

Data sharing in the biosciences: A sociological perspective

Outcomes of a Workshop Edinburgh, 26th June 2008

This one-day workshop was sponsored by the BBSRC, and organised collaboratively by BBSRC and the ESRC Centre for Social and Economic Research on Innovation in Genomics (Innogen) and hosted by the National E-Science Centre. The workshop brought together life scientists, bioinformaticians, research policymakers, and social scientists studying these developments, to discuss the implications of technical and organisational change for data sharing in the biosciences. This document summarises the main points of discussion.

What are the main issues and what do we want to get out of the day?

This roundtable session was introduced by Jef Grainger, Robin Williams and Jane Calvert. Jef Grainger pointed to the issues that the BBSRC was confronting with respect to data sharing. Robin Williams stressed his long-standing interest in social scientific research on information technologies, and how this could be combined with research on the life sciences. Jane Calvert drew attention to the 'sociological' dimensions of the issues, such as rewards, credit and incentives, in the context of the broader changes that are taking place in life science research.

All of the participants then introduced themselves and briefly discussed their interests in the issues. The participants included bioinformaticians, policy makers, librarians, social scientists, a bioethicist, scientists from industry, and wet-lab biologists.

Issues that were raised included:

General issues

- the informational turn in the life sciences
- the exponential increase in data production in the biological sciences
- how databases inscribe certain ways of sharing
- who should own and pay for data sharing?

Encouraging data sharing

- how to better engage the scientific community and demonstrate the benefits of data sharing?
- ways of facilitating cultural change
- how to monitor compliance with data sharing?
- dealing with scientists who do not want to share their data
- the potential to address social issues through technical solutions
- how pooling data can also be about building an emerging scientific community

Specific issues

- acknowledging the data creation source
- the administrative load of data sharing and the unpredictability of that load
- the tension between data sharing and intellectual property (e.g. some bodies allow a delay before publication of data to enable patenting, however patentable opportunities may only be identified once the data has already been released and combined with other information)
- the problems of hostile mining of data (e.g. seeking out animal experimentation in existing data)
- how to encourage useful annotation
- retrieving and analysing data produced elsewhere

Tools

- developing tools for data sharing
- using wikis for structured data
- blog-based electronic notebooks
- web 2.0 as research resource
- how to get recognition for tool developers?

Changes in data sharing practices

Introduction to the BBSRC Data Sharing Policy – a funder’s perspective Jef Grainger (BBSRC)

In his presentation, Jef Grainger talked about how there is a great deal of scope for data re-use, which from the BBSRC’s perspective is a value for money issue. He said that the BBSRC wanted to take a leading role in the context of the changing bioscience landscape. The BBSRC’s data sharing policy has been in place for just over a year. Development of the policy has been led by researchers. It started with a consultation in 2004 and drew on existing MRC and NIH policy.

He said that although the BBSRC regards data sharing as a good thing, they recognise that there are still some sceptics. Also data sharing needs are very diverse, since data types and timelines vary radically, so it is not possible to apply a one-size-fits-all approach – best practise is likely to vary considerably across the patch.

He drew attention to the two areas in which data sharing is considered a priority by BBSRC: areas where there is a high volume of data, such as the various ‘omics fields, and areas where there is low frequency of data but long a timescale of research, for example in the field of agricultural land use research.

He noted that in some fields there are established high standards of data sharing, and said these areas could act as aspirational signposts for researchers working elsewhere.

He said that data sharing was science-driven and community-led, but it was important that it is cost effective and high quality. He also pointed out that the BBSRC will support a data sharing component in grant applications (although this does not mean that data sharing is the same thing as long-term archiving, which raises different issues in respect to regulatory requirements).

The BBSRC's policy statement says that there should be timely availability of data (e.g. not later than publication). He pointed out that intellectual property protection should not unduly disrupt or delay data sharing, but he also stressed that the BBSRC was flexible, and that cases could be made for exceptions.

Questions from the audience brought up issues about collaborative awards with industry, and Jef said it was still necessary to address data sharing intentions in these circumstances. The question was raised about whether curators are available, and Jef said that the BBSRC recognised that it is necessary to support curation, even though this type of research is more methods-driven than traditional hypothesis-driven research.

Jef explained how now BBSRC applicants have to submit a data sharing statement – including resources – and that reviewers must address this as part of the assessment of applications. Data sharing must also be addressed in the final report at the end of the award. The question arose as to how to give this policy more teeth for policing of whether applicants fulfilled stated intentions for data sharing. It was noted that BBSRC has set up a data sharing monitoring group, which will be examining the success of implementation of the policy, and applicants' responses to it.

Jef said that the community was responding well, although there was a varying quality of responses, as to be expected, and an element of consternation amongst those just beginning to consider data sharing beyond established scientific publication routes.

Some concerns were raised about the possible quality issues surrounding the release of data to the community that may not have been subject to the rigours of the peer review process, which published data will be subject to. One potential benefit to counter this was that negative (null) data may be shared, which often do not enter the peer review process or the outputs of scientific publications, and this might be a good thing. There is also concern about the possible impacts on the broad administrative burden for researchers and institution, and perhaps some concern about the possibilities for 'hostile mining' of data.

One of the audience suggested that laboratory records cannot be effectively shared if they are recorded only in laboratory notebooks; the use of a LIMS is a precondition for sharing experimental data, for example for protein production.

Future challenges were discussed, including measuring the success or impact of the policy on research behaviour. Is it reasonable to expect best practice? Is this harder for some areas than others? Are more explicit guidelines needed? What are the barriers? What are the unintended effects of policy? What are the consequences of thinking in the context of a global science system, particularly taking India and China into

consideration? For example, is there a danger of a 'one-way street' for data sharing and the associated benefits?

Lessons from the RNAi consortium

Peter Ghazal (Edinburgh)

Peter Ghazal's talk focused on interference RNA (RNAi), a 1998 discovery. These RNAs act as key control mechanisms between the cellular environment and the genome. RNAi has many applications, and the whole area has moved very fast from the basic research to the clinical trails (for example, anti-viral therapy is already in phase 3 trials). It is also possible to do high-throughput genome-wide RNAi screenings.

In the late 1990s there were DNA microarrays in public and private labs, but there were problems when companies tried to mine publicly available microarray data, because the data was dispersed and often in different formats; it was hard to find and even harder to analyse. It was very difficult to compare results from different labs.

The aim of the RNAi global initiative, which involves nearly 30 laboratories, is to bring people together, to develop tools and define which data should be stored. They wanted to set data standards, not about how to do experiments, but to specify what data was necessary to fully describe an experiment. The aim was to make it possible to send and store data and make it available for statistical analysis and data mining.

The data to be stored included both raw and processed/refined data (the results of experiments). Differences between these types of data could be understood through the analogy between the 'pig' (raw data) and the 'sausage' (processed data). Metadata is also needed to track how data has been created and transformed and the implications for how it might be interpreted and used. This includes the experimental context and the properties of the data, to try to explain the sources of experimental variability.

Peter Ghazal raised the issue that you cannot record much of this kind of work on your CV, even though journals now require data to be published (and this is very time demanding). He stressed that cultural change is necessary.

He did say that as databases are used more and become better structured they will become more established. Groups are producing better algorithms and this will improve utility of the information.

Peter Ghazal's group undertook a pilot screen. They looked at 1000 RNAis across 10 labs, and focused on variation in the best assays. They found that there was good reproducibility within the group, but poor reproducibility between groups.

They have designed a conceptual model, formalised it, developed a data exchange format (XML-based) and implemented supporting tools. The standard that they use, Minimum Information About an RNAi Experiment (MIARE), is not a guide for best practice, but should be used to enhance best practice.

Questions from the audience pointed out that if a tool is going to be useful it has to correspond to people's tacit knowledge, but that eliciting the tacit information is hard, and experiments are often hard to reproduce. What are needed are data standards describing the experimental context, which will enable proper interpretation of results,

replication, meaningful comparison of datasets and possibilities for retrospective analysis.

It was pointed out that other benefits of data standards is that they lead to best practice through standard operating procedures because these reduce operator discretion (e.g. for micro-arrays it is necessary to police standards strictly to check that the same results are obtained from different operators). However, a level of alignment is not often achieved and there is often confusion and resistance from the scientific community. Furthermore, peer review often confuses data standards with standard operating procedures.

There was also discussion of how it was necessary to use informatics tools to lower the barriers for other groups, e.g. enabling biologists to use Excel based files for data input.

Peter Ghazal said that there is a learning curve. People start regulating themselves and this leads to pressures on the non-compliant.

Today's challenges

Standards development: Necessarily a Two-Way Street **Chris Taylor (EBI)**

In his presentation, Chris Taylor described how there is a progression from data and metadata to distilled data and then to established knowledge. He said that there are many formats and vocabularies but few standards.

He made the point that a standards-generating body is heavily dependent on the community it seeks to support. One problem is that people are usually not paid for standards development, but are doing it in their spare time on an *ad hoc* basis. Another issue when developing standards is that it is necessary for experimentalists to volunteer their time to assist with design and testing (although users should not need to get involved too much in the internal machinations of new tools). Again lack of resources is the 'rate limiting step' in such scenarios.

On usability: Some tools are built to run on Excel, and others make use of features of Excel – an application many scientists use regularly – so that experimentalists can enter their data into familiar-looking spreadsheets. Another application under development at the EBI, which also employs an Excel-like spreadsheet for some tasks, remembers previous patterns of usage (e.g., previously used ontology terms or spreadsheet structures) in an attempt to reduce workload by automating some aspects of commonly performed tasks.

There are now a number of standards-supportive services available to the community:

- The Proteome Harvest tool (BBSRC) supports a number of proteomics standards
- ArrayExpress scores MIAME compliance to support reviewers
- ProteoRED produces MIAPE-compliant reports
- Minimum Information for a Biological and Biomedical Investigations (MIBBI) provides a shop window for MI checklists and will push for their integration

- The Open Biomedical Ontologies Foundry (OBO Foundry) hosts a number of *de facto* standard ontologies that are being re-engineered to work better together to facilitate the semantic (*i.e.*, content-based) integration of data sets.

This list shows that there is growth in both the number of, and in the level of enthusiasm for Minimum Information (MI) checklists. Because MI checklists are often developed independently they frequently overlap in scope. Those overlaps feature arbitrary differences in wording and structure, making integration difficult. It is desirable to align groups that are developing ‘minimal information’ checklists to avoid duplication and reduce the effort required to effect harmonisation.

Some scientists object to standards such as MIAPE, saying things like: ‘Why should I dedicate resources to others? What’s in it for me? Is this just ‘make work’ for bioinformaticians?’ There is also a perception that there is no money to pay for this type of work (which is incorrect). Furthermore, scientists are concerned about accreditation and citations, and a particular concern is that the criteria of the Research Assessment Exercise (RAE) are such that there is insufficient recognition or rewards for activities that are not directly focussed towards formal publication outputs in scientific journals, such as annotating a data set for release. Other people say ‘I don’t trust other researchers’ data – I’d rather just repeat the research’. And critically, at present there are almost no free mature tools with which experimentalists can generate experimental reports that employ the various (candidate) standard formats, vocabularies and checklists.

There are challenges facing the bioscience community. To increase their efficiency (specifically, where data volumes are large); to extract greater value from data through sharing; and to reduce the risk of loss of information through, for example, staff turnover. It was noted that there is a strong relationship between data sharing and curation: good curation techniques can lead more naturally to sharing of data. It was also noted that the advanced experiences of the ‘omics communities with respect to large-scale data curation and sharing provided potential lessons for other research areas and data types. This included the need to retain high-quality raw data, as derived datasets are more likely to become redundant as new and more sensitive methods to derive the data become available.

Chris concluded that standards are useful because they support deeper scrutiny of data, potentially enhancing both confidence in, and therefore re-use of those data. It is also good management practice to use such standards; for example, by making it easier to go back and check results. However, there are currently few incentives (a small number of sticks, no carrots to speak of) to put data in to repositories.

Data Sharing in Model Organism Biology **Sabina Leonelli (Exeter)**

Sabina Leonelli talked about problems of classification and ontologies in high-throughput biology, drawing attention to some of the factors that can make data sharing difficult.

She argued that the aim of standardization is to make data usable beyond its context of production, and in this way bioinformatics is a service to biology.

She said that a major problem is that disclosure through journals does not work (because journals only provide disseminated claims, which are not the same as

datasets), but disclosure through repositories is not rewarded (even though it is labour intensive). This means that some datasets are kept secret or discarded. Further issues are that the conditions for disclosure not clear (i.e. raw or processed data? Which kind of meta-data?) and IP issues also arise.

Solutions include the establishment and enforcement of clear disclosure policies; shifts in the reward system, currently focused too narrowly on quantity of publications and not taking into account other scientific contributions (for example, data donation to repositories needs to be rewarded as much as publication in scientific journals, a proposal that requires a radical shift from competition to cooperation among data providers); and the establishment of standard disclosure formats for both data and meta-data, a practice that has been partially implemented with excellent results in some model organism communities (such as the Arabidopsis community).

Classificatory problems do arise, particularly regarding the stability of classificatory categories compared to the dynamism and diversity of research practices. The question here is can classification contribute to collaboration without stifling innovation?

Are bio-ontologies a solution? These are controlled vocabularies, they follow rules, they try to establish objectivity, and they rely on the best available knowledge in the field. They are well-curated, flexible and dynamic, although they do require manual curation. The issues that arise with ontologies include: what constitutes the best available knowledge? And how should we institutionalise dialogue between curators and users? (It is not only the tools but also the user behaviour that is important).

In order to institutionalize dialogue it is necessary to introduce enforcement mechanisms, and feedback amongst curators, including curator meetings and blogs. Currently there are bio-ontologies content meetings and user training.

Sabina concluded that it is necessary to institutionalise data sharing and to enforce dialogue among bioinformaticians responsible for the creation and implementation of standards, biologists producing and re-using data, and both private and public funding agencies. Institutions can act as independent platforms to discuss reward systems and disclosure – but it is not clear which institutions could serve that purpose (it could be funding agencies such as the BBSRC, but also prominent journal boards, e.g. Nature, and specific institutes, which could devote more of their budget to ensure that biologists devote time and effort to the sharing of materials).

Breakout session

Breakout group topics

1. How to incentivize, reward and resource data sharing?
2. Risks of data sharing and the dangers of a two-stream system between high-throughput researchers and the rest of the biosciences
3. Peer-to-peer resource sharing
4. Data ownership: the tension between IP, data sharing and credit

Feedback on breakout group 1: incentivising data sharing

This group agreed that data are an important research output, but that data sharing is not adequately rewarded. A central problem with incentivizing data sharing is the tension between inter-personal data sharing and institutional incentives.

They discussed the issue of recognising individual effort. The possibility of tracking people's use of data and offering micro-accreditation was raised, and of perhaps linking this to the RAE.

Another concern was that no one gets credit for tool development, but that these people are needed; curators as well as researchers and research leaders. Curation still revolves around people – but at the moment they have to develop new projects rather than just doing maintenance.

A question that was raised is whether it is adequate to just tweak existing scientific practices, or if it is necessary to create entirely new incentive structures.

It was stressed that it is necessary to create a strategic environment that supports and rewards certain kinds of behaviour, and to be aware of the possibility of unintended effects. There was no consensus on whether a carrot or stick approach is necessary. There was agreement on the point that funding is crucial, but a problem here is that funding is national, even for international projects.

Implications for Policy and Practice

Chair: Robin Williams

This final discussion was a round-up of the issues discussed during the day. The point was made that the success of these large scale facilities has come to be seen as critical to biological research.

Suggestions for encouraging good data sharing practices were made. For example, it might be a good idea to have a journal issue dedicated to biological results from shared data.

It was stressed that better tools were needed to ease the burden of deposit, and that it was also necessary to track users so that they give credit to contributors (e.g. to require that the user provides a data citation if they then go on to publish).

Others pointed out that it will be difficult to handle the volume of information, and that the audit trail will be very complex. It is necessary to sugar coat the pill of data sharing, but there is still a burden of work that is incompressible.

Jef Grainger said the BBSRC's data sharing policy has been in place for 1 year, and the BBSRC will continue to monitor how the policy is working, and any obvious outcomes. So far there have been few applications for funds for data sharing, but this is expected to change as more people become attuned to the ideas and opportunities. The MRC is further down the road in this respect but they have not been inundated with requests for funding.

Appendix 1: Workshop call



Data Sharing in the Biosciences: a sociological perspective

Date: 26th June 2008

Location: National e-Science Centre, Edinburgh

This workshop aims to explore the changes in the generation, utilisation and governance of information in the biosciences; to consider the implications of these changes; and, to provide advice as appropriate. The workshop will bring together scientists who are having to deal with data sharing issues in their own research (in biology and other relevant disciplines), and social scientists who are studying the impacts of data sharing on scientific practices.

The last decade has seen the generation of increasing quantities of biological data, driven in part by large-scale research efforts such as the human genome project and assisted by advances in automated analysis. Researchers are increasingly likely to be utilising datasets produced elsewhere. These developments mean that life scientists must develop new rules and governance procedures regarding the release and sharing of information, and to do with the standardisation of data, models and experimental protocols. Additionally, public funders of research have introduced data sharing policies that scientists are obliged to follow. These changes have consequences for research practices and for the knowledge that is produced.

This workshop will explore these issues by addressing questions such as:

- What have been the major developments in the role of information technology and data sharing in the biosciences over the last decade?
- What examples are there of emerging experience in working with biological data, and what changes has this brought about?
- What implications do these changes have for scientific practice, community behaviour and associated infrastructures? (For example, will we see the dominance of 'dry' over 'wet' biology?)
- Can experiences in other research sectors (e.g. cosmology, particle physics and climatology) throw light on potential upcoming challenges for the biosciences? Do they have useful experiences or tools that could be translated into bioscience?

This workshop is sponsored by the Biotechnology and Biological Sciences Research Council, in collaboration with ESRC Centre for Social and Economic Research on Innovation in Genomics (Innogen) and will be hosted in Edinburgh by the e-Science Institute.

Places at the workshop are limited. Please send expressions of interest to Robin Williams, Jane Calvert or Jef Grainger using this email address:

research-data-workshop@lists.ed.ac.uk

Web page: <http://www.genomicsnetwork.ac.uk/innogen/events/workshops/title.3589,en.html>

Appendix 2: Workshop programme



Data Sharing in the Biosciences: a Sociological Perspective

26th June 2008

Cramond Room

National e-Science Centre
15 South College Street, Edinburgh, EH8 9AA

9:30 – 10:00am	Registration and coffee
10:00 – 11:10am	What are the main issues and what do we want to get out of the day? <i>Chair: Jef Grainger</i> Introduction: Jane Calvert, Robin Williams, Jef Grainger Round table (all participants talking for 5 minutes each about their interests in the topic)
11:10 – 11:30am	Coffee
11:30 – 1:00pm	Changes in data sharing practices <i>Chair: Robin Williams</i> Jef Grainger (BBSRC) 'Introduction to the BBSRC Data Sharing Policy – a funder's perspective' Peter Ghazal (Edinburgh) 'Lessons learnt from the RNAi Global Consortium' Discussion
1:00 – 2:00pm	Lunch
2:00 – 3:00pm	Today's challenges <i>Chair: Jane Calvert</i> Chris Taylor (EBI) 'Standards development: Necessarily a Two-Way Street' Sabina Leonelli (LSE) 'Data Sharing in Model Organism Biology'
3:00 – 3:40pm	Breakout session
3:40 – 4:00pm	Tea
4:00 – 5:00pm	General discussion <i>Chair: Robin Williams</i>
5:00pm	End

Appendix 3: Link to BBSRC's data sharing policy

http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.html