

5th International Digital Curation Conference

December 2009

Multi-Scale Data Sharing in the Life Sciences – Some Lessons for Policy Makers

<Graham> <Pryor>,
<University of Edinburgh>
<September><2009>

Abstract

Drawing on the final report of a recent series of case studies in the life sciences at the University of Edinburgh, this paper explores the attitudes and perceptions of researchers towards data sharing and contrasts these with the policies of the major research funders. Notwithstanding economic, technical and cultural inhibitors, the general ethos in the life sciences is one of support to the principle of data sharing. However, this position is subject to a complex range of qualifications, not least the crucial need for sharing through collaboration. The kind of generic vision for data sharing that is currently promoted by national agencies is judged to be neither productive nor effective. Only close engagement with research practitioners, in the identification of bottom-up strategies that preserve the exercise of informed choice - a fundamental and persistent element of scientific research - will produce change on a national scale.

A Duty to Share

Most of the major UK research funding agencies have published data sharing policies that explain the responsibility of grant holders to manage their data. This ‘duty to manage’ refers not only to the collection and safeguarding of data produced from publicly funded research but, in the wider sense implied by the practice of effective curation, extends to the provision of measures to enable access. The detail of what is meant by access differs between the domains, but broadly there is an intention that opportunities for the unrestricted re-use of data should be provided by those delivering the data, or their agents. The majority of grant applicants are therefore required to submit a statement on access, management and the long-term curation of their research outputs at the proposal stage.

All the major funders have signed up to the RCUK statement on open access to research outputs,¹ which declares that the *outputs from current and future research must be preserved and remain accessible for future generations*, but their focus currently varies when it comes to data. Whereas the AHRC, ESRC and NERC each require a statement on how resources will be created, on the assumption that this will facilitate their long-term preservation, the BBSRC, MRC and Wellcome Trust focus heavily on the data sharing potential of research resources, with the expectation that data will be made available with as few restrictions as possible. Unfortunately, none of them provides explicit guidance in this matter, although generally they acknowledge that different approaches to data sharing will be required in different situations, making it *appropriate for researchers to determine their own strategies for data sharing*;² and none of the research funders’ data policies advocates the concept of wholly *open* data.

It was in the context of this ostensibly coalescing policy environment that the RIN-funded³ Case Studies in the Life Sciences⁴ sought to provide a broad evidence base about information practices across the life sciences research domain. A partnership between the Digital Curation Centre (DCC) and the Institute for the Study of Science, Technology and Innovation (ISTI), the study team examined a wide array of cases, across differing areas of life sciences research, with the aim of providing a concise analysis of how researchers in the life sciences make use of information sources and services relevant to their research; how they analyse, evaluate and manage the information they acquire from such sources; what measures they take to create, gather, manage and communicate new data and information; and the mechanisms they apply when presenting and disseminating this new information. These studies involved drawing a systematic exposition of distinctions, commonalities and contrasts between the practices and needs of researchers in different individual fields of research, as well as an indication of changes that researchers themselves anticipate in information management practices and requirements in their fields.

The case studies were not, therefore, a specific investigation into data sharing practices and requirements, but in each of the seven cases the topic of data sharing emerged as a range of complex issues including genuine concerns about the risks from

¹ Research Councils UK’ updated position statement on access to research outputs, <http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf>

² Recognised, for example, in the BBSRC Data Sharing Policy, http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf

³ Research Information Network, <http://www.rin.ac.uk/>

⁴ Case Studies in the Life Sciences: Understanding Researchers’ Information Needs and Uses, October 2008 - June 2009, <http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/>



data misuse, sensitivity about ethical constraints and the need to safeguard valuable intellectual capital. Since the three funding agencies that have the more developed ethos of data sharing are those with the most direct bearing on funding for research in the life sciences, it quickly became evident that key messages were being recorded, and that these should be used to inform official data sharing policies and their strategies for implementation.

A Case of Diversity

The seven life sciences groups studied were all Edinburgh based and represented a diverse range of laboratories and research, ranging across studies of humans, animals and plants. They were also illustrative of different kinds of research context, encompassing analytical laboratory-based research, field research and in-silico research. The seven groups comprised:

- Animal Genetics and Animal Disease Genetics
- Transgenesis in the Chick and Development of the Chick Embryo
- Epidemiology of Zoonotic Diseases⁵
- Neuroscience
- Systems Biology
- Regenerative Medicine
- Botanical Curation.

There was diversity too in the nature of the data used by these groups, which included quantitative data, image data, field data (including national botanic collections), clinical data and laboratory-derived data. In some of the cases research is being conducted almost entirely within the digital realm - as demonstrated by the mathematical modelling programme of the Zoonotic Diseases team, or the MRI image scan processing undertaken by the Neuroscience group – where, as observed by the study team, tools and instrumentation do not merely enable the research but *are* the research.

Even where the research process is not exclusively computer-based, the research process was often found to involve the use of data produced from a range of sources, with data generated in the laboratory being complemented by imported data. The Animal Genetics group, for example, works primarily with the analysis of quantitative data created by industrial or research partners; the focus of the Zoonotic Diseases group is upon obtaining good quality data from a number of different sources, including field data collected by the team or its collaborators, spatial data (with GPS tracks, GIS map layers, satellite imagery and data from government agencies) and numeric disease data. Each group is, by the nature of the research being undertaken, operating in a culture built upon an underlying ethos of data exchange. For the Systems Biology group this principle is fundamental, for their goal is to take existing knowledge (often in the form of large datasets and static or kinetic models) and use it to generate new knowledge. For the Botanical Curation team, based at the Edinburgh herbarium, research on plant specimens supplied from around the world leads to the exchange and supply of data for a reference collection that is used on a global basis.

Vigorous and Participative Data Use

The impression gained, therefore, was of seven groups in which data use and generation is recognisably dynamic and participative, with most groups exhibiting

⁵ Zoonotic diseases are those which are transmitted from animals to humans



Figure 1 – Information Flow Map for the Zoonotic Diseases Group

The research data used and created by all of the life science groups we studied were categorised according to the range of digital research data detailed in the RIN publication *Stewardship of Digital Research Data: a Framework of Principles and Guidelines*⁸. Examples of the following five categories of data types were found to be distributed across the seven case studies: scientific experiments, models or simulations, observations of specific phenomena at a specific time or location, derived data from processing or combining ‘raw’ or other data, and canonical or reference data. As we note in the final report⁹, it should be possible to categorize any kind of research data into one of the above types from a *curatorial* perspective, although the categories themselves need not be mutually exclusive. Importantly, our studies confirmed that data can represent both input and output to the research process; they can also be re-purposed, and they may be positioned at more than one point on the research data lifecycle, dependent upon who uses them, as well as how and why they are used. The dynamic nature of research data use as described above may therefore also be expressed as the intrinsic quality of interdependency, which tends to position the process of research in a broader community than is defined by the individual research team.

When considering their attitudes to the data sharing imperative described in funding agency policy documents, we should of course turn our attention more closely to the *outputs* from our study groups, where knowledge transfer and communication cover not only the traditional publication of scientific papers and the regular dissemination of information about their work through presentations, but also raw data from experiments. These data are subsequently and selectively developed into formatted data, re-emerging as data processed for analysis, before leading to the extraction of summary results for discussion. For our seven groups the series of data constituents represented here can include derived products such as graphs, figures, tables, quantitative data, qualitative data and geographic maps. Outputs are created for and from analysis, and can include a broad range of formats and types such as statistics, images and imaging data, photographs, gene sequences, protein sequences and models. Some of the groups also create software and program modifications that include scripts, code and algorithms. When one becomes aware of the scale and range of these outputs, the simple objective to share (or the decision not to share) quickly permutates into a more challenging cluster of questions: what to share, with whom and why. Furthermore, the selection and dissemination process has immediately become considerably more complex.

Our strategy for identifying suitable case studies had deliberately excluded groups such as bioinformaticians, whose main function is information sharing and collaboration. This meant that when considering data sharing attitudes and behaviours our focus could be directed to life science researchers with no specialist inclination or role in information or data management, from which perspective their views and practices would be expected to reflect only the needs of the research and the researcher. Here we found that, in principle, life science researchers are very much in favour of sharing data; indeed, they proved to be remarkably willing freely to share quite valuable information and experience where this would facilitate each other’s research, including the sharing (on request and depending on the circumstances) of

⁸ <http://www.rin.ac.uk/data-principles>

⁹ *Op Cit*, Understanding Researchers’ Information Needs and Uses

standard operating procedures, plasmids, computer programmes, scripts and statistical analysis tools that they have written. In fact, methods and tools appear to be much more readily shared in the life sciences than experimental data, where we found a number of barriers repeated across the case studies.

Data Sharing – a Clutch of Caveats

The first inhibitor to data sharing is associated with cost, and this is made manifest in a number of ways. In the life sciences, typically, collecting data can take a number of years, especially when to obtain the required data depends upon building relationships of trust amongst an extended cohort of different stakeholders. Data may also be difficult and expensive to collect and exchange, particularly where contributors are working internationally in the field, as is the case with the Zoonotic Diseases team, with their research collaborators distributed throughout Europe and Africa; or it may depend upon expensive technological processes for the production, processing and storage of data, as in the case of the image analyses conducted by the Neuroscience group. In the case of the Animal Genetics group, the products of their research are consequent upon years spent building relationships of trust with commercial collaborators; for the Neuroscience and Regenerative Medicine groups, similar amounts of time had been dedicated to negotiations with clinicians and the establishment of a concord with patients.

So there is an obvious economic investment here of both time and money, which will influence decisions over sharing. More notably, perhaps, data produced by a research team also represents intellectual capital.

Researchers expect to be rewarded for the successful conduct of research, not for collecting and distributing the data they produce, although the measurement of their success will more often than not depend on the quality of the data produced. They are therefore naturally reluctant to share data that comprise the main means of adding value to their own research and, by corollary, their careers. So we found them extremely wary of giving away their data when this could lead to a competitor being handed an opportunity to apply further analyses and obtain kudos without having made the initial intellectual investment. From our studies we could identify high importance being attributed to this perceived scale of data value and its influence upon decisions about sharing; conversely, when data do not constitute added value (for example, geographic information or gene marker information), they are readily shared.

There is, then, a career obligation at the heart of this resistance to the open sharing of data, which has not been fully respected in prevailing data sharing policies. Individuals will seek to retain control over their knowledge and information, as this is important for formal and informal recognition as well as – in terms of both money and professional standing – reward. This condition is openly visible as a fundamental aspect of professional, discipline and institutional structures, affecting the incentives and rewards experienced by individuals and groups. Certainly, the issue of recognition and reward is not a matter of contention for the funding agencies. The BBSRC, for example, in its Data Sharing Policy supports the view that *those enabling sharing should receive full and appropriate recognition by funders, their academic institutions and new users for promoting secondary research*¹⁰.

Yet, as our study groups made plain, researchers are not funded to collect and organize data for sharing; they are funded to undertake research. This may of course include working with novel data, where the intellectual property of a project resides in

¹⁰ *Op Cit*, http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf

the raw data, and the researcher's research activity will add value to that data. But as argued by the Zoonotic Diseases group, research value can be lost as soon as the data is shared; further, researchers in the Regenerative Medicine team pointed out that whilst working on the development of a product with commercial potential they were required to keep all data private.

To be fair, research funders that require the sharing of scholarly output have to some extent addressed these sensitivities by allowing the deferral of release. Again, taking an example from the BBSRC's policy document, we find that it

*recognises that different fields of study will require different approaches. What is sensible in one scientific or technological area may not work in others; therefore the policy aims to achieve the sharing of data in an appropriate manner and not to be overly prescriptive.*¹¹

The BBSRC also accepts that

*researchers have a legitimate interest in benefiting from their own time and effort in producing the data but not in prolonged exclusive use of these data. Timescales for data sharing will be influenced by the nature of the data but it is expected that timely release would generally be no later than the release through publication of the main findings and should be in-line with established best practice in the field (where best practices do not exist release within three years of generation of the dataset is suggested as a guide).*¹²

The MRC position is similar if less distinct, communicating an expectation that data should be made available in a timely and responsible manner, and securely maintained for a minimum of ten years after completion of the research (a stipulation echoed by the Wellcome Trust). But the message on which they seem to agree is that it is reasonable only to allow a limited and defined period of exclusive use of data for primary research.

That the risks from premature sharing might be overcome by the careful timing of data release was in principle an acceptable option for the members of our case study groups, and their general preference seemed to reflect the BBSRC's proposition to delay sharing until after researchers have had sufficient time to complete their analysis of the data and to obtain the first publication; but in practice they proved reluctant to define how long the period of embargo should be. They are fully aware that methods of scientific analysis will improve, and with a sense of realism they feel the need to hedge against future opportunities to revisit their data and improve or extend their research. By referring to the example of sequencing, which has over time developed and now established its standard mechanisms, this caution is actually underwritten in the BBSRC data policy document, which suggests that *data sharing practices will change as areas of research develop and become more mature.*

Of course, practices and exigencies were found to vary quite considerably between the seven groups. Researchers in the Systems Biology group, whilst echoing some of the concerns held by the other groups, were found to have fewer reservations about sharing their data; indeed the whole rationale for systems biology is based upon the sharing of information and data. Because they are working in a very innovative field, with a focus on the development of new techniques, their sharing of data is very much more an issue of timing. For the systems biologists it is therefore not a question of should they share, but when will they share. Initially, their sharing will be achieved through publication. Similarly, the very ethos of the Botanical Curation group is all

¹¹ *Ibid*

¹² *Ibid*

about sharing information with other herbaria around the world, with taxonomists, scientists, amateur botanists and the public, where there will be *open access and no payment*. They provide information on request and loan specimens to taxonomists, and currently the group is collaborating in an international project with several hundred other herbaria, sending ‘type’ data to a foundation in New York. But these are exceptions that prove the rule: data sharing is complex and context-driven.

Ownership and Explication

Particular issues around sharing raw experimental data seemed to be well understood, not least the essential requirement for assigning good quality and highly specific metadata. This descriptive information will be crucial if the experimental data is to prove at all useful to third parties, and will include not only an explanation of provenance (where the data has come from and how it has been produced – a highly significant issue for the Zoonotic Diseases and Systems Biology groups, whose use of contributed data was pivotal), but also a range of interpretative notes and labels.

This in turn raises questions about how much time should be spent annotating data. Researchers in the Systems Biology group illustrated the problem by referring to their current and previous experience with micro arrays, which demonstrated how, in a developing field, it may take a very long time for standard methods to be developed and implemented before they can be said to work effectively. In that context, the provision and maintenance of meaningful metadata can prove an onerous task.

In all of our case studies we found that researchers in the life sciences express a keen sense of ‘ownership’ towards their data, which frequently (and perhaps unfortunately) emerged as an attitude resonant of protectiveness. It should not be taken as negativity. They feel responsible for the data they have generated and are genuinely concerned for the consequences of someone outside their immediate research orbit applying any inappropriate re-analysis. Rather than making their data freely available, they want first to know who is going to use the data and for what purpose. Even then, when any data are shared there is a perceived loss of control about how the data are subsequently used in a potentially extending chain of re-use, and measures to ameliorate that loss are important. One suggestion was that data being offered for re-use should be subject to a formal application rather than being made freely available, and that the researchers who originally collected or produced the data should have a role in determining whether the data are being released for only a specific instance of re-use. In many cases, the use of research data produced or collected and processed by other researchers was not favoured by our groups, on the basis that credible re-use would prove problematic given the numerous differences in experimental design and data collection, not to mention the lack of standardisation in data formats.

This view was shared by members of all of the groups except Botanical Curation, who in their conduct of ‘citizen science’ belong to a highly structured yet open domain. For the other groups, only collaborative arrangements where differences can be clarified and understood through direct contact were seen as realistic and preferable to making data freely available for re-use. Here we found a general respect for the view that intricacies of experimental design and data are not necessarily easy to understand, and direct contact is highly desirable. The Systems Biology group, whilst being linked with six other systems biology centres in the United Kingdom, acknowledge that they are each focused on different biological systems and the ways of modelling them, which determines that whilst they may share a philosophy there are practical limitations to the sharing of methods and data. Even after screening,

checking and corroboration through multiple software programmes, instances of data misinterpretation were cited, including that from the Animal Genetics group where poorly annotated genetic markers meant that data they had acquired from outside the group could have multiple meanings. Neither was this leveled as a criticism at the data source since, as already indicated, it is generally accepted that annotation and indexing are time consuming activities, which by implication reduce the time available for the principal objective of carrying out research.

Controlling the Trust-Share Balance

Notwithstanding these caveats, we established that a willingness to share remains a crucial element in the ethos of research in the life sciences, although individual researchers feel they must and will exercise choice about what to share with whom and when. Personal relationships are key here. In terms of sharing data externally, our studies found that the nature of the relationships that have been developed have a strong influence not only on whom a scientist might be willing to share data with, but also the manner of sharing, which might be realised through research collaborations, joint funding bids, or other practical scientific justifications.

Some kinds of data are shared on a highly restricted basis, with privileged access being the rule; others may be more freely exchanged with peers; and there is ambiguity in researchers' preparedness to share standard operating procedures or programmes, where sometimes to protect individual or team 'know-how' they are not shared at all, or are shared only within the research group; and where a novel technique has been developed, post-publication sharing is understandably the more likely preference.

In the biomedical life sciences, there are always particular sensitivities and issues of confidentiality that apply to the sharing of specific types of data (for example, the Neuroscience group spoke about their work with brain images in the context of health-care data, whilst others cited the sensitivity of data derived from animal experiments). But apart from these unarguable socio-ethical considerations, the potential impact from uncontrolled data sharing on business ethics, investment and profit is not inconsiderable. Where commercial organisations are collaborating in the research or where there is a potential for patenting (an imperative from university authorities as well as commercial partners!), then the protection of data from premature disclosure becomes extremely important, and will impact on any altruistic leanings towards the concept of 'open science'. In that context, the Regenerative Medicine case study provided the illustration of current work to create a particular therapeutic product, which it is predicted can be developed commercially. At this time the group does not want to share data from the project, since a high potential to patent means that no project work is shared outside the group. This is standard procedure, and the researchers in that team explained that other groups working in the same field would also not share data or information about their programmes except by publication, and patenting has to be completed prior to publication.

Lack of trust in the wider cyberspace environment was also found to be a large and pervasive issue across the seven cases. Some researchers declared their concerns about the perceived (if not always substantiated) risks from posting data on the Web, or from making data available 'in the Cloud', a nebulous landscape well outside their own more closely-defined arena of operations, where unknown others might secure access to their data and work on them in an equally unknown fashion.

The internal sharing of data with colleagues was described as a much more comfortable experience and one that is rich with productive activity. Our studies recorded examples including *informal discussion and email, formal presentations and*

meetings, the formal recording of data in lab books and the deposit of documents and data in shared folders.¹³ In addition, the sharing of experimental processes and methodologies are part of an ongoing, almost continuous internal discussion, where processes are carefully logged and recorded but rarely shared beyond the group.¹⁴

It has to be conceded that external sharing with peers is no less vigorous and can also include informal discussion and email, formal presentations and meetings, and the production of formal documents such as reports and published papers, although the dissemination of new techniques may be delayed until work is published. But there is a further more marked difference between sharing data internally and externally that is not explained as a distinction between formal and informal sharing (indeed, much of the sharing with external collaborators was reported as being done quite informally). Rather, we deduced from our studies that what is at issue is *the qualitative or experiential nature of the data being shared*.¹⁵ This is territory that generic data sharing policies have tended not to explore.

Conclusions

Achieving a balance between open access to research data and the need to protect intellectual work for future use is an issue that is expected to engage research groups for some time to come. The life sciences researchers who participated in this study held a range of views on their information needs and some mixed feelings about current and future challenges; yet from discussions in both individual interviews and focus groups it was clear that most of our subject researchers have reservations about open data sharing, whilst remaining in favour of the principle of sharing data. As a rule, their distinct preference is for data sharing to be executed through collaboration and in communication with other research groups or individuals, with regard both to the sharing of their own data and in respect of gaining access to the data of other researchers.

Overall, and in the face of pressure from the funding agencies, two strong provisos emerged that should govern the sharing of data: first, researchers must have an opportunity to publish the results of their initial research in a peer reviewed paper; and second, sufficient time must be given to allow the completion of their analysis of the data. Our study groups were, however, unable to prescribe the length of time a researcher or research group could or should be able to hold data for their own use before release to open access. One group offered a retrospective example of four years of intensive analysis on a particular set of data as sufficient for them to have extracted as much as they could from it, but they also believed it possible that a new method or tool could become available that would enable a fruitful return to that data at some undetermined point in the future. Members of the Animal Genetics and Zoonotic Diseases groups actually declared it impossible to predict how much time would be needed for analysis before data could be deemed *finished with* and available for open sharing.

The data sharing activities we observed were driven primarily by the needs and benefits perceived by research active life scientists. They included reciprocal and altruistic exchanges within peer communities where these did not cut across other incentives, such as the need to exploit intellectual property or publication opportunities. Data sharing also takes many forms, including significant levels of informal exchange, which may themselves include the exchange of scientifically

¹³ *Op Cit*, Understanding Researchers' Information Needs and Uses

¹⁴ *Ibid*

¹⁵ *Ibid*

crucial information and experience. These are forms of exchange that are often overlooked by policies promoting formal exchange.

In terms of the promotion of data sharing activity, including the broader underlying requirement for good data curation, the foremost conclusion drawn from our case studies was that if national strategies and policies for research data are to be effective, they must be *informed by an understanding of the exigencies and practices of individual research communities*.¹⁶ A single or generic vision will be neither productive nor effective. Our message, therefore, is that *practical and human issues governing the restriction of data exchange will persist in the life sciences and in other domains, and only further close engagement with research practitioners, to identify and qualify the caveats to data sharing, as well as to preserve the exercise of informed choice that is fundamental to science*,¹⁷ will produce change on a national scale.

Given the limited understanding of which forms of data sharing are most effective and beneficial, and under what circumstances, the view taken by our study team was that it would be helpful if funders could *adopt a more pragmatic and experimental policy that recognises the multiplicity of contexts*,¹⁸ often founded upon informal sharing around the recognised and mutual needs of research groups. If the life sciences can be taken as typical of other scientific domains, such a bottom up view is essential if we are to attend to the practicalities and circumstances of research data sharing, the better identification of its benefits, the labour that must be expended to achieve it, and the barriers and drivers for change. Given that the research culture is not, ultimately, driven by concealment and reticence, would it not be a reasonable beginning for researchers themselves to be encouraged to spell out which data sharing strategies they may wish to adopt?

Acknowledgements

This paper has been developed from the final report from *Case Studies in the Life Sciences: Understanding Researchers' Information Needs and Uses*, which was presented to the Research Information Network in September 2009. I am indebted to my co-authors of that report – Robin Williams, Director of the Institute for the Study of Science, Technology and Innovation, and researchers Ann Bruce, Wendy Marsden (ISSTI) and Stuart Macdonald (EDINA) – for the wealth of original material that enabled this paper to be written.

References

- [Policy document] BBSRC. (Undated). Data sharing policy. Retrieved June 2009 from http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf
- [Report] Humphrey, C. (2006). e-Science and the Life Cycle of Research, University of Alberta, Canada. Retrieved June 2009 from <http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc>
- [Report] Research Information Network. (2009). Case Studies in the Life Sciences: Understanding Researchers' Information Needs and Uses.

¹⁶ *Ibid*

¹⁷ *Ibid*

¹⁸ *Ibid*



- [Policy document] <http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/>
RCUK. (2006). Research Councils UK' updated position statement on access to research outputs. Retrieved June 2009 from <http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf>
- [Report] Research Information Network. (2008). Stewardship of Digital Research Data – a Framework of Principles and Guidelines. Retrieved May 2009 from <http://www.rin.ac.uk/data-principles>