

Driving universal data format input and translation using CIF dictionaries

James Reginald Hester¹

¹*Acns, Ansto, Kirrawee Dc, Australia*

E-mail: jxh@ansto.gov.au

Dealing with multiple file formats is an inescapable part of scientific data handling. Scientific facts do not change when expressed using a different format; consequently, programmers are able to write separate input routines for each file format, extracting the useful information into uniform data structures for further processing. However, in this case information about the relationship between items in different data specifications is usually buried in code and must be redetermined from scratch by other software projects. The alternative approach presented here encapsulates the relationships between datanames from overlapping data specifications in plain text CIF DDLm dictionaries. A "format adapter" with a simple and uniform interface is responsible for reading and writing dataname values in each format, and the dictionaries describe the sometimes complex value transformations and restrictions in machine-readable form using dREL[1]. This system can be used both for universal file input and datafile interconversion. A working example implementation capable of reading and interconverting imgCIF and NeXus nxMX files, and easily expanded to include simpler image file formats, has been developed [2]. CIF dictionaries are used to describe the semantic content of data files in arbitrary formats, and there is no requirement that either the input or output data files are in CIF format.

[1] Spadaccini, N., Castleden, I. R., du Boulay, D. and Hall, S.R. (2012) *J. Chem. Inf. Model.* 52 pp 1917-1925

[2] Hester, J. (2016). [jamesrhester/PyFormatTransformer 0.9](https://github.com/jamesrhester/PyFormatTransformer)

Zenodo. <http://doi.org/10.5281/zenodo.154459>

Keywords: [CIF, data format conversion](#)