Accessing (raw) chemical data: a peek into the CIF format.

Henry Rzepa

Published July 21, 2017

Citation

Rzepa, H. (2017). Accessing (raw) chemical data: a peek into the CIF format. *Henry Rzepa's Blog*. https://doi.org/10.59350/h00r6-y7f18

Keywords

Chemical IT

There is much focus at the moment on how to ensure experimental replicability in *e.g.* the molecular sciences. An important aspect of that is having access to **FAIR** data; data which is findable, accessible, inter-operable and re-usable. One of the "gold standards" in chemistry is the data associated with crystal structures. Here I take an inside peek into the standard file-type for carrying crystal structure data, the CIF file (the Crystallographic Information File).

CIF is a tightly managed format, with utility tools such as checkCIF to validate the files and check for errors. It is also what is called a processed data format, created from structural analysis of the raw image data that emerges from a diffractometer, and is therefore what might be described as a lossy format. Discussing these aspects with our crystallographer here (thanks Andrew!), I began to realise that there are at least three distinctly different versions of a CIF file, each carrying a different degree of data loss.

I am going to take as my illustration of this structure[cite]10.1039/C6DT03810G[/cite] known by three different identifiers; AZUJOW, CCDC 1406199 or DOI: 10.5517/ccdc.csd.cc1j6888

1. The CIF originates with the authors and this version is 449KB in size. I have deposited it and the other two at DOI: 10.14469/hpc/2752 for you to inspect and compare them. This file is relatively large since it contains the so-called structure factors or *hk*l information, a snippet of which looks like:

```
_shelx_hkl_file
;
   0   0   1 108882. 1066.19   2
   0   0   2 320.055 130.609   2
   0   0   3 18538.0 806.608   2
   0   0   4 173192. 2808.03   2
```

2. This information is removed using a utility known as shredcif to produce a second version, known as the name_x.cif version and reducing the size to 27KB. This retains information about properties such as thermal ellipsoids and bond length and angle information but loses the *hkl* information.

3. After the CIF is submitted to CSD, it emerges as AZUJOW.cif, which is now just 7KB in size and is now missing the bond lengths and angles etc.

The original raw image data for this structure is not publicly available, but you can see a set of structures for which it IS available at DOI:10.14469/hpc/2297 (published as [cite]10.1021/acsomega.7b00482[/cite] and where the file sizes are typically 200-600 MB (they can get much larger).

So a CIF can vary in data content between 7- 449KB, and the original "raw" data can be ten thousand times larger than this! To acquire all the flavours, you have to access both the CSD and contact the original authors (unless of course the latter have deposited their versions in an open data repository, as above).

Fortunately for most chemical applications, even the "lossiest" of the CIF formats is more than adequate. But for the gold standard in chemical data, you should be aware that you may still be losing access to a lot of original data in the CIF formats and of course to all of the raw diffractometer data. I think it fair to say however that there is now momentum to increasingly retain as much of this data as is possible for posterity.

https://orcid.org/0000-0002-8635-8390