



**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**

2 August 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 fourth 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. 2 August 2008 Status

**Overview:** The major activity in this quarter was on both the CBFlib and CIFtbx infrastructure for validation of data values against the new DDLm data types.

**Staffing:** The PI/PD is Professor Herbert J. Bernstein. The work reported for this period was done by G. Todorov, E. Zlateva, N. Darakev, G. Darakev and H. J. Bernstein. Mr. Todorov graduated shortly after the start of this quarter. G. Darakev and the PI will both be at the IUCr Congress in Osaka in August 2008.

**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 new code in CBFlib and CIFtbx preparing for release 0.8 of CBFlib and release 4 of CIFtbx, both with code to read and validate data items against expanded DDLm dictionaries. The necessary code infrastructure to parse bracketed items and to test data items against the DDLm regular expressions is now in CBFlib. The regular expressions themselves are being validated and adjusted as necessary and CIFTEST2 is being expanded case by case to provide test cases for likely incorrect uses of the new DDLm types. The current pre-release of CBFlib 0.8 in the repository on [blondie.dowling.edu](http://blondie.dowling.edu) is also being used as the CIF processing component in the current pre-release of RasMol to help ensure adequate testing against existing CIF data. We will post as complete a version of CBFlib 0.8 to the web as is ready just before the Osaka IUCr meeting later this month. The CIFtbx infrastructure to parse bracketed items and to test data items is also ready, but the testing will have to be at a simpler level than in the C-version, not using regular expressions, but simply testing on a coarser level similar to the one adopted for testing against the DDL2 data types. That code will also be posted to the web before the IUCr meeting.

The work on CIFtbx4 has raised an interesting issue of how not to break existing Fortran applications in the transition to DDLm. The problem is that older Fortran applications cannot accept strings of arbitrary length. To deal with this, we are following the approach already used in CIFtbx to deal with text fields – delivering the text in chunks of limited length with a flag set true if there are more chunks to be examined. This will allow existing applications to view bracketed constructs as if they were text fields presenting one item per line. For applications that can be converted to be DDLm aware, new variables giving the depth into a bracketed construct and the index across on the current level should provide appropriate control.

**Creation of Xchek2 based on Xchek and Cyclops:** In the prior report, we wrote, “The lessons learned in the coding for C have caused us to rethink the code currently in Fortran. In the past we have preserved all comments in a CIF while doing the Fortran parse, so that the original CIF with all comments can be recreated even when reformatted. As we have discovered in the C code, it is important to strip the embedded comments in the bracketed constructs, and it also may be necessary to have a full tree-expansion of nested bracketed constructs. Maintaining a full tree structure and three-fold replication of all the bracketed constructs is workable in C and even in Fortran-95, but is a non-trivial change in Fortran-77 if reasonable performance is to be achieved. We are exploring alternatives and will resolve the matter in the next quarter.” We have explored the alternatives, which for Fortran-77 would require extending the current use of direct access files to store the tree. The performance hit was too great and we plan not to bring the full tree-structure into the Fortran version, but to stay with the less-demanding event-based logic discussed above. We should discuss this in Osaka.

Work by H. J. Bernstein

#### **Preliminary adaptation of vcif2 and the test suite to DDLm and dREL:**

In this quarter the paper on `vcif2` was accepted by the Journal of Applied Crystallography [G. Todorov and H. J. Bernstein, “VCIF2: extended CIF validation software,” J. Appl. Cryst. (2008). 41, 808-810]. Checking of data values against DDLm dictionaries in CBFlib began. New test cases have been added to CIFTEST2. The

external specifications of the methods checking are to be discussed at the COMCIFS meetings in Osaka, and we expect to implement what is decided during the coming quarter.

Work by H. J. Bernstein, G. Todorov, E. Zlateva, N. Darakev.

**Infrastructure Issues:** In this quarter the remediation of hardware problems that were addressed in the previous quarter were completed. The backup server was replaced with a system with 3 TB of storage. In the previous quarter the CBFlib CVS was replicated from the GFORGE server to sourceforge in the cbflib project, and that sourceforge CBFlib is now heavily used. As noted in the prior report. By the time of the Osaka meeting we expect to complete the move of the code and web pages for this project to sourceforge for the convenience of the community. The primary development activities will continue on the GFORGE server.

Work by H. J. Bernstein, G. Todorov and N. Darakev

**A note on the DDLm import logic and dictionary layering:** There will be a discussion of DDLm import logic and dictionary layering at the Osaka meeting. The view has been presented that dictionaries should be handled in small segments with multi-pass real-time web-loading of multiple dictionaries to compose on virtual dictionary. An alternative approach is for applications to use complete, "expanded" dictionaries. There are problems with both approaches. In the first case, there are serious platform and performance issues. In the second case, there are serious issues of dictionary synchronization. In this project we are addressing these issues in a modular way. We have started modifications to the open source web-page mirroring program wget to allow it to handle web-based dictionary caching into local mirrors. We are adopting the logic in the original Xcheck to then work from local-file dictionary mirrors to local expanded dictionaries. This should allow appropriate local choices in balancing the performance and synchronization issues and should make it much easier to support a wide range of platforms. We will discuss this further in Osaka.

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

Respectfully Submitted,

Herbert J. Bernstein  
Professor of Computer Science

Cc: Brian McMahon  
Arthur Perri  
Erik Paulson