Data Standardization and Exchange in Macromolecule Crystallography

Ethan Holleman

November 10, 2020

Abstract

I spent the summer of 2020 developing Polo: A graphical user interface for high-throughput crystallography, the vast majority of work for which was done over a period of about 12 weeks. Over that short time I encountered a number of problems but one that stood out to me was the lack of consensus regarding digital representations of crystallization experiments and outcomes. Here, I express a few of my thoughts on the topic.

1 Where is Crystalization's pdf File?

This passed summer, I spent much of my quarantine time working as a summer intern in the BioXFEL internship program. My goal was to build a side project I had been working on previously, a graphical user interface integrating automated image classification algorithms, to fruition. If you are interested in seeing the results you can download the program from the GitHub page. I got to work under the supervision of Dr. Sarah Bowman, the director of the high-throughput crystallography center at Haputman-Woodward Medical Research Institute.

One of the first challenges I faced was just writing code that was robust enough to parse all the different file formats that translate what the lab was doing day-to-day for its users to a digital record. This included csv files of chemical data, xml files describing screens, and of course image files. Behind the schemes of Polo there are hundreds of lines of code that gather this information, connect it, and then present it to the user in a way that makes it easy to assume everything is tightly woven together when the bit-level reality is that this is mostly an illusion. An organizational apparition that evaporates into the RAM ether once you close the program.

To make my own life easier and allow Polo users to save their work in an exchangeable way I began looking for accepted file standards used in the crystallization side of X-ray crystallography but didn't find much outside of what the lab was already going, at least not anything that brought everything I wanted together into the same place. I specially wanted a format that would

- Directly associate image data to classifications and other metadata
- Describe chemical information in an easily parasable format
- Reflect the design of a typical crystallization experiment
- Be implementable into my program in a week or less

The last item is not longer a constraint but I included it since it weighed significantly the direction I actually went in. The best solution I found, and arguably the best solution currently exiting was the json formatted API responses from HWI's own Xtution database. While Xtution is a tremendous resource, I needed a slightly more integrated solution and ended up putting together (some would call hacking) a custom json format that is precipitated directly from the data structures Polo holds in memory. This had the advantages of working and being fast to implement (see item 4 above) but created a file that is not easily generalization beyond its use in my own program.

My thought during and increasingly so now as the time crunch of my ten-week internship is no longer a driving factor of my decision making, is where is the crystallization pdf? The format that I can just open up and know what I am going to expect without having to read documentation or do any guesswork? The diffraction side of crystallography seems to be way ahead and has been for a long time.

- Protein Data Bank format (.pdb), introduced 1977 [1]
- Collaborative Computational Project Number 4 format (.ccp4), introduced 1979
- Crystallographic Information File (.cif), introduced 1991 [2]

My theory is that traditionally, crystallization experiments are small and infrequent enough where you can generally get away using a spreadsheet or just making a table in your lab notebook. The problems with these methods are increasingly revealed as scale and frequency increase, *i.e* the high-throughput crystallography setting. However the relative small-scale of this data does not confine the problems it creates to this same domain. The main problem I see that is created by a lack of standardization is that it forces the development of custom "in-house" solutions (to which to some degree, Polo is apart of). Researchers not seeing an obvious solution follow what they believe to make the most sense at the time and for their laboratory. This could mean tables in Excel, slides in PowerPoint, pictures in a lab notebook or drawings done in MS Paint. This makes it difficult to share, validate, and as time passes even locate (as anyone who has searched for a years-old word file knows).
Before the French Revolution, there were an estimated 250,000 different units of weight and measure in use. From my limited viewpoint, I think this may be an apt, although slightly exaggerated metaphor for the current state of crystallization experiment data representation. Adoption of one, or a small collection of generally agreed-upon and well documented file formats would allow for the easier development of universally accessible software suites that would in turn make managing, tracking and sharing of crystallization data consistent and trivial.

2 Towards Standardization

2.1 Looking Back

2.2 Approaches

I believe there are two primary needs that need to be balanced, data density and human-readability. High-throughput enters will prefer the former while low-throughput users will likely prefer the later. For this reason it seems that having a "rossetta stone" intermediate format would be useful. While itself may not be used commonly in applications its consistent formating would allow easy conversion to more specialized filetypes.

2.3 Exchange Format

2.4 Maximizing-Readability

Low-throughput users will likely not require the data density of binary, optimized and compressed file formats and will value being able to easily view their results in an accessible manor. A potential solution for this is HTML. HTML xtal files could include JavaScript that allow a single file to provide an interface similar to a graphical application while at the same time retaining a consistent and machine parasable structure which would be amenable to creation from and reversion to the standard exchange format.

2.5 Maximizing Density

In highthroughput settings, the volume and frequency of crystlization experiments makes the size of data much more of a factor than in a small scale setting. It would therefore be benefical for medium to high throughput users to have access to a format that minimizes memory at the cost of human readability. This format would be more invovled to parse but maximize memory efficiency.

3 Making Standardization a Reality

Things that will be needed

- Software (ideally online) for CRUD operations on existing file formats and for converting between files. This should be focused on low-throughput users as as throughput increases the compitation ability we can expect from the users increases as well.

References

[1] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, 112(3):535–542, May 1977.

[2] S. R. Hall, F. H. Allen, and I. D. Brown. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47(6):655–685, Nov 1991.