StORE: Source to Output Repositories

Biochemistry report

Suzanne Tonkin
UCL Library Services
Gower Street
London WC1E 6BT
Biochemistry departments at UK universities were identified by using the list of departments submitting to the 1996 Research Assessment Exercise under Unit of Assessment 12: Biochemistry available at http://www.hero.ac.uk/sites/hero/rae/rae96/1_96/t12.html. The 1996 listing was used as this distinct Unit of Assessment for Biochemistry was discontinued during RAE 2001.

A total of 10 departments were selected and the contact details of members of academic and research staff and research level students were identified using the institutions’ websites. The questionnaire was sent by email to a total of 505 contacts (this figure excludes ‘bounced’ messages) during March 2006 (see Appendix *), with a follow up email sent to non-respondents a week before the closure of the questionnaire in the first week of April 2006.

**Questionnaire responses**

There proved to be some difficulty identifying which disciplines certain respondents belonged to, particularly between Biosciences and Biochemistry. After some discussion 46 respondents were identified as Biochemists. This included 2 respondents from institutions not included on the original list of contacts.

<table>
<thead>
<tr>
<th>Role of respondent</th>
<th>Number of responses</th>
<th>% of total Biochemistry responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>University academic staff</td>
<td>32</td>
<td>69.9%</td>
</tr>
<tr>
<td>University research assistant / post-doc</td>
<td>3</td>
<td>6.5%</td>
</tr>
<tr>
<td>Postgraduate students</td>
<td>8</td>
<td>17.4%</td>
</tr>
<tr>
<td>Contracting researcher</td>
<td>2</td>
<td>4.3%</td>
</tr>
<tr>
<td>Independent researcher</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>1*</td>
<td>2.2%</td>
</tr>
</tbody>
</table>

*This respondent was an academic-related member of support staff (Bioinformatics Officer)

*Distribution of Biochemistry respondents by role*
Biochemists’ perceived value of bidirectional links

<table>
<thead>
<tr>
<th>Perceived value</th>
<th>Source to output</th>
<th>Output to source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant advantage to my work</td>
<td>57</td>
<td>59</td>
</tr>
<tr>
<td>Useful but not of major significance</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>Interesting but not particularly useful</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Not sure at this point</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Question 7

Source repositories submitted to by biochemists

<table>
<thead>
<tr>
<th>Source repository</th>
<th>Percentage of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenBank</td>
<td>22.8</td>
</tr>
<tr>
<td>BioMagResBank</td>
<td>12.0</td>
</tr>
<tr>
<td>Protein Structures Dbase</td>
<td>30.4</td>
</tr>
<tr>
<td>Brookhaven Nat. Labs</td>
<td>5.4</td>
</tr>
<tr>
<td>Other</td>
<td>9.8</td>
</tr>
<tr>
<td>None</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Exactly half of all respondents had never submitted to a source repository. This may be because the respondent does not generate the raw or source data that could be deposited within the banks that already exist. Others bemoaned the lack of source repositories for their research area. One of these stating that “As a Bioinformatics lab, we don’t generate raw data that could go to these databanks. We create derived data, online resources, etc”
Of those that had submitted to one or more source repositories, when compared with the more frequency-specific responses of question 8 it appears that the responses to question 7 are unreliable: 8 respondents stated that they used GenBank in question 7, however using the responses to question 8, a count of the number of biochemists that had submitted to GenBank at least once results in a total of 13. The table above therefore takes an average over responses to both questions.

Submissions under other included Uniprot, microarray databases, NERC, and the Consortium for Functional Glycomics.

Biochemists' perceived value of source to output links relative to frequency of submission to Protein Structures Database

<table>
<thead>
<tr>
<th>Frequency of submission</th>
<th>Significant advantage to my work</th>
<th>Useful but not of major significance</th>
<th>Interesting but not particularly useful</th>
<th>Not sure at this point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequently</td>
<td>6.5</td>
<td>2.2</td>
<td>6.5</td>
<td>2.2</td>
</tr>
<tr>
<td>On several occasions</td>
<td>6.5</td>
<td>2.2</td>
<td>6.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Once</td>
<td>4.3</td>
<td>28.3</td>
<td>13.0</td>
<td>6.5</td>
</tr>
<tr>
<td>Never</td>
<td>4.3</td>
<td>2.2</td>
<td>6.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Never but intend to do so soon</td>
<td>2.2</td>
<td>2.2</td>
<td>6.5</td>
<td>6.5</td>
</tr>
<tr>
<td>No answer</td>
<td>13.0</td>
<td>6.5</td>
<td>2.2</td>
<td>2.2</td>
</tr>
</tbody>
</table>
Biochemists' perceived value of output to source links relative to frequency of submission to Protein Structures Database

Frequently | On several occasions | Once | Never | Never but intend to do so soon | No answer
--- | --- | --- | --- | --- | ---
6.5 | 6.5 | 2.2 | 13.0 | 2.2 | 8.7
2.2 | 4.3 | 2.2 | 28.3 | 2.2 | 15.2

Frequency of submission

Significant advantage to my work | Useful but not of major significance | Interesting but not particularly useful | Not sure at this point
--- | --- | --- | ---

GenBank

Biochemists' perceived value of source to output links relative to frequency of submission to GenBank

Frequently | On several occasions | Once | Never | Never but intend to do so soon | No answer
--- | --- | --- | --- | --- | ---
2.2 | 2.2 | 2.2 | 2.2 | 2.2 | 4.3
2.2 | 2.2 | 2.2 | 2.2 | 2.2 | 15.2
8.7 | 6.5 | 26.1 | 2.2 | 15.2

Frequency of submission

Significant advantage to my work | Useful but not of major significance | Interesting but not particularly useful | Not sure at this point
--- | --- | --- | ---
Source data

The questionnaire responses from biochemists indicated that the following data types are generated in the course of their research:

<table>
<thead>
<tr>
<th>File types</th>
<th>No. of biochemistry respondents</th>
<th>% of biochemistry respondents</th>
<th>% of all respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>1</td>
<td>2.2%</td>
<td>3.7%</td>
</tr>
<tr>
<td>Databases</td>
<td>15</td>
<td>32.6%</td>
<td>48.5%</td>
</tr>
<tr>
<td>Derived data</td>
<td>17</td>
<td>37%</td>
<td>33.4%</td>
</tr>
<tr>
<td>Drawings, Plots</td>
<td>26</td>
<td>56.5%</td>
<td>57.2%</td>
</tr>
<tr>
<td>Gene/protein sequences</td>
<td>21</td>
<td>45.7%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Geophysical data</td>
<td>1</td>
<td>2.2%</td>
<td>7.1%</td>
</tr>
<tr>
<td>Images</td>
<td>26</td>
<td>56.5%</td>
<td>51.7%</td>
</tr>
<tr>
<td>Instrument data</td>
<td>12</td>
<td>26.1%</td>
<td>24.1%</td>
</tr>
<tr>
<td>Photographs</td>
<td>12</td>
<td>26.1%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Plans, Maps</td>
<td>1</td>
<td>2.2%</td>
<td>15.6%</td>
</tr>
<tr>
<td>Qualitative questionnaire data</td>
<td>2</td>
<td>4.3%</td>
<td>9.8%</td>
</tr>
<tr>
<td>Quantitative questionnaire data</td>
<td>2</td>
<td>4.3%</td>
<td>10.3%</td>
</tr>
<tr>
<td>Radiographic data</td>
<td>5</td>
<td>10.9%</td>
<td>2.9%</td>
</tr>
<tr>
<td>Raw data</td>
<td>19</td>
<td>41.3%</td>
<td>42.1%</td>
</tr>
<tr>
<td>Remote sensing</td>
<td>0</td>
<td>0</td>
<td>4.0%</td>
</tr>
<tr>
<td>Spectra</td>
<td>23</td>
<td>50%</td>
<td>32.9%</td>
</tr>
<tr>
<td>Statistical data</td>
<td>10</td>
<td>21.7%</td>
<td>34.2%</td>
</tr>
<tr>
<td>Synthetic data</td>
<td>2</td>
<td>4.3%</td>
<td>12.7%</td>
</tr>
</tbody>
</table>
## Types of source data produced by biochemists

These responses indicate that, as expected, the file types that biochemists produce more than average are gene/protein sequences and spectra.

Biochemists also produce a relatively large proportion of ‘other’ file types. The free text entries from biochemists identified the following additional types:

<table>
<thead>
<tr>
<th>File types</th>
<th>No. of biochemistry respondents</th>
<th>% of biochemistry respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crystallographic diffraction data</td>
<td>1</td>
<td>2.2%</td>
</tr>
<tr>
<td>FLASH-based animations/movies</td>
<td>1</td>
<td>2.2%</td>
</tr>
<tr>
<td>Functional genomics data</td>
<td>1</td>
<td>2.2%</td>
</tr>
<tr>
<td>Lists of enzymes</td>
<td>1</td>
<td>2.2%</td>
</tr>
<tr>
<td>…results from model simulations stored in a binary self describing standard data format (netCDF)</td>
<td>1</td>
<td>2.2%</td>
</tr>
<tr>
<td>Molecular structures</td>
<td>1</td>
<td>2.2%</td>
</tr>
<tr>
<td>Protein-protein interaction data</td>
<td>1</td>
<td>2.2%</td>
</tr>
<tr>
<td>Protein structures</td>
<td>4</td>
<td>8.7%</td>
</tr>
<tr>
<td>X-ray diffraction data and images</td>
<td>1</td>
<td>2.2%</td>
</tr>
</tbody>
</table>
When biochemists were asked ‘Are the data you generate sometimes a combination of different data formats?’ the responses indicate that on average researchers within biochemistry are slightly more likely than average to combine different data formats:

<table>
<thead>
<tr>
<th></th>
<th>Biochemistry</th>
<th>All respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Often</td>
<td>20</td>
<td>143</td>
</tr>
<tr>
<td>Sometimes</td>
<td>15</td>
<td>134</td>
</tr>
<tr>
<td>Rarely</td>
<td>8</td>
<td>57</td>
</tr>
<tr>
<td>Never</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>Potentially</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

*Bar chart showing the range of file types produced by biochemists and their relative levels of use*
Support

24

Type of support received by biochemists in their use of output repositories

<table>
<thead>
<tr>
<th>Role</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic staff</td>
<td>10.9%</td>
</tr>
<tr>
<td>Research Assistant</td>
<td>21.7%</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>26.1%</td>
</tr>
<tr>
<td>Contract Researcher</td>
<td>6.5%</td>
</tr>
<tr>
<td>Other (please insert)</td>
<td>2.2%</td>
</tr>
</tbody>
</table>

- Documentary support
- Personal support provided by an intermediary
- Repository-enabled support
- No support is provided
- Unknown
- Other

Types of professional support used by biochemists

<table>
<thead>
<tr>
<th>Role</th>
<th>Number of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic staff</td>
<td>13</td>
</tr>
<tr>
<td>Research Assistant</td>
<td>10</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>11</td>
</tr>
<tr>
<td>Contract Researcher</td>
<td>13</td>
</tr>
<tr>
<td>Other (please insert)</td>
<td>4</td>
</tr>
</tbody>
</table>

- Provision of documentation
- Formal training & documentation
- Online / telephone help
- Assistance with structuring of specific searches
- Assistance with the conduct of specific searches
- Full intermediary service
- Unknown
- None

NB. Respondents chose all the types of support that applied.

Interview stage
Of the 46 Biochemistry respondents, 11 indicated that they would be willing to be interviewed. Each potential interviewee was contacted during May 2006 and 10 interviews in total were conducted during June and July 2006. Of these, 5 were telephone interviews and 5 were conducted face-to-face. Each interview was with a single individual apart from the final face-to-face interview, which was conducted with two members of a research laboratory.
Appendices

Invitation to complete questionnaire

Dear Colleague,

I am writing to invite you and your research group to take part in a national survey investigating new ways of linking academic publications with databases of primary research data.

Project StORe (Source to Output Repositories) is a JISC (Joint Information Systems Committee) funded partnership of seven U.K. universities and Johns Hopkins University in the U.S. We anticipate that StORe’s work will benefit researchers by delivering enhanced opportunities for information discovery and for the curation of valuable research data. On a practical level, this will include new methods of tracking the use and impact of your published research, plus the prospect of surveying research publications together with their associated source data across an entire discipline or within a specific research theme.

The first stage of StORe is a national survey of seven research disciplines - including biochemistry - in order to gain some understanding of how researchers currently use databases both of research data (“source” repositories) and published research materials (“output” repositories) and their expectations as to future developments. If you are willing to participate, then we would be very grateful if you would complete the online questionnaire available at:

http://www.survey.lse.ac.uk/store

The questionnaire should not take more than 15 minutes to complete. The closing date is 14th April and all respondents will be eligible for entry into a prize draw to win either an Apple iPOD or book tokens to the same value.

Thank you very much for your time. If you have questions about the survey or the project in general, then please do not hesitate to contact me. Please feel free too to forward this email to staff and postgraduate students in your research group.

With best wishes,

Suzanne Tonkin
StORe Project - http://jiscstore.jot.com/WikiHome
University College London
suzanne.tonkin@ucl.ac.uk
Biochemistry interview script

Interview reference:
Interviewee:
Institution:
Contact details:
Date:
Start time:
End time:
Interview length:
Format:

**Introduction**
Check that the interviewee is happy to start.

Thank you for agreeing to be interviewed today.

Explain what the project is about:
- potential interactions between output repositories of research publications and source repositories of primary research data.
- purpose of the interviews is to explore the needs and opinions of researchers in particular disciplines (inc. biochemistry) in terms of their generation and use of research data and how this relates to published research outputs

Aim to finish within 40 minutes.

Check whether the interviewee is under any time constraints?

Inform the interviewee that notes will be taken throughout the interview, but that any comments will be anonymous when it comes to reporting.

**A. Identities**
For each interview it is important to identify the discipline group and the role of the interviewee to enable subsequent comparisons.

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

- Biochemistry
- Biosciences
- Chemistry
- Physics

A2. Employing organisation

A3. Role

A4. Research process: to set context, briefly describe a typical research project workflow i.e. what steps do you go through e.g. from generating data, then analysing it and eventually publishing. [Prompt A4 if necessary: How is the data created, what do you do with the data you create? How do you go about accessing data, discovering literature and how is this incorporated into your own research?]
B. Source data

Definition of source data = primary research data on which publication will eventually be based.

B1. Could you confirm the types of electronic source data you produce? (7) (Eg. Databases, gene/protein sequences, spectra, text based files)

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

B3. How is this data produced?

B4. How is this data stored? (e.g. files on university/dept network, own pc, source repository, CD)

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

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

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

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

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

B10. What kind of data?

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

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

C. Source repositories

Definition of a source repository = a place where source data produced during a research programme can be stored, maintained and accessed.

How frequently do you submit data to any of the following source repositories? (12)

- Brookhaven National Laboratories
- Genbank
- National Crystallography Service
- Protein Structures Database
- SuperCOSMOS Science Archive
- Uniprot
- None
- Other (please insert)
C1. Why did you choose these repositories?
C2. How easy was it to submit data?
C3. Anything that you particularly liked or disliked about the process?
C4. Do you download data from any source repositories?
C5. If yes, which ones?
C6. What kind of data?
C7. How frequently?
C8. What were your experiences of this process? (eg How easy was it to find what you were looking for?)
C9. How in your opinion could source repositories be improved?

D. Metadata

Definition of metadata = metadata is simply ‘data about data’ i.e. the information or labels that you use to identify and describe your data. Its purpose is to make it easy to recognise, access and retrieve data. The use of standard metadata sets can also help in the sharing of resources.

What metadata fields do you consider important to describe your 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

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 other fields missing? [Prompt if necessary from the suggestions submitted on the questionnaire:

- copyright
- protein sequence
- experimental method
- chemical entity
- software used
- reliability eg. calibration of instruments
- time & date]
D4. At what stage are metadata assigned to your research data? (14)

D5. Who assigns this metadata? (15)

D2. 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?

D3. Do you know of any standard metadata sets that are used to describe your data?

E. Data access and sharing

E1. What measures do you use to make your research data available? (*prompt from list given in questionnaire*) (16)
  - Data are posted or passed by hand in printed format
  - Data are exchanged by email
  - Through the exchange of portable media (disks, CD-ROMs, USB drives)
  - By the allocation of passwords to network drives or data files
  - By the provision of a publicized URL
  - Via a publisher
  - Through a source repository
  - I undertake no measures to may my research available
  - Other

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

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

E4. What kind of formal restrictions do you apply to the release and/or access to your research data? (19) *ie caveats and rules*
  - No formal restrictions
  - Time-related embargoes
  - Restricted to immediate research team/programme members
  - Individual enquiries/requests for access are judged on their merits
  - Data is flagged confidential/commercial-in-confidence (or other caveat), for authorized access only
  - Other

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

F. Output repositories

Definition = an output repository collects together published research outputs e.g. of research papers in the form of journal articles.

F1. What kind of output repositories do you use when looking for information to draw on in your research. (21)
  - Institutional [eg. Eprint repository]
  - Discipline [eg. arXiv]
F2. Which output repositories in particular?

F3. And which output repositories do you use to source material for teaching? (22)

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

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

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

F7. How could output repositories be improved?

Open access = Journal articles and other research outputs such as conference papers can be made freely available by publishing in an open access journal (the author-pays model) or through the deposition of the paper into an institutional repository.

F8. Would you consider depositing your research papers in an open access institutional repository such as the *** Eprints repository?

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)

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

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

H. Reprise of project aims – Source repositories

H1. If a standard feature of repositories containing 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)

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

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)
Older, error prone, entries often remain in repositories even when newer, more validated, data are submitted. There should be some means of directly appending new data to old data files
- Links between different source repositories (not just source-output)
- Source data from a particular experiment organised in date order to map its development

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)

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. Reprise of project aims – Output repositories

I1. What functionality is missing from the output repositories you have used? (31)
- Ease of administration of paid-for repositories
- Ability to generate best-fit superpositions of related structures
- Standard formats for spectra
- Unified formats

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)

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)

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

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)

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

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)
Example interview transcripts
### Biochemistry Scenario 1

<table>
<thead>
<tr>
<th>Title</th>
<th>Depositing data in a source repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>Suzanne Tonkin</td>
</tr>
</tbody>
</table>
| **Narrative**  | John has generated a new protein structure using x-ray crystallography and wishes to publish a paper describing it. The journal to which John submits his paper demands a Protein Data Bank (PDB) entry code as a requirement of publication. John must therefore submit the protein structure to the Protein Data Bank (PDB).

John first uses PDB’s pdb-extract tool to prepare his co-ordinate and structure factor files for deposit; he uploads his co-ordinate file from his final structure refinement and follows the instructions given to him on the page to extract the structure factors and the key details about the structure. John saves these details locally and then uses PDB’s ADIT (AutoDep Input Tool) to deposit his structure; he goes to the ADIT site and selects ‘X-ray’ as his experimental method and ‘Protein’ as his structure type and clicks ‘Begin’. John is then prompted for the co-ordinate file name and the structure factor file name. On the subsequent deposit screens, John is prompted for metadata, categories including but not limited to ‘Release status’, ‘Chemical description’, and ‘Sequence’. John adds the relevant metadata where applicable, but leaves some screens blank. After he has completed each category, he selects ‘Save’ to move him to the next deposit screen. After completing the final deposit screen, John clicks the ‘Deposit’ button and is given a summary of his data, an indication of its completeness and is given the option to either return to the deposit tool to edit his input or to ‘Deposit now’. John chooses the latter and a PDB identifier is presented within the browser window.
Finding and accessing data in a source repository

Suzanne Tonkin

Jane’s research group is conducting a project examining the role of the anticoagulant Protein C in different species. At the beginning of the project, the group requires an accurate definition of Protein C and decides to query UniProtKB/Swiss-Prot via ExPASy. They take this decision as the database provides a high level of annotation including a description of protein function. The group consider the best way to extract all the proteins with the same name; as Swiss-Prot uses identifiers that are constructed using the format - gene name_species - the group decides to use identifiers as their way of retrieving relevant records.

As the group requires information on a specific protein, rather than using a simple search that may generate some irrelevant results, Jane decides to use the more directed Sequence Retrieval System within Swiss-Prot. Jane runs a gene name search on the protein using PROC (the gene name of the protein) as her keyword and specifies that she would like the Swiss-Prot identifier as her output field to enable her to verify the results quickly.

The database presents Jane with a list of results that include identifiers she had expected to see such as PROC_HUMAN and PROC_HORSE, however Jane is also presented with entries for PROC_ECOLI and PROC_MYCTU (Mycobacterium tuberculosis).

Jane clicks the link of one of the unexpected results to obtain the full entry. Jane discovers that although her search was for PROC, her results also contain entries for proc, a different protein. Swiss-Prot will not allow case-sensitive searches and Jane’s group must therefore rethink their search strategy.
Biochemistry Scenario 3

<table>
<thead>
<tr>
<th>Title</th>
<th>Depositing a manuscript for publication*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>Suzanne Tonkin</td>
</tr>
</tbody>
</table>

**Narrative**

John is a biochemist with a paper he wishes to submit to the journal *Biochemistry*, published by the American Chemical Society (ACS). He uses the Society’s Paragon System which allows him to submit his paper online.

John firstly registers with Paragon online, by agreeing to the Society’s ethical guidelines and supplying a userid, password and contact details. He then downloads the appropriate manuscript template from the ACS website. He uses the template to create a new document and copies and pastes the text and graphics of his paper into the relevant sections of the new document. He deletes the template’s sample text as he goes along and finally formats his references using the Vancouver system preferred by *Biochemistry*, abbreviating journal titles by referring to *Chemical Abstracts Service Source Index* and inputting punctuation by reference to the *Guidance for Authors* on the *Biochemistry* website. He saves the document and prints out a proof copy to ensure that all parts of his manuscript are legible and present.

John then submits his manuscript: he completes the ‘Author Contact Information’ and ‘Manuscript Information’ screens and clicks ‘Upload Documents’. Within the ‘Required Documents’ section, John clicks the ‘Browse’ button to open a Windows dialogue box and selects the manuscript’s filename, clicks ‘OK’ and then ‘Attach’. He repeats this process to upload his cover letter. John then chooses the ‘Review and Submit’ option, checks the information he has submitted and submits his manuscript.

* None of the Biochemistry interviewees had deposited papers in an output repository, however one interviewee highlighted the American Chemical Society’s manuscript submission process as a paradigm hence its inclusion and the obvious parallels for submissions to output repositories.
### Biochemistry Scenario 4

<table>
<thead>
<tr>
<th>Title</th>
<th>Finding and accessing publications in output repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>Suzanne Tonkin</td>
</tr>
<tr>
<td><strong>Narrative</strong></td>
<td>Jane is a biochemist starting a new research project examining the functional genomics of a particular organism. She begins by examining the work that has already been done in the area by performing a literature search. She uses PubMed as her starting point and performs a simple keyword search, including the name of an author whose work she has already come across. Jane examines the results of her search and chooses what appears from its citation to be the most relevant and up to date article. Jane clicks the title and is presented with the abstract. The article looks interesting and Jane clicks the ‘See all related articles’ link. Icons indicate that some of these related articles are available free of charge via PubMed Central, some free of charge from the publisher’s repository and others via Jane’s local institutional subscription. Jane works down the list, viewing the abstract of each article and where relevant clicking the icons that link to the full text of that article. She prints out those she wishes to read and exports the citations to EndNote for future reference.</td>
</tr>
</tbody>
</table>
