eIUS: Crystallography Experience Report 1

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

Interviewee profile

Researcher EPSRC UK National Crystallography Service (NCS): Responsible for the service; conducting research in his own area and to improve the service; doing promotion outreach and training

Time spent in research

Primarily employed as a researcher at Soton: about 50% time for service provision for a research community and 50% "divided up into research into developing new technologies to enable providing that service; I classify that as research"; this includes 25% e-Science enabled service provision and 25% covering my own personal research interests (maybe again with 5% admin and teaching)

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)

Personal research interests (also combining service with research and computer science with chemistry): a) "development of e-Infrastructure methodologies in application to providing a scientific based service as part of research, but that's me with the more computer scientist hat on than a chemist's hat on, so that is developing new systems to communicate the results of scientific experiments, to facilitate doing scientific experiments and currently I'm just starting a couple of projects which are about taking some of the outputs from the scientific research experiments and feeding them back into teaching the next generation, and a lot of that is done in virtual worlds, so virtualising the learning experience, so broadly speaking, that's the kind of areas I'm interested in, in terms of developing a computer science type development of the enabling of science (..)."

b) 25% personal research time over the last years, and this will take "the next five or ten years probably": "With my chemist's hat on, I'm interested in the determination of chemical structure and how that structure alters as you alter the environment of a particular material so I can change temperature or squeeze it by applying pressure, and that actually has effects on the structure which has consequent effects on the property so I'm interested in being able to control the properties of materials by

influencing their environment" and "trying to push the barriers so that one can get a much higher degree of information to make this link between the structure and the property of a material". The current models do not provide this high level of information so new ways of looking in more depth at these structures have to be developed.

Synthesis between a) and b), personal research and service: "broadly speaking, the outputs of what I'm doing in my personal research is pushing the boundary and that has an effect on what we say we can deliver as a service"

Research Lifecycle

Literature Review – Start of the research process

Personal research: "If it's my own personal research yes, you're kind of aware of where the gaps are. I do generally keep up with what's coming out in the literature; because our subject is quite multidisciplinary if you like, there are different ways of going about it".

Service/researchers submitting samples: no literature review is necessary for the NCS, but the researchers sending the samples will have done some: "You're relying on the fact that they've done all the literature work, the people contributing this to you are asking you to do the characterisation stuff; they are experts in that particular area of chemistry, they're fully aware of what's going on in the literature and therefore they're providing you with something that's new and novel and it's worth doing".

Data collection process

Workflow of the service: As shown to the interviewer in a tour of the facilities, the samples come in by post and are labelled and stacked up in boxes for processing. "When a sample comes through our letterbox or someone brings it down the corridor into my office there are a number of different paths it can take depending on who's giving it to us and what the sample is and what kind of result we're looking for out of it, so even though it's providing a service it's not necessarily totally formulaic." **Electronically** there is a "e-Infrastructure which is a loosely coupled set of web services which track the samples coming into the lab, where they physically are in the process and enable remote users to monitor that and get the results when they come off". (also see tracking of samples using web interface, certificates for users, SQL database under data collection in the crystallography exp. report 2)

Service (example of process): "In terms of the process of the experiments that we do, it's only for others in the service; we receive samples, very physical entities, the crystals are very fine dimensions, normally around about nought point one of a millimetre square; they come in from all areas of the country and all areas of chemistry as well, and they're being submitted to us either by people who don't have any facilities for determining or examining the structure of materials, or my colleagues in crystallography who don't have state-of-the-art equipment or access to advanced facilities, so we get a mixture of things either from people who know what they're doing, supposedly, or from people who just want the outputs, the result, as a confirmation of what they've done, so we have to manage an awful lot of samples coming through the door and track them through our system and also be able to inform our users of how that's working, so we developed e-Infrastructure for opening

up our operations, making what we're doing more transparent to our users and enabling them to see what's happening to their samples or to access data at various parts of the process."

Service (example of process continued): Using the defractometer at Soton: "The process itself is one of zapping crystals with x-rays and looking at the way the x-rays bounce off and recording that information; we record each data collection; data collection can vary from about fifteen minutes through to two days almost, depending on the quality of the sample and the nature of the experiment". "A dataset comprises of something like two hundred or three hundred binary images of x-ray beams coming off a crystal", the crystal is move into new positions and images are taken each time and the stack of images is called the "process dataset". This dataset is processed by the instruments and the connected proprietary system which software gives out an ASCII text as a condensed version of the data. This ASCII text file is the actual data set to be worked on/with.

Raw data storage: The raw binary data is stored in the large object data store (the Atlas data store at the Rutherford Appleton Lab) to be able to go back to this data after e.g. five years and reprocess it or hand it over to the researchers themselves for storage etc. ("being a service provider I have to be able to guarantee that I can go back to the data that I acquired").

General storage/backups: there is currently now centralised backup, but a number of portable hard drives and the laptops and machines of the NCS people.

Plans to link to e-Crystals repository (also see under dissemination): "We do think now about having one centralised backup but it's more effort and more work, but that's part of this repository system so everything goes into this e-Crystal system, it's one place, and we're currently trying to link datasets that are in the laptop backup management system if you like or in the crystals repository with the binary raw ones".

Data analysis

Service (example of process continued): Using the ASCII text file: "From that we can then start building our models of chemical structures and comparing our models with what we see in the experiment and finally we'll come up with a representation or a model of what was happening in the experiment of the material under investigation. That result is extremely useful in a number of respects actually, so there are databases with half a million of these things in that you can data mine, so just getting that kind of information into these databases is useful", e.g. for dissemination (see continued example under 'Dissemination' picking up from this quote).

One part of the service is to just send out the condensed data (ASCII) file to researchers who do not have the equipment, expertise or man power to analyse the crystal sample in the first place. But they then can use the data for their means and further analyse and process it ("work out the result, they propose the models and work up the result from there").

Or: "Chemists without any support from a structural perspective locally will use us as the whole service so we will provide them with the results and work in collaboration so there's a number of different ways in which an experiment is done".

From raw data to secondary data (including defractometer software): Proprietary software controls the instruments, data inside the system is without meaning for the researchers and this raw data cannot be accessed; they can only make corrections via the instruments, i.e. changing how data is recorded and corrected on a binary settings level ("we make a number of corrections to the binary data inside the software sweep that is supplied with the defractometer"). When the raw data is produced as an ASCII file (an understandable, exchangeable format) it can be processed further. The proprietary system is basically running on servers and Linux boxes (written in Phyton, those contain "all the correction data files which basically each data slice that is collected needs to be corrected in about six different ways") and the produced ASCII file can be taken from there to a common computational environment, and one can even work on it "on your phone on the train these days, so we shift from being based in the lab, tied to an instrument and a proprietary software system, to being much more out and about in a format that we can exchange and is very well understood".

"From that point we then work up the result and the result is even more well understood and even more machine process-able or understandable, and so the workup of that ASCII text file into our final structural result can be done on a number of different software platforms. There is one which is prevalent or predominant, about ninety percent of people use, so I would say we more or less have a de facto standard across the community", called **SHELX (standardised and portable software solution**: http://www.ncs.chem.soton.ac.uk/data_coll.htm & http://shelx.uni-ac.gwdg.de/SHELX/) – in the rare cases of other software used these can import the SHELX format. (see also below under 'Software used to process the ASCII files')

In the ASCII file "you barely get any information about (..) how the instrument was operating; what collections it performed; what corrections it then made".

{{**distinct difference between the two environments lab and office** ("general usability and accessibility of that data, the fact that it's tied to a particular piece of software and a particular instrument"): "There's the stuff that goes on in the lab and then there's the workup and analysis of the data outside of the lab and a lot of the infrastructural systems that we've been developing of still making that distinction and a lot of information is not propagating or coming through from the lab into the office or on to the laptop and that's actually a real issue and I'm spending quite a lot of time at the moment thinking about how you knit the two or bridge the gap or actually throw away what we've got and say well, we have a specification here but we really need to build something that's much more joined up or weaved together so that we're getting much more information about what's happening in the lab at least onto the laptop".}}

Software used to process the ASCII files further: In the "office I basically use software that's freely available so we don't have to buy software to workup our ASCII files and develop and generate our results; there are two or three bits of software available, they're all written by academics in their spare time and generally speaking they're written by the golden age of people in the field (..) and most of this software comes as a compiled piece of Fortran" – which means that sustainability and support can be an issue in the future, when the programmers of those apps are not available anymore (part of this has been written as early as in the 1970's; "the current stable version is from 1997"). **Names of the software**: "SHELX is the big one; it's a suite of small Fortran programs which was complied into executables and you'll never

see the source code". "Basically it's just command line DOS prompt sort of start, so there are some relatively nice interfaces and again, there are a couple of academics who have made a name by wrapping this code and providing nice interfaces"

"Another one called Crystals (.) came out of Oxford and that has quite a nice Windows GUI".

For the **analysis on the researchers laptop (general example, also metadata)** "you propose a model, you calculate what that should look like and compare it to actually what the data does look like; that process goes on on the laptop as well so you have intermediate steps, outputs which are data files, so you have your initial proposed one and then you have several iterations or refinements of the model that you are comparing with the real data."

Crystographic Information Framework file (CIF): "The final results file, the CIF, the Crystographic Information Framework file is a well understood format, it has a massive amount of metadata which it has plucked out from various different places, or users edit in by hand, unfortunately, but there's a mass amount of metadata about the result and a certain amount about the experiment, although that's somewhat lacking in comparison, and then it has the result underneath".

Sidenote: "the paper describing that format was written in very late 1991; by 1993 the whole community had adopted this as a standard and we still use it today as one hundred percent the way we exchange our results information" – a rare occasion of the quick adoption of a standard in a whole community, something lacking in the view of the interviewee for concepts of open science, like including documenting the whole information of the experiments (also failed ones; steps in the chain) in the research process. (also see 'Open Science perspective')

Collaboration

The NCS consists of the Director of NCS, the service manager and three post doctoral research assistants.

Collaboration with researchers submitting a sample takes place, when NCS is involved beyond the mere data provision service: "if I do the whole job myself then I'm very much involved in a collaborative publication where I present the results and write discussion and the very conventional way of doing things". (quote also under 'Dissemination')

Discussing and sharing results/findings (also contains dissemination): According to the interviewee there are different ways this can happen, e.g. "writing a collaborative publication, so someone's made this thing, they've sent it to me for analysis, we've done some analyses; right, let's write the paper and because I'm a National Centre, most of the time this doesn't come from Southampton, this is coming from other Universities, so we have to collaborate in writing a paper and the normal way; (..) a manuscript gets passed around so generally, the group who originated the thing, synthesised it, are the driving force, they'll be the principal authors on the publication; they'll prepare the framework for the journal article, out their contribution in there and then send it round to others." Usually a **word document** is used and exchanged **via email** for this. This kind of collaboration with researchers can lead to "a whole spectrum of things just within that which can happen; either they can say right, here's your section, you fill that in and send it back to me, which I do, and

they just take it as read you know, that guy knows what he's doing, that's all that needs to be said, that's that bit done, let's send it to the publisher, or there's a whole load of bits and pieces all over the place where they say well can you comment here, can you comment here and you end up with a track changes document where the changes that you're tracking are enormous and that can bounce to and fro quite a bit and you get everything in between, and so there are some people that just don't want anything to do with it and send it to us, we fill in a certain section and then send it back or we're quite heavily involved and that involves god knows how many versions of the document flying around all over the place and the usual problems of collaborative authoring that are bound".

Also see publication example under 'Dissemination' below.

F2f and phone: Sometimes "you might pick up the phone; there's not much more. For some of the more closer collaborations I have (..) if they're close by I'll get on the train and taking a laptop these days has made things a lot easier in that respect but there are no collaborative authoring tools based on the web infrastructure around. None."

Dissemination

Service: "The conventional way we go is through journal publications; you can publish an experiment such as the one I've described on its own, or you publish it as part of a bigger chemistry picture." **Example (combining service and research in collaborative dissemination):** "So my colleagues are synthesising new chemical compounds, I'm doing the structural work and saying this is what you've made, and together we combine to generate a conventional journal article, so there are a number of different ways in which our data or results can be used, and a number of different ways in which one might be able to get them out there in order for them to be used."

I.e.: "if I do the whole job myself then I'm very much involved in a collaborative publication where I present the results and write discussion and the very conventional way of doing things" (quote also under 'Collaboration'). As a mere data provision service (just sending out the ASCII dataset) "we are kind of relinquishing our rights to be involved as authors in the publication".

Also see discussing and sharing results/findings under 'Collaboration' above.

Journals: "Journals nowadays require that results file as well, that supplemental information, and that happened from the database people; the database people spoke to the journals and said you're publishing these results, we need them, so let's work together and make sure that it's mandatory that whenever you're discussing a crystal structure that you have to provide these results from; so that was the way the database got their information conventionally, so not only do I have to prepare some sections in a journal article in a conventional word doc, I also have to prepare the results file for publication as a sort of supplementary material or some such thing or other, but we do also have an arrangement with the database on a number of journals whereby you actually deposit that results file with a database, they provide you with an identifier which you then quote in the paper, but some journals want to have everything and so they ask for the results file, they do all their review process and when the whole is accepted they send that data to the database, so there are a number of different routes".

Publication example (also includes collaboration): “Let's go to one publication scenario, or one results discussion scenario, the next one is (..) when you start to be more involved in structural work, so this is the bit between my research work where I'm looking at how structures change and how the conventional service stuff that we do, which is just saying this is what you made; in between there is some collaborative ground where you can do collaborative work, so we have people synthesising for stuff on demand almost, so we say, well, wouldn’t it be interesting if we had a whole family of materials; they kind of look the same but subtle differences and we want to compare how these look structurally and how their properties are; it’s a bit like these pictures of the crime scenes where you can change someone’s eyes or their nose or something like that so it’s a bit like having the same face but just changing the nose and seeing how it changes the look of that person; it’s the only analogy I can think of; and you might have a series of that ten then and you want to compare and contrast then and that’s when you have to be discussing a lot more with your collaborators, so I do quite a lot of work that is actually very collaborative and there are a couple of free tools that allow you to visualise these results files, so the database people, in order to get people generating stuff in the right form and using it in the right form, provide free software, just graphic visualisation stuff, that allows you to render and look at these results files and so everyone within the crystallography community knows about these and uses them, but I have to train my users, I say well get this piece of software from this site, install it and then load up this results file that I'm talking about, look at this feature and then you can query it this way, that way, however and that’s what we want to discuss in our paper; it can be quite a painful process; you can have twenty emails go to and fro before you’re sure you’re both looking at the same thing and the same piece of software of the same orientation, god knows how many different parameters can vary, and then you start plugging stuff into word documents but the bottom line is we look at these things, visualise them and then chuck them in to transcribe some values into a word document and write a bit about it.” **And (visualisation software)**: “There are lots of pieces of software that allow you to visualise chemical molecules. There are a number of different ways in which you can generate them.” E.g. drawing packages (“tens, maybe hundreds of structure visualisation packages”), for Crystallography (visualisors of crystallographic CIF/results files) in particular enCIFer (http://www.ccdc.cam.ac.uk/free_services/encifer/) and Mercury (http://www.ccdc.cam.ac.uk/free_services/mercury/), both open and free as well as robust and easy to use, even for non-Crystallographers: “that’s pretty much what the community by and large uses, it’s great for a quick look at a result and query a few things and you can even look at the underlying data file rather than just the rendering of these things in a pretty graphics format”.

Open Science perspective: “Ultimately my goal is all about streamlining the process of getting data into these databases, whether it’s through streamlining the journal publication process or whether it’s just fast tracking straight to the databases; the fact is, the experiment’s been done, that piece of data should be in that database and at least eighty percent of it isn't”, for two main reasons: 1) “I generate more of these things than I can possibly write up in journal articles” and 2) collaborators sometimes do not want the results to go out, e.g. because they did not achieve the desired result (and might still be in the process of getting it right). Additionally it can take a huge number (e.g. 50) of single steps to get to a result, but in the end only the result is

published, i.e. the 50th dataset – "the other forty nine along the way get lost". "I'm particularly interested in trying to get all these lost orphaned datasets into the public domain so we can actually start doing some exciting new science with confidence. This science has been going for about ten years but it's very ad hoc and there's massive holes in these data mining datasets because the data's not getting into the database." (also see 'e-Crystals repository' below)

{{"(..) the big problem we have at the moment in publishing our data is that most of it's not getting out there, mainly due to the fact that you have to write a journal article and we're generating more data than we can possibly write journal articles (..)"}}

{{"Partly what I'm doing is generating results that are telling my colleagues and collaborators what they've made and we'll write [a] journal article and say look at this nice thing that we've made; the structure result will go into a database. It sits in a database with half a million other structure results and of course you can start doing data mining across those looking for patterns and similarities and that kind of thing in it, so you can start doing science on top of the science and it's like a by-product, and because we're not getting out all the results from our instruments and our labs into these databases, that data mining science hasn't developed like it should've done or like it could do, so I also see my role as one of trying to enable getting out more of these results so we can have better populated databases which we can get more reliable results from data mining exercises out of."}}

e-Crystals repository (project, see http://ecrystals.chem.soton.ac.uk/): a separate community is involved; the aim is to get "stuff from the lab or my laptop out into the big wide world": It is a "community run subject repository or database that's been going for forty years; it has half a million results in it but it should have maybe five million but because of all these lost things along the way, so the notion that a lab of a university or an organisation can have a repository which is basically storing everything that they do in their lab but at the flick of a switch you can make all that data available and it will immediately be harvested and deposited into the subject database so rather than having to write up the forty nine missing ones, the forty nine orphaned structures out the fifty structure story and always taking more steps back than you take forward, you can easily, within seconds almost, effectively publish that dataset and that's the whole idea behind it."

Important journals:

IUCr journals and tools: "There is a whole series of journals for crystallographers", run by the International Union for Crystallography (IUCr, http://www.iucr.org/). "You can actually write a paper in this CIF format and they're fairly well advanced because they're crystallographers, so you basically write the paper in the results file and their software renders it" as a PDF paper/file. And: They also have a "fairly neat software which is two windows side by side, this is the raw file if you like, and then they have a word style rendering and you can edit the word type document and it actually updates the underlying results file, or if you know the speak or the mark up language of the underlying results file you can edit that and it changes the word rendering". A "tool written by established organisations within the community, for the community".
This free **software is called publCIF** (http://journals.iucr.org/services/cif/publcif/).

Other important elements about/in the research:

{{"(..) I found from playing around with the infrastructure (.) that trying to involve our collaborators with the experiment more helped in their understanding and it enabled me to realise that actually, we need to collect a lot more information about the experiments that we do and I think we're very, very guilty of this across chemistry".}}