StORe

Interview Transcript

Interview reference: Biochemistry 2
Role: Member of academic staff
Interview length: 47 mins
Format: Telephone
Questionnaire respondent? Yes

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?

We do proteomic analysis. For example we take a mutant and a wild-type bacteria, culture them and compare the proteons using 2D-gel Electrophoresis. Modulated results give data sets which we put into tables. Each line of the table would compare the mutant protein with the wild-type protein. This is then analysed in house. We dump the dataset on a server at Cambridge and it's made publicly available when the corresponding paper is submitted for review (as the reviewers will need access to the data).

B. Source data

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

[The interviewee had already indicated via the questionnaire that s/he produced: statistical data; and functional genomics data.]

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

Don't know. This is done by the post-docs. It's largely tabular data i.e. raw data that has been run through a statistics programme. We run both univariate and multivariate analyses on the data.

B3. How is this data produced?

See A2 and B2.

B4. How is this data stored?

See A2: *stored as files on a university server.*

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

Don't know.

B6. You indicated in the questionnaire that you sometimes generate a combination of differing data formats. What formats might be combined? (9)

We're looking at this at the moment i.e. combining metabolomic and proteomic data. Using one set of data and comparing with the other to enable us to see links.

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

This is an interesting question. All data should be useful. The problem is with the metadata: most researchers don't routinely include the type of detail that is necessary to make the data useful, but if people wanted to cross-correlate then there is some utility there. Most people are not aware of the detail needed e.g. time of day the data was harvested.

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

It should be easy. Anyone with a Mac or PC should be OK. Our data typically is an Excel spreadsheet with numbers in it. However, I am not keen on dumping data: it would have to be data we would be happy to stand by.

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

Increasingly I don't. In functional genomics, minor variations in experimental protocol can lead to major differences in results. On the plus side there is beauty in diversity I suppose, if I knew the full metadata, for example if I knew the rate at which the flask was shaken. On the minus side it makes it difficult to compare data: there are variations in the way different labs do things. I wonder at the validity of the data.

We would use data if we felt it was usable.

B10. What kind of data?

n/a

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

There is now a functional genomics database of the organism I study. This was created and curated by the community and the data is peer-reviewed. It's called PseudoCAP [Pseudomonas aeruginosa Community Annotation Project http://www.cmdr.ubc.ca/bobh/PAAP.html]*.The community got together to make sense of genomic information and agreed to submit peer-reviewed data. We use it everyday. We haven't deposited anything yet, but we have been invited to and will do so soon, probably in the next two or three months.*

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

See B11.

C. Source repositories

In the questionnaire, you indicated that you submitted data to: Protein Structures Database. (12)

Which Protein Structures Database in particular?

PDB.

Is there any particular reason why you have only done so once?

I have only solved one crystal structure!

C1. When you did submit, why did you choose this repository?

It's the only one in town.

C2. How easy was it to submit data?

Very easy.

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

No, it's very easy – like greased lightening!

C4. Do you download data from any source repositories?

Occasionally. We do so as a research group. From the Protein Data Bank through the department server.

C5. How frequently?

About once a month from PDB. We access Pseudo-CAP every day and download often.

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

It's very easy.

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

A year ago, I would have said better links to one another but now it's good.

You've mentioned metadata as a problem already. Is this applicable here?

Good metadata is not as important for data in the Protein Data Bank. It is however for functional genomics websites. The difficulty is what to include (if you ask for too much people will be put off submitting), and how to access it. For example if we're talking about mutants in luria broth, people are not really interested so generally we don't have enough information on growth conditions. This is really important in metabolomics where the cell responds to different conditions.

D. Metadata

NB. See comments above.

[The interviewee had already indicated via the questionnaire that the following highlighted metadata fields were important to describe his/her data] (13):

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

Time of day, Day of week (I'm not joking here - some studies have shown that researchers collect data differently when the weekend approaches...), ALL relevant experimental data, Room where data were collected, proximity to other experiments etc. All of the above is important for eg, metabolomics studies.

These comments taken from the questionnaire.

D1. Do these fields describe the resources you deposit accurately enough to allow others to determine the file's contents? i.e. are there any fields missing?

See comments above.

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

[The interviewee confirmed that this was correct.]

D3. Who assigns this metadata? (15)

Whoever collects the data! (why is this (most obvious) option not listed above?!!!)

[The interviewee confirmed that this was correct.]

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?

Yes. Metabolomic data would require a more in depth description than proteonomic data.

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

This is an area of active research i.e. how to incorporate metadata and identify key types. [He then gave me the contact details of someone actively working in this field.]

E. Data access and sharing

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

[The interviewee had already indicated via the questionnaire that: data are exchanged by email; and that those who ask, receive (most of the time).]

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

[The interviewee had already indicated via the questionnaire that all of the factors listed on the questionnaire would encourage him/her to share research data*, provided that valuable time did not have to be wasted in dumping the data on an accessible site. In my opinion, ALL data should be held confidential until it has been peer-reviewed (not all data is good data, and I would like to be sure that any data I dump or access is of high quality)*.]

Maybe I operate a naïve policy in that I feel publicly funded research should be publicly available. I generally release data to anyone who asks, provided they give a valid reason for wanting access to it. The exception would be if our direct competitors asked for it, however they never do!

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

Increasingly IPR considerations. If our data supported a patent then I may be cagey about sharing it.

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

[The interviewee had already indicated via the questionnaire that: individual enquiries/requests for access are judged on their merits.]

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

[The interviewee had already indicated via the questionnaire that: I only let people that I know (and trust) have access to our pre-publication data.]

You mentioned earlier that your data is uploaded to a server in Cambridge. Would the data then be openly available or restricted in some way?

I'm not sure. If the data is referred to in a publication, then it is openly available but I'm not sure prior to that as I've never tried to access it.

F. Output repositories

F1. I'd like to clarify the types of output repository you use to find and retrieve information for use in your research (21)

- Institutional [eg. Dspace @ Cambridge repository]
- Discipline [eg. arXiv]
- Publisher [eg. ScienceDirect]
- None
- Other

I usually use PubMed and follow the links through to resources like ScienceDirect. I recently tried Google Scholar but I didn't like it as much. My search procedure in PubMed is usually using a keyword plus the author's name if I know it, then following the link to the publication.

F2. Which output repositories in particular?

See F1.

F3. And to source material for teaching you would use? (22)

I would very rarely use published material for teaching. I tend to put teaching material together myself as it avoids all the copyright problems.

F4. When making your own research papers available, how do you choose where to publish or deposit?

I publish in the most appropriate forum for the area, in the journal that promises the widest distribution and that is also reputable. It's getting a balance between these three considerations.

F5. In which output repositories do you deposit your research publications? (23)

None.

F6. Of the output repositories you have used, what were their good or bad points?

I don't really know as I always just use PubMed.

F7. How could output repositories be improved?

n/a

F8. Would you consider depositing your research papers in an open access institutional repository?

Yes, why wouldn't I? I want my paper to be read. Provided that the medium satisfies the three criteria I've already mentioned then I would be happy to.

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 haven't really used any support. Not to denigrate the efforts of our librarians, but I think PubMed is really easy.

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

In as much as one can use the web efficiently, yes. I might ask for the ability to do more focused searching, but this is a minor point.

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

Sure, loads. I'm quite happy with the functions I use already. They meet my needs.

H. Reprise of project aims – Source repositories

H1. You indicated on the questionnaire that if it were possible to link from repositories of source data to the publications developed from this data, it would be of "significant advantage to your work". Why is this? (5)

One always likes to look at the source data. In functional genomics, it's not really the vogue to publish the raw data but it is really important. For example, 2D-gel electrophoresis data is always worth looking at in terms of quality. The supplementary information to published articles is always worth looking at.

[N.B. The interviewee placed emphasis upon the output-to-source, rather than the source-to-output direction that the question specified although he did comment that *"there is no reason not to have two-way linking".*]

H2. You indicated on the questionnaire that if it were possible to go directly from within an online publication (electronic journal article or other text) to the primary source data from which that publication was developed would be of "interesting but not particularly useful". Why is this? (6)

See above H1.

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)

Nothing really. Websites are very well linked, so well linked in fact that you don't need to worry about their reliability. Authors have highlighted relevant links.

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)

I can identify with it. I imagine that it could have a significant implication for funding if you can show that others are making use of your data. It would be an ego trip too! You would need a

way of showing use by others. There was a website called Mutant Bank that you could use to dial up any mutant. It's now folded but if they had had the ability to generate statistics showing how many people were using it and from where then it might have still been available.

H5. Linking to source data from output repositories will require that an adequate range of metadata is 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)

I'm not really the person to talk to about this. There was an international symposium on this just recently.

I. Reprise of project aims – Output repositories

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

[The interviewee had already indicated via the questionnaire that s/he would like to see: unified format; unified metadata.]

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 an incremental advance. It's a difficult question. One would always like more peer-reviewed data but then we've managed so far. Authors publish critically therefore do you need the data too? Probably yes in terms of functional genomics data but prior to this I would say no as it would be diluting the power of authors to criticize their work. There are more places to publish now and I think there has been a real decline in publication quality. I wouldn't want to exacerbate this further.

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)

Useful. It sounds like an extension of when you link to supplementary tables from an electronic version of a journal paper. As long at the data is peer-reviewed and is of high-quality then it would be useful.

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

I'm not a computer jockey! I think it's fantastic at the moment. There are already good and relevant links.

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 respository. What is your opinion of the value of such a concept? (34)

It would be useful, yes. As long as it's not too complex and is simple a la PubMed.

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

It would really depend upon the particular study.

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)

In terms of IPR, I'm happy about open access to published stuff. We wouldn't deposit anything on a publisher's or limited access website unless we thought there were no IPR issues.