eIUS: CombeChem Chemistry Experience Report

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

Interviewee profile

Senior scientist/researcher and leader of a research group at a chemistry department at a UK university; one of the leading researchers in the CombeChem project (part of the UK e-science programme)

Time spent in research

"I would say thirty to forty percent of my time."

Research area/research question(s)

"I've got several, so in Chemistry I'm looking at the properties of molecules at liquid surfaces, I have collaborations with physics and we're developing a laser based laboratory x-ray source and then there's the e-science research as well where we're looking at enabling collaboration in smart environments. That's a brief summary; there are plenty of other things but that will do."

"I would just say I do a mixture of laboratory based, computer based work and that I am very interdisciplinary and I am involved in a large number of projects and a large number of topics which probably distinguishes me from a number of other researchers who are perhaps more focussed, but otherwise, day to day it's very variable depending on which topic, whether I'm trying to do something, a piece of research myself or managing research; it's very, very variable but I wouldn't say I was particularly different other than doing a large number of topics."

Research Lifecycle

Literature Review – Start of the research process (also includes research approaches, examples)

"Most of the research topics that I'm involved in are long running and the literature stuff is essential, but we've often up to date on the literature, the research is done collaboratively with students and they may be looking at some of the literature and background (..); we look at the literature, look for data, talk to colleagues, come to meetings."

No specific sources/tools are used: "If I was looking in a new area I would do as everybody else is, I would look in Google and then I would look in Web of Science. No particular journal, partly because the areas are so diverse and public and even when it is, say, just a chemistry topic, relevant material can be published over a very wide range of journals, so no, I don't think there's any given tool other than the search engines."

Specific research questions/approaches (examples): "In every area (..) there are important questions; if we take the x-ray one then the important question is producing technology that will enable us to look at structures of proteins and protein clusters in

much nearer to the conditions that they will operate in an active cell; that's a way off blue sky limit for what we're trying to do but provides some motivation (..)." "(..) that particular one is a technology driven one; we're at the beginning and we're trying to develop the technology to do something, so the literature work and the background studies and understanding tell you what you need to be able to do and a group of us had an idea about how we can build some new technology that would enable that, and the first step is to build that technology, so then, yes, again, looking in the literature and developing the equipment, making the measurements, looking at it in that way, in no different from any other project that would build a piece of equipment." **There are certain activities in this particular research which cannot be done easily in every laboratory – they are working on solving such problems via new technology:** "(..) we'd have to go to a big facility, make measurements on a nano scale, so there's a window of size and time for looking at physical and chemical phenomena which is missing at the moment and this is what technology would replace (..)."

Other research areas: "In the other areas of chemical laboratories, it's more on fundamental understanding that would lead to understanding of behaviour of liquid surfaces which might have an environmental impact that one of the students is working on. The general e-science area is just generally supporting laboratory work, so just facilitating the collaboration between people, electronic notebook and so on, so that has a more immediate feedback to enable and speed up the collaborations."

Data collection/analysis process (see also under 'Collaboration' for use of tools and electronic lab notebooks)

The experiments in the laboratory have to be documented properly – traditionally this is done on paper sheets; here "(..) making sure the data is sharable across a distributed operation (..) having all the metadata and so on associated with it, that's been our issue." This includes using and further developing technical solutions like in the context of open lab notebooks: "(..) it produces a more efficient, more modern, more rigorous, more adaptable, more sharable version of a laboratory notebook, that would be one area; and then facilitates conversations on top of that; discussions about the material, and then ultimately, keeping a repository of the material and the data to make it more readily accessible along side a publication for example." (quotes used in this paragraph can also be found with more supporting quotes under 'Facilitating research and collaboration..' in the 'Collaboration' section)

Importance of the recording of Metadata as an ongoing process: "(.) you need the context of all the data; (..) how it was recorded, when it was recorded, who it was recorded [by], what the units are, what's it's associated with, what is being done with it; if you've got a graph you want to know what the raw data behind it is and [to] be able to recover that, you need all the information about the conditions and so on it was taken from; all the necessary background to context to provide proper understanding of that data." "Exactly what items of information are needed to provide context will develop as experiments develop (..), experiments change but the need to provide the description of the context will always be there; (..) you need a description and need to be able to keep that linked with the data (..)". "(..) how you describe it will have to evolve because people's understanding of the material will change."

Proper descriptions of experiments stored in repositories/data archive are "(..) important and that's another reason why full context is needed, so that subsequent

researchers [can] recover that context and understand the data so that they can re-use it or check it or re-implement it, whatever they need to do, and therefore comment further on it and therefore contribute to the description of that data as they collect other material that's used together with the existing material, and so the collection grows, but it has to be continuously reviewed in that sense."

Data storage (time), curation, responsibility, tools: "Some data you might decide is to be kept forever, but [for] most experimental data you would not necessarily keep the raw data forever, but you would consider how long to keep it for, which may be a few years, it may be ten years, maybe twenty years, maybe a hundred years, but you make a statement and then it's re-visited at the appropriate time." "I consider it to be a responsibility of the person who creates the data to consider how it might be curated and provide the necessary information (..)." "(..) the tools should enable you to keep the data, keep it in context and make it available."

Collaboration

Facilitating research and collaboration (in the laboratory) through the use of tools developed in projects: "(..) if you develop a collaboration tool and you have a project that's a collaboration, it can obviously use the tool. [But initially] the tool was not developed for that [specific] project." "(..) you have a piece of equipment initially, they would come in a collaboration and use it with the existing people and develop an experiment around that and if it works out well it will become a more routine piece of equipment (..)" – which very likely will work better than something off the shelf. **Use of electronic lab notebooks as a one specific project/research theme**: "(..) making sure the data is sharable across a distributed operation (..) having all the metadata and so on associated with it, that's been our issue." "(..) if you are a collaborative project then you might well be able to use some of the infrastructure that we've thought about for electronic lab notebooks, for blogs, general semantics and so on in chemistry, and those will just be the sort of thing that (..) initial adopters would run up themselves, [those] subsequently may become part of other people's infrastructure (.), more generally, probably the ideas behind it will be used by other people to enhance their products, their software, and then it would become available hopefully as a tool like any other piece of software (..). The concept of having a tool that enables you to collaborate that enables you to store data, understand how to put the metadata in it properly, the initial users will be part of that development of tailor[ing] it to themselves [and] subsequently users would just use it." It supports collaboration in that "some of the tools we do, allow a discussion on that data, it's a blog style discussion on your data or your analysis for anything else that you might want to do." A **project with early adopters** is already established: "(..) it produces a more efficient, more modern, more rigorous, more adaptable, more sharable version of a laboratory notebook, that would be one area; and then facilitates conversations on top of that; discussions about the material, and then ultimately, keeping a repository of the material and the data to make it more readily accessible along side a publication for example." {{"(..) the problem with the traditional paper records is that they're often not very well kept, they're often difficult to read, they're not easily searchable, they're not sharable, if a student leaves it s very hard to correlate the information in the notebook with the files that are on the computer, if the linkages are not there it's very tricky, you can't have multiple users at the same time discussing material that's in a

lab book very easily without a lot of repetition, so it just generally eases collaboration, even between a supervisor and a student and extendable to bigger, interdisciplinary collaborations where you need expertise from multiple sources to be able to analyse the data."}} **Technical/organisational requirements**: "You could run it on a simple server, you could distribute it; we are agnostic on that point; the data that you might run underneath we would believe it should at least be replicated institutionally, each institution a repository, each research group might have it's own repository (..)".

"(..) the **main collaborative tool** is the blog work, which is novel in the sense that data is held there and commented upon where blog doesn't normally have data and images in quite the same way, so that's available for comment; the other side of it is that if we record the laboratory work, record the data in a way that's semantically rich then it's a lot easier for somebody who is collaborating to understand the context that that data was collected in; so whilst with the blog tool we actually do the collaboration, some of the other work was done are making sure that what is recorded will be useable in a collaborative context, even if it itself is not a collaboration tool it allows you to share, so you use another tool maybe to share it, but what you're sharing is more easily understood by the team, or however remote they were from the experiment then it would currently be, so instead of just a set of numbers you've got all the context for example would be there, so instead of just a graph, you've got the graph plus the data behind it that created it."

Dissemination (see also electronic lab notebook example under 'Collaboration': making data available "along side a publication")

"Some of the work's been published in the way the e-science stuff is published, other stuff is published in the subject area journals about the research that was actually done and indicating how these tools were used as just part of the description; so they're not being published in their own right, they're being published because they did some science and there's a note as to how that was done, if you like; so we're preparing some things that go out to some journals at the moment, but again, it's angled on the scientific work that was done, not on the infrastructure development. The infrastructure development is published in some of these e-science conference areas and I suppose some bits in the journals of chemical information for example, so there are some slightly more specific areas but the biggest game is to publish it along side the science that it enabled."

The **vision** is linking the publications "back to the raw data, should you need it, and some will be kept for longer, some less, just because after a while it becomes pointless to keep the raw data, it would be better to re-do the experiment because you could do it better, but up to a point you want that traceability all the way back to the raw data."

Publishing data from experiments/open science practice: "(..) some of the blog work is being recorded in an open science context, that is, the data is exposed as it's collected and there you're obviously free to look at that because everybody is (..), some of those [blogs] have been running for more than a year now so there's quite an archive of data available." Having open blogs also means that the "data is available, (..) it's very much easier to write a talk subsequently because you can find the information you need and you don't have to go and find the person and get the file; it's there, it's available, collaboration meetings are easier because it is easier to see

what's been doing, it's very much easier to find previous information to compare it with, it's all available much more readily than it would be normally if it was in individuals lab books or even collectively on a server; it's harder to find this way; the conversation has been documented so it's easier to go back and find it."