



**Report from
Dowling College, Oakdale, NY, USA
to the International Union of Crystallography, Chester, England
concerning provision of new and upgraded CIF software to facilitate publication in
IUCr journals**

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

This is the first bimonthly 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.

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 as detailed in section I below in order to support the evolving needs of the community for new and upgraded CIF software to facilitate publication in IUCr journals. This is a major set of inter-related projects, expected to take approximately 2 years to complete. However, a phased release to the Chester office of the IUCr of partial preliminary versions of all of these packages would be expected before the end of the first year, and feedback from the Chester office of the IUCr would be used to guide completion of the packages.

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

These include:

- (1) CIFTEST – Creation of a new and extended test file suite and test protocol for validation of CIF parsers.
- (2) vcif2 – an extended version of vcif (a program used to validate the syntax of CIF files). vcif2 will verify compliance with CIF 1.0, CIF 1.1 or mmCIF file formats. vcif2 will also accept an arbitrary list of layered DDL1 and DDL2 dictionaries against which units, enumerations, parent–child relationships and category integrity will be checked.
- (3) CIFFOLD – a new utility to fold and unfold long-line CIFs to allow existing CIF 1.0 applications to work with CIF 1.1 files.
- (4) CIFtbx3 – a new release of CIFtbx to provide support for CIF 1.1 (as well as CIF 1.0 and mmCIF) for Fortran applications and to provide extended integrity checking comparable to that in vcif2.

III. Timetable

The agreement started 1 August 2004. The Agreement will terminate when the work under Item I is complete and this will be no later than 31 July 2006.

IV. 1 October 2004 Status

Overview: The project started on time on 1 August 2004. The project is fully staffed and operational and coding has begun with code for line folding and unfolding (CIFFOLD), and generation of new test cases (CIFTEST). The first code for folding is being tested and should be ready for release along with the test cases for line folding before the next bimonthly report. The code for unfolding has been started.

Staffing: The PI/PD is Professor Herbert J. Bernstein. The students working on the project are Stavros Louris and Kostadin Mitev.

Funding and Administration: Initial funding has been received, and the necessary budget has been established.

Project Activities: The initial software focus is on the line length and line-folding line unfolding issues. Prof. Bernstein is working on the buffer size changes and related coding changes in CIFtbx. Mr. Mitev has produced initial draft filters for line folding and unfolding as a test-bed of heuristics and desirable user options for line folding. Mr. Louris made a first pass at extending the line length in vcif and is working on the test cases for Mr. Mitev's code.

The folding function of CIFFOLD is currently working. The folding procedure determines on-the-fly how to handle the data from the file. There are two nested loops - the first reads a line at a time, the second runs through each character. In the second loop there are if else statements to decide how the data should be handled. The folding logic recognizes loops, tags, text fields, comments and regular data. The formatting is a mixture of the format used in the original file and the most common format used for generating CIF files in human readable format (each non-looped tag and its corresponding data on one line if possible and each reserved word to be on a single line).

Through other funding, the software infrastructure for web site distribution of multiple versions of this software is being prepared. These changes should be ready as this first software development pass matures over the next several weeks and will be used to make the first release of CIFFOLD available for external testing and comments.

Summary: The project is off to a good start. While there is a great deal of additional work to be done, the progress on line-folding provides a good base for further progress.

Respectfully Submitted,

Herbert J. Bernstein
Professor of Computer Science

Cc: Brian McMahon
Dawn Pierpoint