



## **A RESPONSE FROM THE BRITISH ACADEMY TO 'MAKING OPEN DATA REAL: A PUBLIC CONSULTATION'.**

The British Academy welcomes the opportunity to respond to this consultation on 'Making Open Data Real'.

This response is a slightly revised and updated version of the British Academy's submission to the Royal Society's call for evidence on 'Science as a Public Enterprise: opening up scientific information'.<sup>1</sup> The Science as a Public Enterprise project is a major study on the use of scientific information. It aims to examine the principles, opportunities and problems of sharing and disclosing scientific information and identify measures that should be taken by scientists and their institutions, policymakers and others in creating a socially responsible open data regime. The British Academy's former president, Baroness O'Neill sits on the high-level working group for the project. We strongly recommend that the government takes note of the recommendations of this project, once they are available.

### Summary

1. The British Academy supports the release, disclosure and sharing of scientific data, though we recognise that there will have to be an explicit policy to restrict access in certain circumstances. The underlying reason for favouring this approach is the importance of public policy, and public debate about it, being based on the best available evidence. There is a concomitant need to improve public understanding of, and capacity to interpret, such evidence and for the need for data to be accessible.
2. In humanities and social science research, great value can be gained from access to large social science data sets, especially longitudinal studies. We believe that in principle, and across all subjects, anonymised data collected and held by government and public bodies should be made available to other researchers in order that they can assess, test and challenge research findings, or conduct additional research using these data. Where the research has been supported by public or charitable funding, it seems hard to argue convincingly for restricting access to that data. We also believe that privately funded research data should be made available if there are potential public policy implications.

### Access to data and publications

3. The social science community is concerned about obtaining access to government *ad hoc* anonymised social surveys in order to conduct secondary analysis, which is a major part of social research. We understand that, in the wake of some well-publicized losses of administrative and personalised data, some government departments may be denying or delaying access to those who wish to carry out secondary analysis on publicly funded anonymised data. Though the lost data tended to contain personal information (e.g. CD-ROMs and portable hard drives with pensions or welfare benefit details about identifiable individuals), the

---

<sup>1</sup> <http://royalsociety.org/policy/projects/science-public-enterprise/>

consequence of such losses has been that government departments have placed additional restrictions on the sharing of anonymised survey data. Some of these may be a result of over-caution among civil servants or due to the fact that some departmental lawyers are not aware of the appropriate standard for handling anonymised data (which are different from the standards relating to personally identifiable data). Some restrictions are clearly necessary to ensure anonymity and confidentiality, but it is important to find a way to address issues of data security while at the same time enabling secondary analysis – much of which could provide new insight into public policy dilemmas.

4. It appears that there is an inconsistent approach to this across government. Some departments have clear protocols in place; others, it seems, do not. The Government Social Research Unit (GSRU) and the Heads of Analysis have been promoting good practice but there is a lack of clear, consistent guidelines to cover all government departments. In its report published on 28 July, the Science and Technology Committee quoted Professor Adrian Smith, BIS Director General, Knowledge and Innovation:

*There is a great movement now and a recognition of openness and transparency, which has always been implicit as a fundamental element of the scientific process. But the more we collect large datasets, you have to give other people, as part of the challenge process, the ability to revisit that data and see what they make of it with openness and transparency. There is general support these days for the presumption that the research, the associated data and if you have written a computer code to assess it, should all be available and up for challenge and testing validation. In fact, explicitly the Research Councils encourage that, as Government Departments do. However, there can be complex and legitimate reasons for not necessarily, at least in the short term, being that transparent. An awful lot of policy in recent years has meant that we have been trying to lever more out of public investment by joint working with business and industry and leveraging additional funding. Once you get into that territory, you do have commercial and intellectual property constraints on a temporary basis at least, for openness and transparency. The presumption is that, unless there is a strong reason otherwise, everything should be out there and available.<sup>2</sup>*

5. The Committee's report makes a welcome reference to the Royal Society's Call for Evidence for their project on Science as a Public Enterprise. We believe that this is an important opportunity to examine fully the issues involved in sharing data but, more fundamentally, to influence new ways of working. As the Committee concluded:

*Access to data is fundamental if researchers are to reproduce, verify and build on results that are reported in the literature. We welcome the Government's recognition of the importance of openness and transparency. The presumption must be that, unless there is a strong reason otherwise, data should be fully disclosed and made publicly available. In line with this principle, where possible, data associated with all publicly funded research should be made widely and freely available. Funders of research must coordinate with publishers to ensure that researchers disclose their data in a timely manner. The work of researchers who expend time and effort adding value to their data, to make it usable by others, should be acknowledged*

---

<sup>2</sup> Science and Technology Committee – Eighth Report: Peer Review in Scientific Publications, paragraph 197

*as a valuable part of their role. Research funders and publishers should explore how researchers could be encouraged to add this value.*<sup>3</sup>

6. It is hard to see how government will make progress with releasing anonymised administrative data for wider public analysis and debate if it cannot solve the problems of releasing anonymised survey data, via recognised national archives, to researchers who demonstrate that their institutions have appropriate data-handling safeguards and where the researchers have given appropriate undertakings in relation to anonymity. The British Academy understands that these three elements (appropriate and proportionate anonymisation of survey data, data-handling safeguards and researcher undertakings) should be sufficient, but the current practice in some departments is for lengthy vetting of analysis plans and so on, which in effect delay or deny access. In the past, when government departments funded data collection, there were contractual undertakings that the data would be anonymised (by joint agreement of the data collector and the department) and archived at an appropriate data archive within a certain period of the study's conclusion. We would hope that government departments might revert to this process, which would then not only make data more widely available, but would free departmental time to work on the harder issues of making administrative data more widely available as appropriate.
7. We are disappointed that the consultation makes no mention of pseudonymised data, which is essential to much research in public policy and the social sciences, and should be part of any consideration of 'open data' policies. Pseudonymised data is data that has had all personal identifiers (information such as name and address) removed, but has been allocated a code number which enables the data controller to link it back to the individual via a 'key' that decodes it. This kind of data is essential for any research which seeks to understand the relationship between one thing and another (e.g. health status and employment, smoking and cancer, free school meals and educational attainment). Although it is impossible for anyone other than the data controller to identify individuals from the data and so is, to all intents and purposes, anonymous, it is still classified as 'identifiable' data. As such, researchers have either to seek consent from the individuals concerned or apply for special exemptions from the Data Protection Act 1998 (DPA).
8. A lack of clarity about the provisions in the DPA for the use of data without explicit consent has meant that researchers have been reluctant to make use of these special provisions and instead adopt a 'consent or anonymise' approach to the legal framework in the UK. Full anonymisation of data renders the data useless for much research, including any that relies on techniques of data linkage. However, seeking explicit consent can impose real burdens on researchers in terms of the financial costs, cause delays in starting research and it can jeopardise the success of research due to problems with consent bias and incomplete samples. Our Vice President of Public Policy, Professor Albert Weale, published a research report that looked at this issue in more detail, earlier this year.<sup>4</sup> The report concluded that more consideration

---

<sup>3</sup> Ibid, paragraph 203

<sup>4</sup> Clark, S. and Weale, A. (2011) *Information Governance in Health* (London: Nuffield Trust)

<http://www.nuffieldtrust.org.uk/publications/access-person-level-data-health-care-understanding-information-governance>

should be given to the possible justifications offered by the public interest in research rather than automatically defaulting to obtaining consent if anonymisation is not feasible.

9. It is worth noting the issue of qualitative research, particularly the sort done by private opinion polling companies for government and related organisations in connection with public consultations. One concern lies in the company/organisation maintaining the data as a proprietary item. Another problem with such studies is that often the reports summarise consultation responses where there has been disagreement, but it is impossible to know how differences of opinion have been aggregated.
10. In some branches of the humanities and social sciences, there has for some time been the practice of lodging replication data sets for articles published in journals. An example is provided by the British Journal of Political Science.<sup>5</sup> These data sets often contain information about programmes and coding so that replication can (in principle) be carried out. There may be a public interest in discussing this more widely in the case of data analysis that is likely to be contentious. A separate issue is the requirement of all research councils to require the public archiving of data collected with public funding. For instance, the ESRC requires its grant holders to offer any data that result from an award to the UK Data Archive, most of which is then free for others to use. This includes data sets like the British Election Study,<sup>6</sup> the British Social Attitudes surveys, and the ESRC funded birth cohorts. A different tradition has developed in some areas, wherein the data are seen as the investigator's intellectual property. While the Academy appreciates that these are delicate issues, it thinks that after an appropriate period of time all anonymised data collected with public funds should be made available through appropriate archiving (and with appropriate data safeguards) for more widespread secondary analysis.
11. We welcome the general move to 'open data' whenever feasible to improve access to and awareness of the results of research. There are many reasons why it is more desirable, and also more practical, for access to data to be across the board, rather than attempting to limit it to particular organizations or professions. Not the least of these reasons is that there would appear at present to be a widespread public scepticism about science (in the broadest sense). Making data and results more publicly available and usable may help to counter such scepticism; and may also reduce the toxicity of claims that information is being kept secret or manipulated – whether by government, university departments or companies.

#### Areas of concern

12. It is not sufficiently clear what costs will be incurred and who will bear them. If the onus is on the original research data collectors, then there is potential for some organisations and institutions to take on the cost while others receive all the benefit. It is important that this is not seen as a demand for data to be given away to potential competitors with no benefit to those who have spent valuable time and money to

---

<sup>5</sup> The Journal's Instructions for Contributors states that "Where statistical analysis of data has been conducted, contributors are expected to deposit a REPLICATION DATA SET with a major international data archive or on the *Political Analysis* website" [Journal's emphasis] - <http://journals.cambridge.org/action/displayJournal?jid=JPS>

<sup>6</sup> <http://www.essex.ac.uk/bes/>

collect it. However, if fees are charged for access to data, this must not be done in such a way as to disincentivise further research. In practice, we suspect that with appropriate attention to how anonymised data might be deposited with an appropriate archive, with suitable time lags for anonymisation and for data description (which protects the interests of those designing data collection), more widespread secondary analysis would follow. Those funding the secondary analysis could then fund their share of the archiving overheads.

13. There will also be other costs. For example, who should pay for providing metadata and interpretive material? And, how is that to be reconciled with, for example, plans to obtain commercial advantage from data (patenting, licensing, further academic or commercial research)? Should those that collate and curate valuable data retain some form of prior claim over its use?
14. There are conflicts between openness and data protection requirements. Our former president, Baroness O'Neill – who is also a member of the Royal Society's 'Science as a Public Enterprise' working group<sup>7</sup> – spoke on this subject at a discussion event hosted by The Foundation for Science and Technology, held at The Royal Society on 8 June 2011. One particular concern is the requirement for specific re-consenting to enable impersonal secondary use of legitimately acquired and lawfully held data. It is not possible to provide a workable or coherent distinction between personal and non-personal information and the definition of 'processing' of information is so broad that it regulates almost every type of action. There is also a contradiction in people wanting data about their own medical treatment to be kept private while also wanting medical practitioners to base that treatment on data derived from the treatment of other patients.<sup>8</sup>
15. A further area of concern relates to the level of representation of the social sciences in government. In April 2010, in correspondence with Sir Nicholas Macpherson (Permanent Secretary, HM Treasury), the Academy indicated its concerns on this matter, particularly the then government's failure to reappoint when the post of Chief Social Scientist fell vacant. We strongly support the appointment of a Chief Social Scientist, working across all departments, as this might result in more general protocols to promote the availability of anonymised survey data for secondary analysis.
16. Finally, for many years Britain has led the world in terms of its rich data collection, particularly its cohort studies. Studies such as the internationally renowned birth cohort studies provide a wealth of information on the lives of British people from birth and throughout their lives, including information on their health, education, employment and social attitudes. But this investment in high quality data, and subsequent efforts to ensure open access to that data must be accompanied by national investment in greater analytical capability, particularly within government, to be able to exploit these and other data. The British Academy is pleased that the funding allocated to the 2012 birth cohort study was protected and that the

---

<sup>7</sup> <http://royalsociety.org/policy/projects/science-public-enterprise/>

<sup>8</sup> Baroness O'Neill's contribution to that discussion is available at <http://www.foundation.org.uk/events/audios/audiopdf.htm?e=442&s=1207>.

government recognised the importance of also investing in greater data analysis and data analytic skills.

17. We trust that the government will continue to recognise the enormous benefit of data resources such as these and the need to invest in data analysis. Given the availability of data, the government could very profitably deploy twice or three times the number of expert data analysts currently working on applied issues. It should also ensure that social statisticians outside government receive adequate encouragement and support to focus on policy-relevant issues.

18. In summary, The British Academy believes that:

- government and other publicly funded datasets should be made available for secondary analysis, provided that confidentiality is protected
- privately funded research data should be made available if there are potential public implications
- decision protocols for research projects should be publicly available
- work is required to establish clear protocols to deal with issues of cost, data security, data protection, and data commercialisation
- pseudonymised data is essential to much research in public policy and the social sciences, and should be part of any consideration of 'open data' policies
- there is a need for a statement of principles, setting out good practice in this area, to make sure all government departments are working to these standards and to ensure that political decisions do not lead to data being withheld unreasonably
- good practice needs to be shared across all government departments
- the appointment of a Chief Social Scientist to advise government is essential to address these issues
- investment in high quality data and ensuring open access to that data must be accompanied by national investment in greater analytical capability, particularly within government, to be able to exploit these and other data

The British Academy would welcome the opportunity to work with government and others in the future to strengthen our understanding of this important issue.

---

A submission from:

The British Academy  
10-11 Carlton House Terrace  
London  
SW1Y 5AH

For further information contact:  
Helen Haggart  
Senior Policy Advisor

---

*The British Academy, established by Royal Charter in 1902, champions and supports the humanities and social sciences across the UK and internationally. It aims to inspire, recognise and support excellence and high achievement. As a Fellowship of over 900 UK scholars and social scientists, elected for their distinction in research, the Academy is an independent and self-governing organisation, in receipt of public funding. Views expressed in this submission are not necessarily shared by each individual Fellow.*