eIUS: Earth Sciences Experience Report

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

Interviewee profile

Senior scientist/researcher in the area of computational mineral physics in an earth sciences department at a UK university; involved in the NERC National Institute for Environmental eScience (NIEeS); one of the leading researchers in the consortia projects MaterialsGrid (DTI funded) and eMinerals (NERC-funded)

Time spent in research

"There are two main things I do, one is teaching, one is research, and I think probably more than half is research, but of that half, quite a large fraction of it is managing other researchers rather than doing stuff myself (..)."

Research area/research question(s)/example

"In the broad area, what we are interested in is understanding how materials work based on what happens at the level of the atom, so a large part of what we do is computer simulation; but there's a big range of simulations that we do, so we do a mixture of simulations using quantum mechanics where we try and understand the electronic structure of materials and from the electronic structure you can then work out the properties of materials or behaviour of materials; so on one level we do electronic structure, then we do simulations where we have simple mathematical functions to describe the forces between atoms, so we can do another class of problems, in particular, we can do very large simulations of that sort, and then the third type we do is what I would call inverse modelling where we don't actually represent the forces between atoms but we have experimental data and we basically tune the configuration of atoms until it gives the best agreement with experiment, so [these are] the three sorts of area that we work at."

"The sorts of questions we're asking are really quite various but it actually comes down to trying to understand something about the properties or behaviour of materials and always in what we do, we're looking at the atomic foundation, so in terms of examples, one example of what we do (..) where we use really very big simulations, is trying to understand how materials withstand the damage due to radioactive decay, with the application of trying to identify good features of materials for long term nuclear waste storage, so that's one area. Another area we're putting a lot of effort into – (.) it's actually very difficult – is to look at the way in which pollutant molecules bind to mineral surfaces, so trying to understand a little bit about pollution in an environment and how pollution is trapped; and then we look at also how materials transform from one structural phase to another as you change temperature or pressure, and I suppose that's a bit more on the less applied, more on the fundamental side, but I have a long standing interest in understanding phase transitions; and then I'm also interested – again, it's more of a physics side – (.) in understanding disordered materials, so we do a lot of work on glasses, understanding how glasses occur,

understanding how glasses respond to temperature and pressure; we're also interested in [the behaviour of] a lot of materials at a high temperature, there's a lot of structural disorder; sometimes it's dynamic, sometimes it's atoms swapping their positions around, so we study that sort of process by modelling; and then I guess the last area is (..) trying to understand (.) why some materials shrink when you heat them up; see, most materials expand when you heat them up and we have a lot of interesting materials that actually shrink when you heat them up, and so we're using computer simulations to give us insights into that."

Research Lifecycle

Literature Review – Start of the research process

"(.) literature review I think is a phase in the process that's a little bit ill defined because usually you pick up a topic because of the knowledge of the literature rather than say, 'well that's an interesting topic, I'll go and see what's being written'. You tend to now what's written and then you embark, so the project is formulated as a result of knowing what's in the literature rather than the literature review being a part of the project, and I think literature review is often ongoing, so if I were to take a material that I'm interested in, I will often keep track of what's going on in that area."

"Different people have different approaches to actually doing research, some people like to have a very targeted question. I tend to like to fish around first. Starting on a project, I would usually try and do a whole load of trial calculations (..) and then formulate a question based on the sorts of things that we get out."

Long and detailed example involving a material called "Silver Cobalt Cyanide" and expanding into the other areas of the research lifecycle: "(.) we have a group and one of my team had been doing some experiments and had discovered (..) that in one direction it expands when you heat and in another direction it shrinks when you heat; but it turns out that the amount it expands and shrinks is huge; in fact, I think it's possible close to being the record, certainly the record for shrinking, it's probably the record for expanding. So we started (..) looking at this material as a fishing around, then we discovered these numbers and then we started to ask the question why, and it seemed to me that one of the things we should do is some proper calculations on this material, and at the time, I hadn't entirely fathomed out what we were going to do, but it seemed to be a sensible approach just to do some calculations. So we were going to do some calculations, so we had the knowledge that this thing expanded and shrunk but we didn't know exactly why, so we started off; (.) it seemed to me that this material was not going to be amenable to the approach of using simple functions to describe the forces because we didn't have a starting model and it seemed to me a starting model was actually very difficult to obtain so we went straight into doing quantum mechanical calculations and the first you do is, 'let's just have a model of the material as it is'; and when we started doing it we found it very hard to get agreement with experiment, and we do know, we have a good feeling for how close agreement should be, none of these calculations give you prefect agreement but we have a good feeling for what sort of agreement we should get and we were getting a long way away from what should be the agreement with the experiment, but it was agreeing in certain respects; so it was getting all the bondings correct, but we knew from the structure of this material that it actually has quite a significant, like a hinging affect; you kind of take the material and just bend it over and it was getting that bending bit

somewhat wrong, so we then tried out lots of different levels of approximation, and fundamentally we were always getting this wrong and so we also know what sort of affect might be missing out of our calculations, and so we then did a calculation of this effect and indeed, we could see quite easily that if we added that affect we could actually get the experimental behaviour quite sensibly. We realised then that actually because of this not agreeing, but because we know that the calculations always have to work, we therefore felt we had identified something that was missing, but we also know that the missing bit only comes into play as a very subtle affect, and subtle affects only come into play usually when there's a very delicate balance of the forces, so we realised that this was a material that essentially (.) hinges. So you can imagine bending the thing, now we realised that the forces against bending were actually incredibly weak and we mapped out the whole energy surface by doing a lot of calculations across a grid; this is a grid of points rather than a computing grid; so we did a whole load of calculations across a grid and then we back-corrected for the missing affects, but fundamentally I think that we understood exactly why you were getting the coupling between very large positive expansion and very large negative expansion that came out of this, because our grid of calculations gave us the incredibly steep valley of points, but along the bottom of the value it was incredibly flat, so this gave us an understanding of why you'd got the positive and the negative coupled and why, if one was big, the other was big, but we also realised that if you then followed across the bottom of the value it was incredibly soft and we know that the theory of thermal expansion uses that parameter and so by realising that this surface was very soft, we could then understand why the effect was very big. So, effectively we iterated our way towards an understanding of why this material had both huge positive and huge negative thermal expansion and we could identify I think, the components of the interactions that give rise to the energy surface and so I think it gave us a bigger understanding. (..) But in all other respects I think we pretty well understood this material, and we've got two papers out of this; one in 'Science' and one in general physics (..), and these came out earlier on in this year. So the process wasn't well planned because I don't think a plan would work; this is my point about fishing around; so we knew what the question was, the question is why has it got very large positive and very large negative thermal expansion; why are the two coupled and why are the values very large? That was the question. We identified that we would probably get a lot of insight out of calculations, identify which sort of calculation to use and then we did a whole load of trial calculations which then told us stuff we would not have guessed in advance, so having understood the situation through the trial calculations it was only then that we could plan the detailed suite of calculations to undertake, but by that stage we knew pretty well what we after; but this is my point about the fishing around, in my view, for my own way of working, is actually quite important." **Involving the community**: "Now, there is still one factor that is missing in all of this (..); we managed to pin everything down to how the silver atoms interact, but I don't understand silver well enough to know why it's giving the effect that it gives, so that's the next level of question that's probably something I'm not going to try and answer and I'll tell you why in a minute." Answer: "I have learned in science is that it's not always good for the community to answer every single question yourself, so I'm going to throw it out and let somebody who knows more about silver than me take the next step on; what I have realised is that if you find something very interesting that gets you a high profile paper and then you go on to

answer all the following steps, then nobody else picks up o the interest because you've done it all, so I have learned that the best thing to do is to provoke the community and then to step back a bit and let other people get involved and then move on to something else and follow some interest."

"(..) understanding is the key thing we're after, but that's an iterative process that involved looping around back and forth, comparing with experiments; in this case [of the **"Silver Cobalt Cyanide"** example] we knew we weren't matching the experiment, but we were able to identify what was missing (..); the fact that we couldn't properly reproduce the experiment actually gave us an insight, so some people would say it doesn't agree with the experiment, it must be wrong, but we know how these tools work and if it's wrong, it's wrong for a very good scientific reason, not just "it's wrong", so we wanted to understand why we weren't reproducing experiments, because that gave us an understanding of how this material works."

Data collection/computer simulation process

Detailed description of simulation process, used Grids, examples: "(..) my chosen technique is computer simulation (..) so for us, the computers are the equipment. However, we don't own all the computers we use, so our simulations are of many different types and because of this they actually require very many different types of computing, so some of our simulations require big super computing; the sort that you would only get with HECTAR [i.e. high availability clusters] (.). (.) [Our university] has it's own super computer as well, so some calculations absolutely require super computers; there was a time when most of our calculations required super computers but that has changed quite a lot, so at the other extreme, my laptop (.) now is a super computer by some standards; I mean, laptops are now so incredibly powerful compared to how they were ten years ago, that increasingly stuff that was going to require a special computer I can now do on my laptop, so there's a lot of calculations I now can happily do on my laptop and it's interesting; I was benchmarking something for a paper I was writing, I got a figure out, it was something like a few years ago, [on] a computer we bought specially for doing work a calculation that took three hours now takes less than second on my laptop; (..) so increasingly we think of super computers as being for the very big simulations; ones that require a million atoms or so. There are a lot of calculations I can happily run on my laptop by myself, then there's another class of problems which is where I actually want to do something that requires a certain amount of parallelisation but not a lot, and so I might for example, take something like the National Grid Service where they've got clusters with quite fast interconnects and so I could run a job with eight processes where the job itself doesn't demand the big HECTAR facilities but it's slightly bigger than I would run on my laptop so I would use the NGS for something like that, or increasingly, what's going to happen and we're doing this here, you now buy four core processor boards so you can put MPI onto a four board processor board and I can run many parallel jobs on a single cluster node. I'll come back tot his in a minute; the other resource thing that I want to do, is I often want to do a lot of repeated calculations which all run happily on a single processor, but I might want to run several hundred of them, and that's where grid computing comes really very useful, so (..) [at our university we have a local] grid and then we have the NGS we use a lot (.); well, I don't use it now but we also used to use the North West Grid; in all these cases you can get small

single processor jobs or jobs with small parallelisation, so on [our university Grid] (.) I might run a lot of four processor tasks, on NGS I might also run a number of small processor tasks. Increasingly I do work where I actually like a graph that's got two hundred points on; (..) There is a change I think in the way we are going to be doing research; often I want a graph, say, something has a function of temperature, and in the past, they way you would do that would be you'd do a few runs, say, (..) ten runs on the graph, and then you would start to fill in the points one by one as you found the interesting areas to fill in; but now with the grid technologies that we have I can easily create two hundred jobs at once and just send them off to the NGS and it comes back and I have a complete graph of what I want, and rather than filling in the graph by iteration, I now construct a graph from scratch and the nice thing about getting two hundred points is that you can start to see subtle effects, whereas in the past when people drew graphs they would try and do as few calculations as possible to prove the point they're trying to make; now I can produce a graph that has got so many points on that I can see things that I wouldn't have seen had I chosen the points one by one (..)."

Experimental data, examples (also continuing "Silver Cobalt Cyanide" example): "Sometimes laboratory experiments simply gives us the prompt and then provokes us into doing something; more often, if there are lab experiments we will calibrate, make sure our calculations agree, and there are two things you can do; you can either change the model to make it agree but in the case I was telling you before [in the "Silver Cobalt Cyanide" example], the fact that we had calculations that didn't agree with the experiment actually told us something that was physics, so that was actually quite good. We do a whole series of inverse models, (..) so one approach, the standard approach is you do a simulation, compare it with the experiment and then change the model if it doesn't agree, so you start with a model and then you compare with the experiment, the inverse modelling actually you don't have a model, you start with experiment, produce the configuration that agrees with the experiment and then you develop the model, so there's two arrows in opposite directions; the inverse modelling has been really very very interesting and it's one that I am particularly excited about; (..) I think it's going to be really quite important long term because we're basically getting stuff out that is purely from experiment and not from any model, but we are doing simulations but we are basically using as our model the agreement with the experiment and that I think has been quite good."

Data analysis/visualisation (also see last part of 'Data collection/computer simulation process')

Graphs for visualisation and analysis, example: "(.) we use graphs to understand how the data are varying as I change the parameters in the calculation or in the model. This is a different example now; one of the materials we used is Silica Glass; so, most materials, when you squash them, they get harder, so you know that if you try to squash them you can't squash them any more, they eventually get so hard you can't push things down any more; I mean, if you take a cardboard box and try and squash it, first of all you squash the cardboard box, it breaks very easily and then you keep pushing it and pushing it and then the cardboard starts to push against itself and then you can't push it any more. (..) Well that's how most materials are; (..) but there are some materials and Silica Glass is one of them, that as you squash them they actually

get softer not harder, and so we were trying to understand that and it's a subtle effect and we plotted the volume as a function of pressure; it's an obvious thing to plot, and it has a curve and if you had a few point it would just look like a normal curve, but we realised we had to have a lot f points and we had a lot of points you could actually see the subtle structure on this curve, so for most people it will be a straight line going down, but actually when you look at the points you saw it had little curves in it and what we needed to do was to fit functions to this in order to differentiate the functions because the differentiation of the function tells you about its stiffness and you couldn't really do a simple difference type plot because the effect was so subtle, so we actually felt we had to go and get a lot of plots in order to fill a very good curve in order to be able to differentiate that curve in order to be able to get good results and it's an example of where, if you have a lot of points you can actually do rather more with the data than if you only have a few points, so we went through that process and we actually did discover why the material gets softer as you squash it and that also is written up and published about a year ago. Again, it was all done with grid computing because gird computing gave us that possibility of doing a very high through put of results out."

Use of laptop/desktop/Grid in own work practice: "I do all the preliminary calculations on my laptop or a desktop and then we use the grid for production work; we don't use the grid for testing, so there's probably the case; the real benefit of the grid is not for doing one off calculations but for doing two hundred at once."

Automation of research tasks/Grid jobs (also see 'Data sharing..' under 'Collaboration'): "(..)we're now in a position where we can run so many jobs that you can't possibly look after the data by hand. (..) if I have two hundred points, I can't manage [and analyse] two hundred sets of runs, (..) so I basically want to just have one command that just grabs everything for me and that's where we've managed to get to, and it's all integrated into the job submission process and it's also integrated in the job creation process as well, there's a job submission process, so I can type one command which will set up all my jobs, put all the input files into the data grid, submit the jobs and at the end of the jobs the data will all be archived on the data grid in a proper form, metadata will be collected and I can then go off and actually draw my graph with one command. So two commands can do the entire science."

Collaboration (also see 'Involving the community' under 'Literature Review – Start of the research process')

Interaction (tools): "The traditional approach is email, (..) but I think email doesn't help particularly; it clearly has some role but is not as good a role as people think. (..) in terms of interaction we actually use instant messaging a lot; it seems to me it's a lot better than email (..); we also use things like Access Grid or video conferencing and there are lots of video conferencing tools (..). That's really the interaction level of things."

Sharing data, includes information/examples about output files (xml), standards, tools: "The other level is things like how do you share data? How do you share ideas? How do you share information? The traditional way (..) is to email my output files to somebody, but there's a big problems with that which is that they won't understand the output files. Unless they're an expert, all I'm giving them is a pile of numbers and most output files are not terribly well documented. There are some programs that

require you have the manual to understand the output files because they really are a list of numbers. There are other programs that not only do you have to have the manual but you also have to know what the input file was and you really have to now the insides of input files, so we have done quite a lot of thinking about how you represent data to make the data sharable. So now what we do is, most of our programs now write xml output and then we've developed some tools that allow you to transform the xml into a web based report, so you can do this is xml and it allows you to transform from one xml language to another semi-easily; so we have a tool that transforms our simulation xml into html. So, all our output is in xml so all xml means that if you have a number you've also got a tag to tell you what the number is, and all our numbers are tagged also with a dictionary reference, so we then transform our xml file into a web report and every single variable has a number and a number and a unit and if you put your mouse over the variable you get a dictionary entry that tells you what that variable means; so if you take something like temperature; everyone thinks they know what temperature means, but actually it's defined scientifically in several different ways which all have a common root but in terms of what they mean, they're really very different; (..) for a different application you would have a dictionary that tells you what temperature means and then when you put your mouse over that word you get the dictionary telling you what it means, (..) in a lot of the cases our simulation does things in steps so you start at the beginning and step by step things change and then you can get a graph showing you how it's changed and then sometimes the end result is a calculation of something; you can again get graphs, so we get lots of graphs in this report, so we're transforming from lists of numbers to graphs and then we can also get up the atomic configurations and show that. So what we've gone from is the traditional view of output files as being piles of numbers to web reports that come because we're using xml. So we've developed those sorts of tools and we've developed tools that allow us to extract numbers from xml files quite easily, so when it comes to data sharing we don't just send each other output files anymore, we post these web reports, but actually we don't post the web reports, we post the xml files; the web reports are generated on the fly for people because it's pretty quick to do that. So then you have to share data; now, sending email is a pretty bad way of sharing data, so we might use something like the Storage Resource Broker;"

Use of Storage Resource Broker (SRB): "The SRB is basically a distributed file system so we put our files on to the SRB and then when you want to look at a file then it does the; my collaborator[s] (..) can see my files on the SRB and then they click on the button and they get this transformation into this web report, so I don't actually have to send anything (..), they can find it for themselves; so in terms of sharing data, our view is that data are not what your collaborators want; what they want is the information within the data and so by going down the xml route we're actually able to do this transition from data to information, so our colleagues can understand the information content of data without us having to explain things to them; so the traditional thing is to send a file and say well if you cut and paste that table into your spreadsheet and you draw it, you'll get a graph and it'll look like this – [in our work practice] actually this is all done through the xml instantaneously so there isn't any cutting and pasting (..); the worst thing is when people send you spreadsheets that are not annotated properly. Everything is annotated, everything explained by dictionaries and so life is easy."

A **self-developed and written web tool application including xml libraries** is used for that purpose (as described above), integrating the different formats and functionalities.

Data sharing, integration Grid-desktop/laptop, metadata, archiving, functionalities: "We're at a stage now that it's time for us to start publicising this; (.) currently we use it in a small group (..), but now is the time to actually start to tell people more about this. I mean, this is all to do with the sharing and the sharing of data has been part of this and even if it's just sharing between two people in the same lab; I mean, the classic case is where a PhD student finishes, leaves, supervisor tries to find the results throughout the paper; it's almost impossible, nobody looks after their data properly, and this all allows us to look after both the data and the information in the data by using something like the Storage Resource Broker you can actually probably archive, so if I run a job on the National Grid Service, this is all part of the issue. Most people, when they run a grid job get their data back to their desktop. (..) we have a tool that actually integrates directly with the data grid so that you submit the job, it pulls the data down from the data grid, it pulls the executable down from the data grid, it runs the job and when it's finished it puts the data back into the data grid so you end up with a complete archive automatically of all the input files and all the output files and the xml fines as well, and it's in a way that you'd have to work hard to damage, so if stuff goes back on your desktop it's very easy to lose track of stuff and to mix up the wrong input and output files. The way we do things, you almost have to deliberately destroy your data because it's on a form that you can't destroy by accident so we have a complete archive of everything and we collect metadata so I can go to my metadata browser, be it an online tool or a web page and then I can go straight to the data from the metadata, and that also allows my collaborators to find my data because all they have to do is look at the metadata and they'll tell the collaborator where the data are so I don't have to send an email to the collaborator to say the data are here because these things have happened automatically."

Dissemination

"(..) we are using traditional journals because at the end of the day most people who will read your paper will find it in journals and not from anything else."

Important journals: "I think most journals in physics do that now [linking to data], so if you look at the 'Institute of Physics' or the 'American Physical Society' or the 'American Institute of Physics', these journals all allow you to deposit data. The other journals we look at are mineralogical journals like 'Physics and Chemistry of Minerals' or 'American Mineralogist', and they all allow you to deposit data. In part they are keen on it because it cuts down the number of pages they have to print, so if you have a lot of tables they can all go into supplementary. 'Science' and 'Nature' also do the same thing, so our 'Science' paper, we put a lot of data in as supplementary data. Where I think they're not very good at is I don't think they're very keen on storing the raw data, I think they like to produce tables of data so I don't think we've ever deposited a whole load of raw files."

From the "Silver Cobalt Cyanide" example under 'Literature Review – Start of the research process' (full quote there): "But in all other respects I think we pretty well

understood this material, and we've got two papers out of this; one in 'Science' and one in general physics (..), and these came out earlier on in this year."

From 'data integration..' under 'Collaboration' (full quote there): "We're at a stage now that it's time for us to start publicising this; (.) currently we use it in a small group (..), but now is the time to actually start to tell people more about this. I mean, this is all to do with the sharing" of data.

Other important elements about/in the research:

Biggest goal/impact of projects like eMinerals: "The big grid area. (..) [integrating] job submission with data management, information management and collaboration (..); we've now got a lot of our tools onto the National Grid Service and what I hope to do in the next academic year is to start the ball rolling with getting other people to use the stuff (..). (.) we call it eMinerals toolkit if you like."