StORe

Interview Transcript

Interview reference: Biochemistry 6
Role: Member of academic staff
Interview length: 45 mins
Format: Face to face
Questionnaire respondent? No

A. Identities

A1. Which discipline group best describes your field of interest?

Biochemistry.

A2. Briefly describe a typical research project workflow. What steps do you go through e.g. from generating data, then analysing it and eventually publishing?

Firstly published literature searching. We use PubMed, ScienceDirect and PubMedCentral. For data mining I use NCBI and also PubMedCentral and ScienceDirect, which gives us all the electronic access and then I have to chase up the ones we don't have. So that's for the basic background.

We sometimes look at the structural databases: we look at Gnome a lot; Ensembl; we use EBI as a starting point. Within Ensembl there are so many links we tend to just jump out.

For structural information we start off with the ExPASy protein structure database and from there the ExPASy Proteomics Server that then has a full menu of programmes.

We use other peoples' structures we don't generate our own.

B. Source data

B1. Could you confirm the types of electronic source data you produce? (7)

We do some sequencing of genes. We store them ourselves really. When we published it, it wasn't long enough to be a sequence entry so we published variant. We've used various programmes for sequence analysis.

B2. In what formats are these source data held? (8)

Generally we quite like Word and Powerpoint. We have some sequences in DNASTAR but generally most sequences are available through NCBI now so we just get them each time, it's so fast.

B3. How is this data produced?

PCR to amplify a bit of DNA and have it sequenced and then translate from DNA to protein. But mostly what we generate are cell-based experiments, so we'd be looking at motor activity for example. That would end up as a spreadsheet in Excel normally and a bar graph.

B4. How is this data stored?

On our local network in Life Sciences. We're encouraged to use that.

B5. Generally, how large are the files you generate?

Actually I've just asked for more space. I've got 83gb on my PC. The spreadsheets are quite small, it's scanned images that take up the space. Although most of our data is a spreadsheet and bar graph, we run radioactive gels and then we capture them on a store imager and those images take a lot of space. And the scans, scanned blue gels.

B6. Are the data you generate sometimes a combination or group of differing data formats? What formats might be combined? (9)

It's always mixed which is why we use MS Office. Scanned images and photographs next to a histogram and text.

B7. To what extent do you think the data you generate would be useful for other research projects?

Once it's published. Probably not in its raw format. The other thing we haven't touched on is microarray analysis. We haven't published that yet so we haven't deposited that anywhere. I think it's good when it's available as people generate so much data but they can only highlight a few things in a paper, so it's very good to be able to access it.

B8. Would it need any modifying before dissemination i.e. would it be easy for others to use in its raw state?

We run multiple experiments so it's best to have an average summary at the end. We'd definitely run processing on it first. We will have multiple files of experiments where maybe there will be more controls as well, but without an explanation of what they are they won't mean much to people.

B9. Conversely, why might you wish to access source data generated by other research projects? (10)

I would only do it in database format. It wouldn't make sense to me to try and work out how people have lodged their individual data, unless it's in a specific format and you were used to looking at data in that format.

B10. What kind of data?

Sequence databases are out there anyway. Microarrays – I'm not sure how to use those databases – but I'm sure that would be useful. But for a lot of the assays I'm not sure. You couldn't do direct comparisons as experiments are not standardized. It might be nice to look at somebody's data next to yours but they wouldn't have been done in the same way ever, where as DNA sequence analysis is black and white and that is a difference. Microarrays - they're not all the same but they do have a standardized way of explaining what conditions they've used.

What's out there is good.

B11. How do you find and access it? (11)

See above

B12. How might the sharing of this source data be made easier?

See above.

C. Source repositories

Do you submit data to source repositories?

No.

C1. Why did you choose these repositories?

n/a.

C2. How easy was it to submit data?

n/a.

C3. Anything that you particularly liked or disliked about the process?

n/a.

C4. Do you download data from any source repositories?

Oh yes. PDB – the crystal structure of proteins.

C5. How frequently?

Around three or four times a year. Also the Genome database about once a week. We're downloading bits of DNA sequence from a particular gene, normally so that we can get a clone of that gene ourselves.

C6. What were your experiences of this process? E.g. How easy was it to find what you were looking for?

Easy.

C7. How in your opinion could source repositories be improved?

No.

D. Metadata

D1. What metadata fields do you consider important to describe your data? (13)

- Project reference numbers/identifiers X
- Author / data creator name(s) X
- Title of data set X
- Subject keywords X
- Funding source X
- Publisher X
- Dates of project X
- Date (e.g. of data creation) X
- Format (e.g. PDF or HTML) X
- Project description X
- Project title X
- Other

Do these fields describe the resources enough to allow others to determine the file's contents? Are there any other fields missing?

Yes all useful for searching.

D2. At what stage are metadata assigned to your research data? (14)

Other than papers, no. We don't really assign metadata.

D3. Who assigns this metadata? (15)

n/a.

D4. Does your use of metadata vary according to the type of data you submit? i.e. would you use other fields to describe other types of data?

n/a.

D5. Do you know of any standard metadata sets that are used to describe your data?

No.

E. Data access and sharing

E1. What measures do you use to make your research data available? (16)

When we do it's just by email. The local network is available to the research group.

E2. What factors would encourage you to share your research data? (17)

It's really the relevance. If you're collaborating they'd probably rather see it as a histogram than as raw data. I'd possibly just do it on an individual basis if I wanted someone's opinion.

There are also intellectual property questions, which wouldn't bother us as but the university might object.

E3. What factors would discourage you from sharing your data? (18)

Intellectual property. And sharing with whom, that's the question? We'd share with collaborators.

E4. What kind of formal restrictions do you apply to the release and/or access to your research data? (19)

Not really, it's not that commercially sensitive, so we would give seminars and present at meetings. So none really.

E5. What actual practical measures or processes do you use to control access to your data? (20)

See E1.

F. Output repositories

F1. Which output repositories do you use for information to draw on in your research? (21)

- Institutional [eg. Eprints repositories]
- Discipline [eg. arXiv]
- Publisher [eg. ScienceDirect]
- None
- Other

For research I use NCBI as they have the full range of papers. I try getting them straight from NCBI by following the links. Generally, if we have electronic access we get through that way.

To be honest I'm not quite sure what the point of PubMedCentral is as most things seem to be available after a year anyway. I've got a feeling they're just building up to this now with submission systems. There is one that the Journal of Virology, which I'm on the editorial board of uses, but again it means it has to be mounted and set up differently, I'm not even that keen to put my own pdfs on the web because I have to check through all the publishing agreements with all the different journals we've published in to see whether I'm allowed to and whether I can or not, and go through and do it myself and yet they're freely available most of them so all I'd be doing would be duplicating.

F2. And to source material for teaching you use? (22)

For teaching we use ScienceDirect because we can give references to things we have access to, so we look there first.

F3. When making your own research papers available, how do you choose where to publish or deposit?

The best journals we can get into. Definitely the best journal we can get into, whether public access or not. We decide using impact factors and esteem in a particular field. Of course the

one's we can't get are Cell and Nature – with Cell you can't even get back issues, which is very frustrating.

See F2.

F4. In which output repositories do you deposit your research publications? (23)

See above.

F5. Of the output repositories you have used, what were their good or bad points?

NCBI it says free access with a flashing thing and then you get the abstract page but then you have about three or four clicks and it turns out it's not free access and you can pay $30. I think that's a little bit of wasted time! It leads you down a false path.

ScienceDirect works very well for what it does.

F6. How could output repositories be improved?

The clarity of what's free and what's not on the front page.

F7. Would you consider depositing your research papers in an open access institutional repository?

See F1.

I don't really want to have to go through checking whether I'm allowed to or not. If it were done on an institutional basis then that would be OK. It would only be once they were published and passed publishers' deadlines. I don't think they should be locked away in databases and if an institute wanted to make its own papers available then that's OK.

At the moment you have the journal, then you get them different in PubMedCentral and probably different somewhere else. I think it should look the same.

G. Support

G1. Please would you describe the level of support you receive when using output repositories. This can be from individuals or from online links or advice. (26)

None. They all seem very straightforward.

G2. Do you think you are using output repositories efficiently?

Yes.

G3. Might there be features of output repositories that you are unfamiliar with?

Possibly as I've never gone on a training course or done one of the tutorials. I did go on a training course for Ensembl at Hinxton and that was very good. I'm sure there's a lot more on NCBI that I don't know about that would be useful for me to know but I'm happy with what I do.

H. Reprise of project aims – Source repositories

H1. If a standard feature of repositories of source data was the ability to identify and link directly to the to the publications developed from these data, how advantageous would you find it? Why is this? (5)

Can't you do this already? Actually it would probably just take you to the crystallography paper. Yes that would be useful. Embedded links are useful – I use them in Ensembl a lot.

H2. Similarly, if you could navigate directly from within an online article or other text to the primary source data from which it was derived, how advantageous would you find it? Why is this? (6)

Definitely. I've wasted a lot of time trying to do that.

H3. Having now considered both source and output repositories, and how they might relate, what functionality do you consider to be missing from the source repositories you have used? (28)

I guess you could just have more links, so from a Swiss-Prot file and maybe add a link to a secondary structure prediction to see if it's been predicted correctly. Every database could link to every other, but a link with a direct search. Ensembl is good – it doesn't just take you to Swiss-Prot, it takes you to the entry for that particular gene and you can come back again, so it's not just a list of databases at the end, it's databases with the right links.

I guess sometimes genes have strange names – they have their official nam, which isn't like any of their [indistinct] *names so it can take quite a while, especially if it's a gene family, making sure you're actually looking at the right one. If someone had make a link, as long as it's correct and based on an identifier then that would save you a lot.*

H4. We are exploring ways of providing links from repositories of source data to repositories of published papers because we believe there is a need amongst researchers to identify published (and pre-published) papers that have made use of their source data. In what way can you identify with this perceived need? (29)

Yes. If I'd crystallized a protein I would want to see. How would you do this, by looking at published papers and referring back? Because you don't normally quote PDB even though you probably should. If I were referring to someone else's I would refer back to the original paper.

H5. Linking to source data from output repositories will require that an adequate range of metadata are applied to the source data that will persist over time. What sort of difficulties – and solutions to them – might you anticipate when attempting this? (30)

This would be very hard. There are these unique gene names now, so maybe the proteins would have to use same gene names. Basically there's a major problem now with papers not using the same nomenclature. If everyone used the gene name for every database you could link.

I. Reprise of project aims – Output repositories

I1. What functionality is missing from the output repositories you have used? (31)

See F7.

I2. We are considering building an interface for output repositories that would let you as a depositor, associate newly deposited publications with the data from which they are derived. In what way might this be of benefit to you or indeed others? (32)

No because we wouldn't publish the primary data just the summary. Not all the replicates.

I3. A number of new operations could be supported within an output repository, such as the automatic creation of links, the automatic embedding of source repository data and the presentation of relationships (i.e. showing publications and their source data in adjacent windows). How do you think these features could meet your needs? (33)

It wouldn't be useful for us but again I could see it being useful for some people, yes - crystallographers and genome mappers.

I4. What other features might you expect to be advantageous? (33)

No. We just don't generate that sort of data.

J. Reprise of project aims – Potential solutions

J1. A 'dataset knowledgebase' is an online service which allows the creation of two-way links between source and output repositories. It could resolve questions placed in either direction and could also be enhanced through the addition of features such as stored user annotations, quality assessments or ratings and answers to FAQs about specific sets of data held in a repository. What is your opinion of the value of such a concept? (34)

No opinion. It's new.

J2. Are there specific issues you might want it to address? (34)

I can't really imagine it. I would want to see it.

J3. Some data repositories are open to all enquirers while others are password protected. If we are expecting to design links that will provide access from open repositories to controlled repositories, we shall need to devise some level of validation and temporary access rights. Could you describe the extent to which this is necessary in the context of your own source data? (35)

Who would then have to give out the validation? Who do you stop and who do you allow to do it and who ought to decide that? That would be the question really. If it comes down to individual laboratories giving access to people, it would take too much time.