Interview Transcript

Interview reference: Biochemistry 3  
Role: Member of academic staff  
Interview length: 48 mins  
Format: Telephone  
Questionnaire respondent? Yes

A. Identities

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

Biochemistry – chemical methods in 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, we create the materials that we want to study. We run instrumental spectroscopy. We then spend some time thinking and looking at the spectra. We would then look at the next set of samples. We’re not doing routine analysis but instead we’re looking at something interesting in the data. There comes a point at which we have to decide that enough is enough and at this point we would write a paper and submit figures.

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: lists of enzymes; spectra; raw data; photographs; instrument data; images; gene/protein sequences; drawings; plots; derived data; databases.]

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

[The interviewee had already indicated via the questionnaire that data were held as: word processed files; tables/catalogues; spreadsheets; rich text files; pdfs; plain text files; image files; HTML files; flat files; database files; and XML.]

B3. How is this data produced?  

Through EPR spectroscopy. Sometimes other biophysical methods.

B4. How is this data stored?  

On my own PC as computer readable files. The spectra are stored along with data relating to the instrument parameters.

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

They’re very small. Generally the spectra are about 4kb. It’s the secondary datasets that are much bigger.
B6. You indicated in the questionnaire that you occasionally generate a combination of differing data formats. What formats might be combined? (9)

With spectra, data tend to be produced in proprietary formats peculiar to the instrument. If we then use a new instrument, it can’t read the older data. We have to consider the compatibility of formats.

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

I think very. It represents a lot of work. Sometimes we want to find someone else’s spectra, but if it’s not in the field, then we can’t. The analysis of the data is published but often the same protein has the same EPR. It could be the same spectra but you often can’t tell. Having access to the spectra itself is therefore important.

Within biochemistry there has been some work done on open source formats which are non-proprietary, using generic software. The data would be stored as compressed ASCII files alongside other data, such as how the machine was set up. Several formats have been created by instrument manufacturers and the software to convert spectra from other machines doesn’t always work, therefore we have to write converters or if we’re not careful we will lose data.

Because I’ve been converting my old data to new formats, I now have thousands of spectra on file.

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

My files are stored in JCAMP-DX format, which can be read by anyone. Java can convert JCAMP-DX files to spectra on your machine.

[The interviewee has produced a standards report for EPR spectroscopy in which the type of metadata needed for this work was discussed- Pure Appl. Chem., Vol. 78, No. 3, pp. 613–631, 2006.]

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

This is very little done in chemistry and there is much reinvention of the wheel. Papers are necessarily compressed therefore much data will not see the light of day.

B10. What kind of data?

Spectra.

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

I’d probably write to someone although I’ve never really done this. It’s a bit rude.

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

It would be a matter of asking the community although many would ask ‘why would anyone want to do this?’ There isn’t such a culture within chemistry and so currently there is no systematic way of making this data available.
The best way would be to have a central repository and give recognition to those that have deposited. Even then some will and some won’t.

C. Source repositories

In the questionnaire, you indicated that you have never submitted data to a source repository. Can you explain why? (12)

Basically there isn’t one but also there are issues regarding indexing, cataloguing and metadata.

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?

No, not at the moment.

C5. How frequently?

n/a.

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

n/a.

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

This is starting now, for example the ACS are looking at it. There are copyright problems however when really it would be better if the process were transparent – if there were no application forms when one clicked on a link. It has to be quick and simple.
D. Metadata

[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)
- Title of data set
- Subject keywords
- Funding source
- Publisher
- Dates of project
- Date (e.g. of data creation)
- Format (e.g. PDF or HTML)
- Project description
- Project title
- Other X - Reliability e.g. the calibration of instruments

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?

Instrument parameters. I can’t think that any of the above are not needed. Standards and nomenclature are vital – it’s important that everyone gets on board. The RSC have produced an extensive report on data deposited within journals.

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

Prior to data creation; as part of the indexing process for source data files; during file saving.

[The interviewee confirmed that this was correct.]

D3. Who assigns this metadata? (15)

I decide which terms to use and I assign them; metadata are generated automatically.

[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?

All the above are generic and a lot are automatic e.g. date of submission. More automatic creation of metadata would help.

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

Bruker Spectrospin has software with increasingly complex metadata. I think NMR data will be the model. The second generation of JCAMP-DX is based on pulsed NMR metadata.
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; through the exchange of portable media; by the provision of a publicized URL; through a source repository.]

NB I queried the latter method given that s/he has not submitted to a source repository: There was a database of software for manipulating spectra but the new generation of software meant that these no longer worked. However there is now enough computing power to make use of these programmes. The repository in question was held by the NIMR but it’s now defunct.

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

I would be happy to share data anyway but in more sensitive areas one would hang on until publication.

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

There isn’t a sharing attitude in the U.S. The U.S. also jealously guards software for sale.

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

There are no formal restrictions but the institution is concerned about impact.

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: most people wouldn’t know about it, unless I tell them.]

F. Output repositories

F1. In the questionnaire, you indicated that you used

- Institutional X [eg. King’s eprints repository]
- Discipline X [eg. arXiv]
- Publisher [eg. ScienceDirect]
- None
- Other

NB. On questioning the interviewee it transpired that these questionnaire replies were not valid and the interviewee had not in fact used either institutional or discipline repositories to source information.

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

ScienceDirect. I’m also interested in having a look at MIMAS.
F2. And to source material for teaching you used? (22)

I use my own data. EPR is an obscure technique so no good teaching resources are available. A WIKI would be useful.

[N.B. Despite noting in the questionnaire that s/he used institutional, discipline and publisher repositories to source material for teaching, this is not in fact the case.]

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

Impact factor. The hot journals are controlled by the ACS. I don’t go for Open Access journals at the moment.

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

I’m thinking of putting some things on the institutional server. The problem is that once you’ve written one paper, you’re straight on to the next, so there’s little time.

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

n/a.

F6. How could output repositories be improved?

n/a.

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

I have no qualms about this but it can be problematic as you have to sign copyright over to journals and they hang on to it. Elsevier and the ACS don’t take kindly to this sort of thing.

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.

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

It worries me. Three or four years ago I got some data from MIMAS but I didn’t really know what I was doing and there was no one to help. [S/he couldn’t remember which resource in particular.]

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

Yes, and I do need help.
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? (9)

People would then be able to verify what you said. I sometimes wonder if people don’t like you seeing data as there’s something wrong with that data, although in our area very few students can fake results as you can’t fake a spectra.

[N.B. The interviewee placed emphasis upon the output-to-source, rather than the source-to-output direction that the question specified.]

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 “significant advantage to your work”. Why is this? (6)

Publication to source for verification purposes would be good. You might also think of different things to do with the data although no patents really come from my area.

[N.B. the interviewee was very keen on the publication to source route.]

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)

Visualization and simple upload/download facilities. The software is not well organized and databases require maintenance.

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. It would be very nice if we could deposit spectra and then if others were to republish using that data, it shows that the research has value.

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)

The biggest thing is standard metadata in a maintained format. JCAMP-DX is now up to version 5 but it’s backwards compatible. To apply JCAMP-DX generally we have to get computing manufacturers on board and encourage them to enable their software to use it.

I. Reprise of project aims – Output repositories

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

A good catalogue and good search facilities. I often run searches which come back with no results when I know there should be.
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)

Yes. It’s a bit like supplementary material. The paper represents an interesting sample of which there are more data, therefore where would we put this? You would need to be able to describe spectra well.

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)

I don’t know if I would use raw data for a spectra. This is more of a research too. It would be useful for a broad overview.

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

None I can think of.

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)

Yes. It sounds like a valuable tool, especially if I was starting my career again!

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

The interpretation of spectra, how data are used – this would be tremendously useful.

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)

Personally I hate them all. Passwords and secrecy are unnecessary and only trip you up when you’re trying to access something. The biggest trip up is when it asks you for your credit card details!