eIUS: Crystallography Experience Report 2

In the text: Some barriers in '{{…}}' to provide more contextual information.

Interviewee profile

Researcher at the EPSRC UK National Crystallography Service (NCS): Post doctoral research assistant for the NCS in X-ray Crystallography;

Time spent in research

60% on research (mainly eScience projects with interviewee 1, see Crystallography Experience Report 1) and 40% regarding the service provision

Research area

EPSRC UK National Crystallography Service (NCS, see http://www.ncs.chem.soton.ac.uk/index.html): laboratory-based facilities in the Chemical Crystallography Laboratory at the School of Chemistry, offering services in Full Structure Determination and Data Collection only.

The NCS additionally has a remit to further develop the science of chemical crystallography through research and innovation.

Research question(s)

As a national service "we collaborate with other people throughout other Universities and Institutions in the country, and all they're looking for is structure, a solution. So they'll beaver away in their labs making all these different new materials and they send them off to us to get definitive structural information." "So we'll either run it in the machine [the defractometer] here like we showed you earlier or if the crystals are not good enough we use the National Synchrotron Facility".

Areas of research for the service: "So a structured solution could be some fancy new drug molecule. We do quite a lot of medicinal chemistry work, anti malarial, anti cancer all these kinds of things, right the way through to new Inorganic materials, hydrogen storage and pretty much everything in between. So the work is quite varied and you never quite know if crystals you know sample to sample what you might get. We do a huge amount of organics because that's you know mainly what Chemists are making but we also do the Inorganic material side of things and given the current climate for new materials for all the you know greener energy and all the rest we are seeing a lot of materials pointed in that direction. So Organometallic Cage structures, Zelite type things as well as other Inorganic new materials, so it's quite varied it's quite good fun. You get to see a lot of interesting chemistry."

Research Lifecycle

Literature Review – Start of the research process

Start of the service process (and relation to research background): “We are assigned users to solve structures for, but they you know that’s kind of draws on our previous experience of what areas we’ve previously worked in. So I do a lot of the Inorganic stuff because that’s what my background is in. Whereas (.) one of the other guys on the service will look at more the Organometallics on a lot of the more difficult structures because he has all the experience in handling those. So we kind of split it up. So it's almost projects because we stick to sort of our areas of expertise almost”. **General example**: “Looking at the structures of a drug molecule it's completely different to a big Inorganic framework and how you approach the data collection and the structure solution is different.” Therefore **usually there is no literature review for the service process** and “we basically get a sheet with a formula what they think the structure might be and we take it from there so we’ll check crystal quality, well it's locked into our systems first so we’ve got a huge back end database and everything is logged into that and that’s basically our sample tracking system.”

Data collection process

Service (tracking of samples using web interface, certificates for users, SQL database): “We track our samples through our system through a big web accessible database, it’s all certificate security protected. So we can access all the data as service personal. Users get their own certificate which means they can access all their own data, and we can track all the samples through our system so we know exactly when it’s been, when it arrived, when it left and what happened to it in between, who looked at it, when it was collected what machine all that is collected on a big SQL database we’ve got downstairs.” **Example of how the system works and interacts with the users**: “When it's logged in it enters it into the database, generates an email, fires it off to the user who’s sent the sample to tell him that we’ve got it and it’s logged in, and throughout the whole progress when we first put it on the machine to look at it that generates another event in the database and generates another email which can go out to the user and say oh well you know we’re looking at your sample now, and then form that it’ll tell, well you know it’ll go through and generate further information say when we’ve solved it, when we’re re processing it, when we’re looking at it, when we’ve finished and when it's available to download off the servers”; “(..) recently we introduced what we’re calling our interact service and it's basically another, it's a big upload database. We put on all the data up to it, issue security certificates out to our users so that only they can see their own data and nobody else’s, and then they can log in and download their data directly so we don’t even have to send the data out to them now”. {{A part of the applications is still send on paper by post as “we need to accommodate everybody because we’re a national service.”}} (also see workflow of service under data collection in the crystallography exp. report 1)

Submission form on service tasks: “On the little submission form we have pretty much their reaction on what they’ve done and what they’re expecting the structure to be and if they want any other information or you know any special experimental conditions run.”

Defractometer (proprietary closed system; collects raw data from the experiments: “you need their software to interpret their raw data”): “Yeah a very

closed industry closeted system; the data collection that's fairly straight forward that doesn't vary, manufacturer to manufacturer it's all the same principles. It's how the software handles what you get off it's completely proprietary, entirely specific to whichever manufacturer you run in there, you know the machine is built like that. We've pushed for years to get industry open standards on raw data and they're just not interested."

The **binary output as an ASCII file (general example)** "gives basically an x y z co-ordinate almost, and an intensity number and a standard deviation calculated standard deviation, and it does this for the whole sphere of data. So every reflection you see has a location in space and how bright it is and nearer, and that basically is put into one giant text file which the solution and any programmes can interpret and all the solution programmes can interpret this because it's a fixed form text file".

Data analysis

Analysis of the ASCII file (general example continued, includes timescales of analysis): "So once we get to that then it's just a case of running that file through a set up programme to set up your initial structure files and then you can run your structure files to solve the structure and then you loop through that to refine your structure to get the model you want, which basically runs composing a model to the data to see how good a fit it is. So you run a few cycles through the computer, it will compare what it gets to what is already in through the data see how well it fits give you a, basically an error percentage and you adjust your model run it again and it compares your new model. So you basically you're refining it altogether as close a fit to your experimental data as possible. Sometimes it takes ten minutes; sometimes it takes two days or longer. If you've got a particularly problematic data set it can take weeks. But the nice routine ones can drop out, you can run the experiment in an hour, drop it out to final solution in half an hour and be done an hour and a half on a nice day. A lot of the stuff we get at the service is not that nice, and (..) the data collection takes 15, 16 hours and then the solution will take (..) two weeks trying to figure out what's going. So it's those kinds of timescales we operate in."

Analysis/service as a research process (experience matters): "Crystallography more than anything is all about experience" and the difference in "experience is quite noticeable when you get a really difficult problem. It's how you've handled it before or you know how to handle it from previous experience." I.e. evaluation activities are needed to assess the process throughout and "it requires well three years of training at least plus any post-doc experience you've got (..). So (.) it can be quite challenging at times".

Data types in the process: Raw propriety data ->> "processed data which corrects the raw data for all of the machine variables, detector variants [of e.g.] how far it was away from the crystal [or the temperature]" ->> the ASCII file (processed raw data) ->> "from that you run your solution" on a PC ->> visualisation, output etc. files.

Collaboration

The NCS consists of the Director of NCS, the service manager and three post doctoral research assistants.**Discussions**: f2f between the team ("(..) unique position having

basically four or five of us in the office that we can bounce ideas off and show structures to and generally talk about (..) the work we're doing (..)");

Discuss/collaborate via email: "We discuss with the user or whose sample it is, because it's meant to be a big collaborative thing. So they'll send us a sample and we'll collect the data and we'll talk to them about the initial result. If it's what they want – if it's not what they want we send them back and say 'well you know (.), you've got starting material that's made something completely different, how would you like to proceed'. But more often than not it's what they want, and so you know they have a few discussions (..) [on] what they want from the data and how we should work it out, (..) what they want to get from it, at that point you get a bit more specific."

Dissemination

Data collection only: "For just data collection we collect the data check [if] it's solvable and send it back out to the user, and for that if they do write it up into a paper we might get a mention like an acknowledgment". ("Sometimes they don't bother with that either".)

Full structure analysis: "If you go to completely to the other end where it's a full structure analysis it's almost like a collaborative effort. We get all the chemical information and the chemistry involved that we need from the user and we use that to interpret the results we get to give them their final crystal structure. So we sit there with our crystal structure that we've spent maybe two days (.) collecting the data and solving for, and (.) usually we provide a full structure report which the user can use but writing the paper which (..) for Chemists (.) [is] still the only way to publish." The team at NCS helps with writing "experimental sections, explaining the structure for, depending on which journal and which paper (..); they all have their own specific requirements." (more details see crystallography exp. report 1)

e-Crystals repository (open and free; fostering publication of data including negative results/experiments): Associated data files (the structure files) which might not be used that much in publication can now be stored in the "e-Crystals like structure repository as another route to get these out (..) because before they'd just sit there and (..) three other people that are going to do exactly the same experiment get exactly the same crystal structure (..); and that's the thing with science, negative results aren't reported you'd never read a journal (..) [saying] 'we tried this and it didn't work' which is sometimes as just as useful. It saves the next person (.) spending research money and doing exactly the same experiment and getting the exact same result and going: 'well that was a waste of time'."

The NCS plans to support and advocate to the community "an institutional repository that we can upload our full structured details to all the refinement files, make the whole process completely transparent, because it's not being peer reviewed". Not everybody is up to this in the community: "It's meeting with some resistance [some] people like to keep hold of their data, they don't want other people to see it". (more details see crystallography exp. report 1 on e-Crystals and open science)

Used as well in the community: **The Cambridge Crystallographic Data Centre** (**CCDC**, see http://www.ccdc.cam.ac.uk/); website quote: CCDC "builds, maintains and distributes the Cambridge Structural Database (CSD), a searchable database of organic and and metallo-organic crystal structures. The CCDC also produce and

distribute software products which make use of the data contained in the CSD." And "that used to be a route to get just crystal structures out you could submit CIF files which is all the structural information to them and it would go into their database". "It's a subscription, so you pay to use it".

Important journals: (see crystallography exp. report 1)

Other important elements about/in the research:

Future plans on electronic open lab notebooks: "(..) the electronic lab notebook is something we've been trying to sort out for a long time, and (..) the technology barriers have been there (..), Chemists are reluctant to use it in the lab when they're running their experiment they prefer the whole pen and paper. But we're pushing for it (..), we've got some ideas on how we can move forward with pretty much moving all our sample tracking experimental workflows and everything electronic, because we think that's definitely the way things should be. You need complete transparency of what you've done and how you did it". "If we can produce a standard for all this and farm it out to the community it will be taken up quite quickly, as it has been done before but it needs to be clear and concise that people can understand, so that's pretty much what we're going to be looking at next."

Modern computers (quicker processing): "There's been huge improvements on that so actually solving and refining the data is getting far more trivial on the whole computing just because of the power of the more modern computers" – "thirty, forty minutes an hour of cycle now take less than a second".

Video conferencing, collaborative workspaces on servers "all over the world (..) [with] collaborative document editing, all this is brilliant if the people are willing to use it, and we've tried it with a few of our users in this country and they're not really all that interested."