RSE Young Academy of Scotland Working Group Report on 'Science as a Public Enterprise'

11 December 2011

This working group, consisting of members of the RSE Young Academy of Scotland, is pleased to contribute to the call for evidence from the Royal Society of London on the important area of sharing scientific data (including information). This paper attempts to discuss some of the benefits, ethics, motivators, concerns and practicalities relevant to the study.

The imperative for responsible data sharing

Sharing data has been critical to scientific progress, and the cost of not sharing is significant; as Sir John Sulston notes, "From sharing, discovery is accelerated in the community" [tinyurl.com/bukdrcv]. The prompt and freely available sharing of data advocated by Sir John (e.g. the human genome sequence), requires trust amongst nations and has already enabled huge scientific progress. While the clinical benefits of the human genome sequencing project are just beginning to be realised [*Nature* 2010;464:649], restrictions on the availability of information clearly restrict the advancement of knowledge. For example, it is unethical to withhold critical decision-making data, such as information relating to drug safety and efficacy as documented for Oseltamivir [*BMJ* 2009;339:b5351].

Sharing enables meta-analysis and data repurposing to explore hypotheses unrelated to the thinking that first generated the data. There are many examples where this could be compelling, such as protein/gene expression data [e.g. *PNAS* 2004;101:9309] and tracked position against time data in colloidal crystals/glasses. Indeed, repurposing both reduces duplication of effort and has a strong ethical motivation when compared with potentially unnecessary collection of equivalent data (especially where patients are subjected to lengthy procedures). Related to duplication of effort, interaction and sharing of information amongst funding bodies helps prevent parallel sponsoring of similar projects. Availability of data (and meta-data) is also crucial for the independent verification of published work. For example, the ArrayExpress gene expression database [www.ebi.ac.uk/arrayexpress] assists peer review (e.g. for scientific journals) by scoring the compliance of deposited datasets with community standards, including the availability of data provenance [*Nat. Biotech* 2006;24:1321]. Wider access to data may encourage cross-fertilisation between disciplines, increasing opportunity for application of methods from one discipline to generate or test hypotheses in another discipline. Also, timely access to data is important for up-to-date, evidence-based decisions (for example, in public policy). Furthermore, a global community participating in open scientific data sharing may contribute to greater economic and political cooperation.

However, it would be unethical to share information without careful consideration of the potential consequences. In some instances, scientific developments can have major effects on the power of nations. Political cooperation can be difficult, especially considering ongoing hostilities and economic disparity among nations. Accordingly, a significant risk and ethical concern in global data sharing is that knowledge brings power, which may be wielded destructively; for example, knowledge relevant to making nuclear warheads raises serious concerns [*Science* 2009;324:1499]. Indeed, game theory suggests that conditions of distrust and alienation equate to uncooperative, potentially violent behaviour [e.g. *Am. Econ. Rev.* 1993;83:1281]. With the above in mind, free disclosure of scientific information such as the smallpox virus genome [e.g. www.ncbi.nlm.nih.gov] brings inherent risks. Nevertheless, cooperation and mutuality confer significant benefits to both groups and individuals: "nice guys finish first" [Dawkins 1989, *The Selfish Gene*, pp202-233]. Therefore a current challenge is to nurture conditions that encourage cooperative, globally advantageous behaviour and so promote the welfare of all. In scientific data sharing, this challenge is faced in multiple overlapping contexts. These include international relations, various academic communities (e.g. global, national, within institutions) and commercial enterprise. One factor that may encourage cooperation in data sharing is an individual sense of participation in and ownership of knowledge advances. Interestingly, ownership of scientific data/information, realised through Intellectual Property (IP) law, impacts significantly on the affordability of the fruits of science; a recent article highlights economic, scientific and ethical drawbacks to the current IP system [*Prometheus* 2011;29:325].

Confidentiality, anonymity and informed consent are overriding considerations in collecting and sharing human data. Primary ethical responsibility rests with the scientists actually performing the research, as well as other key stakeholders (e.g. data repositories). Leading figures such as Sir Joseph Rotblat and Sir John Sulston have called for scientists to take an equivalent of the Hippocratic oath. Indeed, the UK government,

universities and funding councils have clear scientific ethical frameworks (e.g. Sir David King's seven principles [tinyurl.com/cph6884], UKRIO Code of practice [tinyurl.com/cjlz7jg]). Sharing human data has significant complications, including the ethical context of some studies whereby participants may withdraw their data at any time, which would be practically impossible where data have been released into the public domain. On the other hand, we expect many cases where informed consent may be reasonably obtained in order to enable data sharing. Inherently personally identifiable data, such as videotaped interviews, are of particular note and require special consideration. However, even in the case of non-human data, issues regarding data security and context have to be carefully considered, as demonstrated by the 2009 'Climategate' controversy [tinyurl.com/38wtk5v, tinyurl.com/338tb24]. It is clear that different data types have distinct constraints on sharing, both ethically and practically. For example, some areas of Life Sciences stipulate an ethical approval process as precondition for data release and so critically require data security. Accordingly, some data - even if publicly funded - should not be made freely available.

In summary, the open sharing of scientific data has many potential benefits, such as increasing the rate of scientific progress, ensuring more efficient use of time and financial resources, and ensuring that information critical to public policy decisions is available to all. Development of an environment that encourages sharing and reusing data is important for realisation of these advantages. Nevertheless, available data must be reliable. Flawed data can be worse than no data at all, especially if errors contaminate years of downstream work. Therefore, quality control is an essential part of data sharing and a responsibility of both producers and receivers of data. Furthermore, any benefits must be carefully weighed against potential costs, such as when the imperatives of confidentiality, anonymity, and security conflict with those of openness. As benefits and costs will vary among disciplines and even among projects within disciplines, procedures for formally weighing these considerations should be paramount to any plans for increasing data sharing.

The economics of openness

The economic implications of data sharing also warrant careful consideration; withholding data offers a route to maximise immediate economic returns to the country or organisation that has funded (or ‘owns’) the research. Of course, working for individual short-term economic gain runs against an ethical imperative to promote the welfare of all. Global economic progress is impeded where effort is duplicated, and where ideas are not communicated across research communities. On the other hand, it is important to avoid expending resources to accommodate the sharing of data that is never retrieved. Furthermore, commercialisation can energise the translation of research into practically useful outcomes that impact directly on quality of life. Therefore, a more transparent framework seems to be required in order to synthesise openness with commercial activity; indeed, current activities may be operating far from the optimal point of balance between these superficially opposing concerns. For example, researchers can patent their inventions in the US up to one year after publication of their findings, whereas patenting in the UK is required before publication [http://www.ipo.gov.uk/grace.pdf]. It is not necessarily clear whether the latter practice is to be preferred from the point of view of innovation. There seems to be a good case in support of legislation requiring companies to release more data into the public domain; clinical trials data are a particularly compelling example, such as the documented withholding of data on the drug Oseltamivir (Tamiflu) [*BMJ* 2009;339:b5351, *CDSR* 2011;CD008965], referred to above.

Public engagement and data sharing

Firstly, we list below some of the ways in which non-specialists may interact with science (including interactions between scientific researchers from different fields):
i) As participants, i.e. through being involved in the actual research. For example, (a) as subjects in medical tests, (b) as data collectors [bigbutterflycount.org/, fold.it/portal/], (c) by doing virtual experiments [ve.soton.ac.uk/, faceresearch.org] and (d) by analyzing data or providing resources for the analysis [setiathome.berkeley.edu/, climateprediction.net/].
ii) As recipients of knowledge generated through scientific research. For example, through facilities and/or projects such as the Science Museum [sciencemuseum.org.uk/], Lab in a Lorry [labinalorry.org/] and Cafe Scientifique [cafescientifique.org/], or as students.
iii) As consumers of innovative products based on scientific outputs. To give just a couple of many possible examples: (a) using e-readers employing electronic ink based on colloid science [‘Colloids and colloid assemblies’ by Frank Caruso] and (b) taking anti-influenza medicines such as Relenza.
iv) As adjudicators of scientific knowledge, e.g. to position oneself (politically) on issues such as: climate change, geo-engineering (Claire L. Parkinson’s “Coming Climate Crisis?: Consider the Past, Beware the Big Fix”), genetically modified crops, stem cells for medical treatments and/or cloned meat.
v) As funders of scientific research, e.g. via taxes.
vi) As agenda setters for science, through public participation mechanisms that seek to consult citizens on what kind of scientific work should be funded and how it should be governed [e.g. tinyurl.com/bls5uc5].

However, sharing data (including information) with non-specialists remains a significant challenge, largely due to the implicit knowledge gap, leading to possible misinterpretation of results—and sometimes misrepresentation of scientific endeavour [e.g. tinyurl.com/3sf495t]. A significant part of communicating research and data involves interactions between scientists and the wider public. However, mistranslations and misunderstandings are also an issue where investigators from different scientific fields come together to discuss the work and share data. This issue is further complicated by mainstream media (e.g. newspapers) seeking to sensationalise research findings, intentionally or indirectly. Therefore, particular care is required, and individual researchers must do their utmost to communicate sensitively with each other and with the wider public; particular benefit will be afforded through close collaboration with appropriate sources of support (e.g. university press offices). Greater support for researchers would be helpful in this regard such as systematic and inspirational training in communication and engagement (e.g. during the doctoral degree), which should lay the foundation for more widespread dissemination and appreciation of scientific data - as well as promoting genuine and reciprocal dialogue between scientists and the wider public.

As noted above, data interpretation and education are inextricably linked; for example, knowledge is required in order to make informed choices about one's own healthcare. Therefore, patients who wish to make informed decisions, but lack prerequisite knowledge about the relevant data, may feel frustrated with the healthcare system. On the other hand, education campaigns may be regarded as intrusive; the information contained within them may lack salience for people's everyday lives. Furthermore, some people are relatively well educated about health and science, and it can be precisely because of this that they are able to challenge the information they are presented with. While this only touches briefly on a complicated area, greater availability of systems to engage with and educate patients about specific medical readouts (e.g. data on haemoglobin levels) would appear helpful. More generally, it seems likely that people will be better able to interpret data if they understand more about how the data is generated. Hence, it would appear to be beneficial to increase awareness of how science is conducted, e.g. by incorporating more material on scientific methods and technologies within the science curriculum. Unfortunately, a significant challenge remains - in that many scientists perceive that they do not have sufficient time to spend on public engagement activities. Greater weighting of public engagement outputs in research funding decisions and performance reviews would help to mitigate this and so encourage data sharing.

The influence of new media on scientific research

The Internet means that communication is almost instant, with wide national and international coverage. For example, the British Psychological Society Media and Press Committee 'tweet' about conference presentations during or shortly after presentations, giving details of their content. Conference presentations often include research that is unpublished, unpatented, and minimally peer-reviewed, if at all. The increase in the use of blogs among scientists is another area where the lack of peer review can become an issue. One prominent example is Satoshi Kanazawa's controversial Psychology Today post on "Why are Black women physically less attractive than other women?", which presented unreviewed analyses that were later shown to be highly flawed [tinyurl.com/cejz6ml]. Although subsequent work can resolve incorrect findings, the initial reporting can often cause serious damage in the meantime, as exemplified by the MMR controversy. Indeed, cases of measles and mumps have risen significantly in recent years compared with 1998 levels [tinyurl.com/5uylxdc], potentially as a direct result of media-induced fears about a link between MMR vaccines and autism. The above issues highlight emerging pitfalls, and the importance of ensuring that the status of findings and analyses are made clearly apparent to the public.

In addition to increased access to direct communication from scientists, the Internet has brought about an explosion in the availability of news media reports on science. These media reports are often written by journalists without specialist training in scientific journalism, potentially resulting in sensationalism and inaccurate conclusions. Often, the authors' names and the journal are not even mentioned, just the university's name, making it difficult for even experienced researchers to access further information. While some of this could be remedied by linking media reports to the original scientific publications (e.g., by DOI), much of this information is barricaded behind paywalls. Therefore, academic journals should be encouraged to make abstracts and non-specialist summaries available for all of their papers, as well as information on obtaining the original data reported. Indeed, we would welcome wider discussions around the role that scientists and scientific institutions may inadvertently play in enabling the sensationalisation of science.

Effecting and policing data sharing

In addition to ethical responsibility, individuals carrying out research have a critical role in deciding how data may be shared in practice. However, a fundamental concern is that individuals and organisations may stand to make significant gains by withholding data. In addition, individual researchers, or their line managers, may feel that the time spent on facilitating data sharing would be better spent on career progression, e.g. by

writing papers. These issues, amongst others [e.g. PLoS One 2011;6:e26828], underline the importance of mechanisms to enforce data sharing. Public domain archiving of data is typically compulsory for research funded by the Economic and Social Research Council (ESRC), who support a database (UK-DA) specifically for this purpose [store.data-archive.ac.uk/store/]. We would encourage a model where research funders can both require and be involved in policing of data sharing - subject to appropriate exemption guidelines (e.g. on ethical grounds). For example, UK-DA archives data from approximately 70% of ESRC funded projects [ESRC *pers. communication*]. We would also welcome more consistency amongst funding bodies in this regard, while appreciating that different disciplines may have varying data-sharing needs—one size does not necessarily fit all. Crucially, research funders have significant leverage to ensure that data sharing is effected. For example, the relatively recent movement towards open access publication is massively energized by the policies of funding bodies both in the UK (e.g. BBSRC, Wellcome Trust, MRC) and abroad (e.g. NIH, CIH). Publishers and data resources (e.g. databases) also have a crucial role in enabling and scrutinising authors' sharing of research results. We would welcome further development of automated systems to facilitate assessment of compliance with data sharing requirements (e.g. for peer reviewers), including appraisal of provenance information, which is fundamental for effective reuse of data.

Models where individuals grant access to data prior to publication can raise concerns due to the current bibliometrics-driven assessment of research output. Making data directly citable could help this to some extent; for example companies such as Reuters could track the use of data (e.g. accession numbers, images) in research publications. Data transfer agreements may also be appropriate—especially where there are ethical concerns about how data is used. However, in some areas data are made available as soon as practicable, with a time-limited public disclosure embargo on results based on the data. For example, the ENCyclopedia Of DNA Elements consortium [*Science* 2004;306:636] operates an embargo for 9 months from the date of data release, which precludes any disclosure of results at seminars or to electronic servers such as journal submission systems. Broadly, 'large-scale' projects involving expensive equipment and/or many collaborators seem to inherently promote a culture of data sharing; communities of this type include astrophysics and functional genomics. Funding bodies are evaluated at least in part through publication of research that they have funded, as well as citation of those publications. Therefore a possible issue with policing of data sharing by funders is potential conflict between funding bodies' own reporting objectives and any policing framework.

Organisations and systems to manage appropriate access to data are an essential practical aspect of data sharing. Examples include the UK National Cancer Research Institute 'Oncology Information Exchange' [http://www.ncri-onix.org.uk] and US National Cancer Institute 'cancer Biomedical Informatics Grid' [https://cabig.nci.nih.gov/], the large-scale digitization of books by Google [e.g. books.google.com/ngrams] and the recent soft-matter data-repository trial 'eComploids' [ecomploids.org]. The cost of maintaining data archives and access is an important consideration that requires forward planning, particularly by funders and individual researchers. For example data from 'deep sequencing' technologies were at one time centrally available in the US NCBI Sequence Read Archive (SRA), which was closed in 2010 due to funding issues [tinyurl.com/bwc95rg] (currently operating a limited service [tinyurl.com/ccghftb]). Moreover, facilities require clear protocols to ensure long-term storage and accessibility (e.g. via robust file formats) of large amounts of data, for example 3D microscopy time-series and/or X-ray diffraction data. If source data are not retained, then the value of the associated research activity is greatly diminished. Moreover, availability of raw data is a critical component in independent evaluation of research findings. Where costs of data archiving are limiting, alternative strategies (e.g. off-line access, minimising redundancy, access-on-demand) should be considered.

Closing remarks

Research funders, both public and private, as well as publishers, have absolutely key positions in setting the norms in science culture, including attitudes and behaviours around data sharing. Importantly, funders and publishers are in a unique position to enforce data sharing as a requirement of accepting a funding bid or an article for publication, as appropriate. However both researchers and publishers may seem conflicted in this area due to specific pressures such as attracting readers or fulfilling reporting objectives. In addition, data sharing (including public engagement) should become an integral part of the training and assessment of researchers, so that data sharing becomes an integral part of future research efforts. For example, the Research Excellence Framework (REF) could give more emphasis to rewarding data sharing. It may be premature to draw strong conclusions, however some of the emerging themes include:

- Sharing scientific data has significant advantages, to summarise a few:
  - Accelerating scientific progress.
  - Fostering public debate in order to guide research priorities.
  - Enabling unforeseen connections between different areas of research.

    - Supporting economic growth.
    - Promoting good relations between parties involved in sharing, such as economic cooperation between nations.
- There are instances where commercial companies should be compelled to make more data available to the scientific community.
- Security issues should be carefully considered to minimise the use of data in a damaging way due to a lack of understanding or malicious intent.
- Ethical concerns must be addressed robustly. Principles of confidentiality and anonymity sit alongside the issue of ownership and the mechanisms by which benefits are realised.
- Comprehensive procedures to formally assess costs and benefits of data sharing appear to be an important area for future development.
- Data protection strategies are crucial for managing ethically approved access to databases.
- Given finite resources (both financial and technical) and the sheer volume of data to be stored, there are practical difficulties with sharing all data.
- Policies should consider lessons learnt from successes and failures in different disciplines, as well as what the leading edge of information technology can offer as tools to facilitate large scale data sharing.
- There are concerns about greater availability of data to the public via the media, including the internet. Serious efforts will need to be made to avoid potentially harmful consequences, such as the premature release of information that could affect healthcare decisions.
- Sharing data in the widest sense remains a significant challenge, partly due to the implicit knowledge gap between specialists and non-specialists.
- Widespread and systematic training of scientists in communication and engagement seems important as a foundation to encourage greater appreciation of scientific data, as well as promoting dialogue between scientists and the wider public.
- Developing a culture of greater data sharing may require established methods of recognition to be combined with tangible rewards for sharing, and possibly penalties for not sharing.
- Perceived and legally defined ownership of data are important factors influencing attitudes to data sharing.
- Academic journals should be encouraged to make abstracts and non-specialist summaries available for all of their papers, as well as information on obtaining the original data reported.
- We would welcome further development of automated systems to facilitate assessment of compliance with community-defined data sharing requirements.
- In some areas, a ‘critical mass’ of data sharing may have to be achieved before the benefits and efficiencies outweigh the drawbacks and overheads.

Please note: both the Royal Society of Edinburgh (RSE) and a working group of the RSE Young Academy of Scotland (RSE YAS) have responded to the Royal Society of London call for evidence on "Science as a public enterprise". These responses were independently produced.

Also, this paper does not necessarily reflect the views of all RSE YAS members or the RSE.

The RSE YAS members involved were: Ian Overton (chair), Lisa DeBruine, Sinead Rhodes, Job Thijssen, Alan Gow, Des Balmforth and Martyn Pickersgill.