Interview Transcript

Interview reference: Biochemistry 8
Role: Member of academic staff
Interview length: 50 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?

Publish when it's ready!

To start at the very beginning, one has to define projects either as proposals for funding bodies or studentship projects. One often has a rough idea of what might make a publication before one starts but usually one is proved wrong as things don't turn out the way you expect. Usually a piece of what looks like publishable work gradually emerges.

The imperative to publish means that it's not always possible to publish a polished and finished piece. Often if it looks as though it's timely and important you need to get on and publish it. If you have competitors or can sense from other publications the way the field is going then you can't sit on this forever. So the thing may not be as polished and finished as you would have hoped.

The generation of data is all laboratory-based analytical work. We do experiments with isolated mammalian cells – extracts of tissue that undergo biochemical analysis. There is a lot of data collection: because of biological variation, experiments have to be repeated a substantial number of times and analyzed using statistical methods.

B. Source data

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

Numbers! The data we generate would be summed up in tables - which is often the way it's published - or simply in things like graphs and histograms. Some data are purely visual such as gels and pictures that tell what the story is.

The raw data will just be numbers that come off instruments, like radioactive incorporation data or enzyme analysis data or various other things.

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

I suspect that my students would probably use something like Excel. I'm so old fashioned that – I don't tend to do experiments any more - I personally would have written them down in a lab book! I'm a bit mistrustful of Excel. I wouldn't dream of doing my stats with a stats package, I still use the tables.

B3. How is this data produced?

See A2 and B1.

B4. How is this data stored?

On a machine in the lab. Increasingly, raw data from analytical apparatus is stored on a linked computer, stored on its hard drive and backed up. A lot of our data processing is simply doing repetitive calculations, which of course is very easy in Excel.

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

Kilobytes. Some of the visual gels will be megabytes - depending upon resolution - but we're not talking big stuff.

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

Yes. See B1.

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

Only, I think, in the finished published form. Raw data – I can't conceive of a situation when it would be useful unless of course we were engaged in a piece of collaborative work. I can't conceive of our raw data being any use whatever until we publish it.

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

See B7.

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

It's possible but it might be for curious reasons: sometimes when I begin to analyze other people's papers in depth, I sometimes think wait a minute I don't quite know how they've got here or how they managed to show that's significant, I can't work that out. It would be useful for validation and in that regard to actually see the raw data and make a judgment. Sometimes people stretch the truth - well perhaps they don't but I wonder what's gone on.

B10. What kind of data?

See B9.

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

I haven't tried. I don't try to. I just don't trust the publication.

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

See above.

C. Source repositories

Do you submit data to source repositories?

I think in our case it's not relevant. Our repository data is essentially formed of postdoc notebooks: then when people leave they leave their notebooks behind.

I've never really had a need as we try to keep up and publish as we go.

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?

Sequence data - with one paper that I was rather peripherally involved with. I'm not sure which one it was.

C5. How frequently?

See C4.

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

n/a.

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

n/a.

D. Metadata

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

I don't really use metadata.

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

n/a.

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)

I don't think anyone has ever asked me for the raw data apart from collaborators. People ask methodological questions - to provide more details on how you did a particular analysis, that's more common. Usually because they can't access the publication themselves.

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

There's no point in sharing until one's put it together into a coherent body, which is a publication… or a conference. I've shared data in a sense pre-publication, at conferences or giving research seminars. But you're careful what you give out.

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

See above.

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

Yes, in something like a research seminar or maybe having an informal discussion with someone at a seminar, you might say well actually we're having some trouble repeating this experiment, or we're having some trouble because they've changed the strain of animals and we're not getting the same results. That sort of thing. But obviously, all those sorts of things are not in the publication.

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

It's all kept in house.

See above.

F. Output repositories

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

It's essentially all through published research papers or conference proceedings. Interestingly I realize that I haven't actually been into the library for three years as everything is electronic. The only time I have been in the library is when something is not available electronically and I've lost the photocopy.

I seem to have developed a fairly effective series of keywords – this is not only accessing published information for my immediate research area and background, but also for teaching - every three months or six months I use them [the keywords] *to go through PubMed and check what's happening, say in 2006. I print it off and then literally whiz through – these are papers I'm not interested in, these are papers I am interested in. I used to do this through Medline through the College library, but found it very cumbersome and full of stuff I didn't want. This is so simple – I just save it in an EndNote library and I can call up the URLs, save it, annotate it and share it with my research group. No visits to the library, no photocopies, no messing about – it's wonderful!*

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

For research-led teaching in the third year, it's exactly the same process.

I've a set of tried and tested keywords. PubMed suits me. As I say, I used to try and do this through Medline in the library, but the system kept changing. You got used to the system and it changed and I just thought I can't work with this.

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

You start with the highest profile journal you think you can get into and work from there really. Not every single paper you produce is earth-shattering but some are and some papers will have more impact so you try the higher impact journals. The tendency is to try to publish everything in the Biochemical Journal – the journal of the British Biochemical Society, which is the most respectable, rigorous, journal but it's not as sexy or high impact as some of the others.

We're most likely to publish in the Biochemical Journal, what was known as the European Journal of Biochemistry but is now called the FEBS journal, and I have published in the Journal of Biological Chemistry. There's one other - FEBS letters for short publications. Those are the main ones.

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

See F3.

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

Yes, one or two: there is a journal called Diabetes, which the College has a subscription to but I don't have a personal subscription so you can't access it electronically until six months after publication. So, I either have to wait six months to read it on the screen or I've got to go to the library to read it. That's very irritating actually. There are one or two other journals but I can't remember which ones that you can't actually get at unless you have a personal subscription.

I haven't come across any problems with getting the archive stuff, just one or two journals that are hot off the press or maybe stuff which is appearing in pre-publication format. It's just one or two.

F6. How could output repositories be improved?

The timeliness of these one or two journals. That's all really. UCL at least in my area has a good range of subscriptions. Reading other publications, one would soon become aware if there were huge gaps.

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

I haven't really thought about it. I didn't know about it. [Ensuing discussion about institutional repositories, how to deposit, benefits thereof.]

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)

I've never needed to. I'm using Endnote which I paid for out of my research grant, which is supported by UCL IT staff.

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

Yes. It seems to me the most efficient way I've ever come up with accessing [publications]. *I feel pretty good about it actually.*

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

I can't think of anything. Basically I'm accessing things to keep up with the literature and of course to produce bibliographies for teaching and publishing papers.

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 publications developed from these data, how advantageous would you find it? Why is this? (5)

Probably not because trying to keep up in biochemistry - an area which is jolly fast moving - you've got enough frankly to read and I think the publisher provides a digest. So I think you

can take things on trust. I've given you an indication already why one might have suspicions about the rigour of certain pieces of work, but I think it can get too fine grained.

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)

I suppose it would be useful if you did want to. I wonder if people did obtain source data, whether this could actually encourage fraud, or encourage plagiarism or encourage people to steal and reuse data – I don't know, am I being a bit funny about this? On the other hand of course, if the data is all actually there out in the open people can then access and see if it's bone fide.

I just don't know what some raw data would actually mean really. There are various stages in the data generation process and what I'm doing, you get raw numbers come off an analytical instrument, which then have to be multiplied up or divided or manipulated in various ways to then give the data that's in the final format. And then you have to repeat the experiment several times and then you would take the different sets of data in the final format and then all the statistics and all the rest of it and then you have something meaningful. Going right back to providing the analytical machine raw data would be gibberish.

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)

None.

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)

Not relevant.

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)

Don't know.

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)

It would be a matter of having the time. It's a low priority.

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)

This is relevant but again it's time. One has to generally take the authors' finished product on trust. Probably in the last nine months or so, I've probably read two or three hundred research papers. To go more fine-grained than that would just be too much. Almost invariably, if a piece of work is sound it's a launch pad for the next piece of work. Someone else will almost certainly repeat it and confirm. Most things of interest are actually confirmed more than once.

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

None.

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)

Don't know.

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

No.

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)

n/a.