) and certain of his students to modify existing software and to create new software in order to support the evolving needs of the community for new and upgraded CIF software to facilitate publication in IUCr journals.

As versions of these packages mature they will be released to the community as open source software without charge to encourage wide use. The software will be released using the GNU GPL or a similar license. "CIF Applications" articles will be submitted to help make the community aware of these new and upgraded tools, and the IUCr will be given first refusal in publication of such articles produced from the work of this project.

### **II. Description of goods and services to be delivered to the IUCr**

New extensions to the capabilities of CIF embodied in the work of Hall et al. on DDLm will necessitate revisions to the software used by the IUCr in the publication of journals. We have started this effort with with two essential subprojects:

1. Creation of Xchek2 based on Xchek and Cyclops. We are creating a new checking program that is aware of and able to use the methods-based checking provided in DDLm, with improved capabilities, such as validating cell volumes against edges and angles.
2. Preliminary adaptation of vcif2 and the test suite to DDLm and dREL. This will augment the test suite we did in the prior funding period to support the new methods-based checking.

### **III. Timetable**

The agreement started on 1 August 2007. The Agreement will terminate when the work is complete and this will be no later than 31 July 2009.

#### IV. 1 February 2008 Status

**Overview:** As noted in the prior quarterly report, “The project was funded by the IUCr in August 2007 and has been started. The initial base for the project was the documentation and software provided by Syd Hall directly and via the IUCr web site, as well as the work done on CIF validation using CBFlib in the prior 2004-2006 funding period, and the work done subsequently on CBFlib. The work draws on and intersects with other work funded by the U.S. Department of Energy, the U.S. National Institutes of Health, the U. S. National Science foundation and Dowling College. Of particular note is the work on the image-supporting CIF dictionary (imgCIF), the related work on the CBFlib API for imgCIF, the work in integration of CBFlib with NeXus, and the creation of a Structural Biology Extensible Visualization Scripting Language (SBEVSL). There are two major threads of software development for the project: one starting from Xchek and Cyclops in Fortran and one starting from the vcif2 code in CBFlib in C. In addition, DDLm requires work in java/python/jython.” In this quarter we have worked on CBFlib and CIFtbx, the C and Fortran bases for the project and have upgraded some of the hardware infrastructure for the project.

**Staffing:** The PI/PD is Professor Herbert J. Bernstein. The work reported for this period was done by G. Todorov, D. O’Brien and H. J. Bernstein.

**Funding and Administration:** Cash flows and burn rates have been appropriate to the needs of the project.

**Project Activities:** The activities for the project in this quarter have consisted of a cleanup of the base level of CBFlib (to be formally released shortly as CBFlib\_0.7.9), and a start on the changes to the CBFlib C-based parser and to the ciftbx Fortran based parser. The major work has been on the design of the logic to handle parsing the “list”, “array”, “tuple” and “table” entities, and identification of most of the mappings from DDLm to DDL2 needed to create hooks to the existing validation code. In addition the main server for the software development activities in the lab (arcib) was upgraded, providing a more stable development environment with more disk space.

**Creation of Xchek2 based on Xchek and Cyclops:** In order to validate CIFs against DDLm dictionaries, we need to parse those dictionaries. Xchek has a minimal partial parse of DDLm, but a full check is best done using a full parse. To upgrade from the existing DDL1 and DDL2 parsers, code has to be added to handle the new syntax for values, which, in addition to the existing unquoted words, single and double quoted strings and “\n,” quoted texts fields in DDL1 and DDL2, includes the new “{ item, item, ...}”, “[ item, item, ...]”, and “( item, item, ...)” constructs, allowing for the possibility that these mechanisms may be nested, requiring some uses of stacks to save state. We are adding new code and variables to CIFtbx to support these new constructs. This raises an interesting issue in the handling of existing DDL1 and DDL2 CIFs – should the new constructs be allowed in parsing such CIFs. Allowing them helps to encourage the community to move up to DDLm, but weakens strict validation of CIFs for compliance with DDL1 and DDL2. The compromise we propose is to make acceptance of the new constructs the default and to add a control flag to allow applications to revert to strict checking against DDL1 and DDL2.

We hope to have the first pass of these parser upgrades to ciftbx for these constructs ready for distribution by mid-February and the upgrade to CBFlib ready for distribution by the end of February. The code will be posted on the [blondie.dowling.edu](http://blondie.dowling.edu), our GFORGE server.

Work by H. J. Bernstein and G. Todorov.

**Preliminary adaptation of vcif2 and the test suite to DDLm and dREL:** Xcheck and Cyclops prepare lists of names from the dictionaries and compare them to the names in documents. In the course of doing that the syntax of the dictionaries are checked, but not the syntax of the document. In vcif3 (vcif2 adapted to DDLm and dREL) we will be checking the syntax of the document as well. Most of the checking is conceptually

identical to the vcif2 checking, but the information is conveyed by different tags or in a slightly different way. The major significant difference is that DDLm and dREL allow methods to algorithmically state the relationships among values of different tags. This last aspect is most easily handled in a C-like context, such as CBFLib (the base for the current vcif2). This requires a rework for CBFLib not only of the parser (a task similar to what is being done to the cifbx parser) but of the data structures to support efficient access to the elements of lists, arrays, tuples and tables. We have begun the parser changes. The data structure and method interpretation changes will almost certainly extend into the second year of the project. In this quarter the current vcif2 has been more fully documented and the design of the new data structures began.

Work by H. J. Bernstein and G. Todorov.

**Infrastructure Upgrade:** In this quarter the main development machine for our lab, arcib.dowling.edu was replaced with a faster, more reliable machine with more disk space.

Work by G. Todorov and D. O'Brien.

**Summary:** The project is on track.

Respectfully Submitted,

Herbert J. Bernstein  
Professor of Computer Science

Cc: Brian McMahon  
Arthur Perri  
Erik Paulson