THE ROYAL SOCIETY

UK Open Research Data Forum

This paper sets out the national and international context for the creation of a UK Open Research Data Forum, its role in stimulating action by the research community, and the recommendations of its first meeting (21/22 January 2014) of principles and priorities for action.

Summary

The massive increase in the rate of creation of new research data creates opportunities for new discoveries whilst posing a challenge to the maintenance of the principle of "self-correction" in science. If the opportunity is to be seized and the challenge to be confronted, there needs to be a radical shift of behaviour by scientists and institutions towards an open data regime in which research data is routinely made open to scrutiny and re-use, as the default position, and processes of data sharing are accepted as normative. The Open Research Data Forum is representative of the range of institutions involved in the publicly and charitably funded research process, with the purpose of articulating the principles and advocating practical steps in developing a UK open data regime whilst mindful of the developing international context. The paper identifies the principles that should govern data repositories, the incentives required to stimulate appropriate actions by researchers and institutions, the standards and tools that exist or need to be developed for an efficient open data regime and the need for skills in supporting it. It summarises the responsibilities for action that it argues should lie with institutions: funders of research, universities and other research-performing institutions, learned societies and publishers, and makes recommendations to Government. It sets out a number of opportunities for immediate action. The initiative draws its inspiration from the Royal Society report *Science as an Open Enterprise*.

A. The Open Data imperative

A.1 The data explosion and the changing the landscape of science and research

1. New means of acquiring, storing and analyzing data have created an unprecedented explosion of digital data in recent decades. Coupled with ubiquitous means of instantaneous communication, they are fundamentally changing the nature of the scientific enterprise. They permit analysis of large and complex datasets to reveal relationships in phenomena that have hitherto been beyond our capacity to resolve. They facilitate new modes of collaboration that increase the creativity of the scientific enterprise through interaction of many brains and many communities unbounded by institutional walls. They enable scientific concepts and the evidence that underlies them to be more effectively disseminated through society and in education, in ways that could change the social dynamics of science, making science a public enterprise rather than one conducted behind closed laboratory doors.

2. There is also a negative impact of the data explosion. It challenges the principle of "scientific self-correction" that requires both concept and evidence (data) to be concurrently published so that the logic of an argument and the integrity of data can be independently scrutinized and supported or refuted through attempts at replication. The widespread failure to ensure that large datasets that provide the evidence for scientific concepts, together with the meta-data (& the software codes) that are needed for replication, are concurrently published undermines self correction. The apparently increasing incidence of published results that are not reproducible is not only because of logical errors in an argument, but also because of the absence or partial absence of the underpinning data, metadata or relevant computer code, making replication impossible. There should be an absolute requirement for concurrent publication of the data underlying a scientific article. Such a requirement would also be a deterrent to what seems to be an increasing trend for the "cherry-picking" or the fraudulent invention of data.

3. The impact of digital technologies is not restricted to science, but creates challenges for the whole range of research and scholarship. In the "digital humanities" for example, research often entails new methodologies and intellectual strategies that are nonetheless grounded in traditional humanistic areas of focus (the nature of authorship, continuity of concepts over time, the social context of artistic expression). The challenges not only apply to data that are born digital, but also across large corpora of text, as well as visual, aural, audiovisual, sensory, neurological and even kinaesthetic forms of information.

A.2 Principles for open data, data sharing and data management

4. Four sets of principles need to be addressed if the above opportunities and challenges are to be grasped:

   a) The data that provide the evidence for the concepts in a published paper together with the relevant metadata and computer code must be concurrently available for scrutiny and must be consistent with the following criteria of "intelligent openness". To do otherwise should come to be regarded as scientific malpractice. The data must be:
      i. *Discoverable* – readily found to exist by online search
      ii. *Accessible* – when discovered they can be interrogated
      iii. *Intelligible* – they can be understood
      iv. *Assessable* – the reliability of their source can be evaluated
      v. *Useable* – they can be re-used
   b) The data generated by publicly-funded research that is not used as evidence for a published scientific concept should also be made intelligently open after a pre-specified period in which originators have exclusive access.
   c) Existing processes, reward structures and norms of behaviour that inhibit or prevent data sharing and new forms of open collaboration should, wherever possible, be reformed so that data sharing and collaboration are encouraged, facilitated and rewarded.
   d) Although the default position for data generated by publicly funded research should be one of openness as defined above, there are justifiable limits to openness. They are: where commercial exploitation is in the public interest and requires limitations on openness; to preserve the privacy of individuals whose personal information is contained in databases; where data release would endanger safety (unintended accidents) or security (deliberate attack). However, these instances do not provide justification for blanket exceptions to the default position, and should be argued on a case-by-case basis.

A.3 International and national trends

5. Science is an international enterprise, and needs highly permeable national boundaries. With this in mind, G8 science ministers in 2013 published commitments to an international open data regime, which is being followed up by the G8+5+1 group.[2] CODATA is promoting the concept of open data, several research intensive countries have developed or are developing open data policies and open data pilots are being developed by the EU as part of the Horizon2020 programme.

6. The development of enabling solutions for open data is inevitably an international exercise. For example the Research Data Alliance (RDA), an international collaboration launched early in 2013 and rapidly acquiring new members, implements the technology, practice, and connections that make data work across barriers and aims to accelerate and facilitate research data sharing and exchange. At a procedural level, a US initiative, SHARE, involving the Association of American Universities and the Association of Research Libraries, addresses the issue of the burden of compliance for individual researchers where there are many funding agencies with distinct policies and procedures. It is developing a single-deposit mechanism that allows principal investigators to focus less on process and more on research, and for researchers to access, reuse, and mine their colleagues' research results and data. In recognition of the vital international dimension, SHARE is being designed to facilitate international collaboration.

[2] Britain, Russia, France, Germany, Italy, Canada, Japan, US + China, Brazil, India, Mexico, South Africa + Egypt.

B. Establishing an open research data regime in the UK: the role of a Research Data Forum

7. The government, as a proxy for the public and national interest, is promoting an open government-data regime through the Public Sector Transparency Board and Departmental Transparency Boards, and more widely through agencies such as the Open Data User Group (ODUG) and the Open Data Institute (ODI). It has established a Research Sector Transparency Board (RSTB) to stimulate creation of an open data regime in publicly funded research.

8. Recognising that the involvement of institutions and communities that are engaged in the public research process is vital to the successful development of an open data regime, the Royal Society convened an Open Research Data Forum in January 2014. The Forum demonstrated a strong consensus amongst research funders, representatives of universities, institutes, learned societies, publishers and other bodies about the importance of open data and opportunities and priorities for establishing a powerful open data regime in the UK, that it was decided to maintain the Forum for the following roles:

    - to articulate the rationale, principles, processes and priorities for the development of an open regime for data produced by publicly and charitably funded research in the UK;
    - to stimulate a coherent approach across the all bodies involved in the research process;
    - to ensure that developments in the UK are linked, consistent with and able to influence international trends.

It will:

    - identify practical steps that should be taken by institutions involved in publicly or charitably funded research in the UK to develop an efficient and operational open data regime;
    - identify barriers to their implementation and how they might be resolved;
    - make recommendations to institutions and bodies that represent institutions that play important roles in the scientific enterprise;
    - advocate to Government any steps that it could take to support necessary changes.

It should not replicate the work of other expert groups, but should use and build on relevant work by others.

9. The Forum's role is distinctive, involving the bodies that have a primary role in implementing an open regime for data produced by and used in research. Creative interaction with the RSTB will be important in achieving the aims of both bodies. It is important to share approaches where relevant to do so across the government data/research data boundary, in areas such as licences; privacy; some technical issues such as core common standards & technical approaches to sharing, serving and accessing data, although there will be some significant differences in the scale & types of data in relation to their purposes, processes & the priorities of interested parties.

C. Principles, processes and technologies

C.1 Principles

10. Understanding this ecology and its self-organising character is important for researchers and for institutions in considering how they should exercise their responsibility for the data they create.

11. Learned societies are well-placed to lead their communities in developing best practice, defining the characteristics of the data produced in their field and contributing to development of a data taxonomy, a key to understanding priorities for data collection and curation and to identifying gaps or overlaps in standards.

12. Data where privacy is an issue needs to be discoverable so that access can be given under clearly-defined conditions. Researchers should be held accountable for their use of such data. As the aim of access is to allow the maximum benefit of reuse not the maximum reuse of data.

13. Data ownership, IP rights and licensing are important issues. Appropriate Open Licences are a key to reuse and to the Open Innovation we aspire too.

C.2 Repositories

14. There is great diversity of data repositories created from publicly- and charitably-funded research. Amongst those that are highly subject specific, they vary from the small scale held jointly by small groups of researchers, national databases, for instance those held by research councils, international databases for specific fields, for example the EMBL Nucleotide Sequence Database, generalist databases such as Data Dryad and databases held in so-called World Data Centres. Institutional repositories that cover a wide range of content are increasingly being developed. They have varied rules and procedures that govern access. Key issues for data repositories and their formal recognition are:

    - Registers or search mechanisms that permit data sources to be discovered.
    - The criteria for the data that they accept ("intelligent openness"?)
    - Development of persistent identifiers and hot links to data in published papers
    - Documented metadata standards
    - Linking databases, ontologies and the literature based on content analysis via text mining
    - Validation, consistency checking, ongoing update of databases by using evidence from the literature via text mining
    - Use of community data and meta-data standards.
    - The sustainability and inter-operability of the database.
    - Criteria for repositories that are acceptable sources of citation in published papers, or procedures for access to data by reviewers of submitted papers where there are issues of confidentiality.
    - Flexible schema for the integration of new data and data types in the future.
    - Criteria and processes for selection for retention ie. what data to store for how long, what to throw out, when and processes of perennial review?

C.3 Incentives

15. Section A argues that open data should be adopted as a normative principle for science and scientists. Pragmatism suggests however that this principle is most likely to be observed if there are incentives for researchers, their institutions and for users to adopt processes that serve this principle.

16. Improving the discoverability of published work is a major incentive for authors. As such, the development and use of efficient means of creating metadata is a crucial priority, to facilitate deposition and sharing so that it is not a burdensome overhead.

17. However, there is a tension between rich metadata (discipline specific and required for reuse) and broad metadata that can be used for discovery across fields. These are not mutually exclusive but serve different purposes.

18. The citation of deposited and openly intelligent data has the potential to be a major incentive for data originators. Devising experiments or observations that reveal hitherto unobserved patterns, relationships or causes is high level creative skill, and citation of the derived data should be valued at least as much as papers published in conventional journals. Indeed there is some evidence that such deposited data, when citable, draws many more citations than the first interpretation in a conventional article.

19. DataCite, founded in 2009, is a global network of nineteen national libraries, data centres and other research organisations that work to increase the recognition of data as a legitimate, citable contribution to the scholarly record. The British Library was a founding member. DataCite provides Digital Object Identifiers (DOIs) for data sets and other non-traditional research outputs, whilst DOI assignment helps to make data persistently identifiable and citable. Adoption of DataCite by UK research institutions would be a practicable and effective step in stimulating data citation and incentivising data deposition.

20. It is important that in data reuse, the originator must be cited, but must not be included as a co-author of the new output without consent. The data provider should not be liable for the outputs of secondary data.

21. The pharmaceutical industry increasingly recognises the importance of releasing all data on clinical trials, rather than partial release. What needs to happen in other industries to bring a similar situation about? Transparency, trust and cost are incentives for action.

22. The role of journals is crucial in incentivising data deposition, by mandating concurrent data deposition, in the development of data journals, by inserting hot links where this can be done, and in using open persistent identifiers using open standards to data in articles. It is important to develop generally applicable criteria for acceptable data deposition and an understanding of the appropriate business model(s) for data deposition and use to ensure that the process is not subject to excessive costs. Persuading editorial boards of journals that open data is a fundamental issue for the future of research is an important priority.

23. Journals should also capture a range of information about how data and papers are being used (views, bookmarks). It is important to know who is using which data and for what purpose.

C.4 Standards

24. Open licenses and open standards should be fundamental principles. Proprietary standards are a serious barrier to an effective open data regime.

25. Open data is defined in the Open Definition (http://opendefinition.org/) which provides key critieria including availability and provision under an open license which permits use for any purpose (including commercial activities).

26. A taxonomy of data could be created for aggregate, non-aggregate, personal and non-personal data etc. There is a concern that the lack of standards should not stifle sharing of data now.

27. Domain-specific standards can be adopted where applicable. Bodies already exist to oversee these: IUPAC (chemistry) and IUPAP (physics) for example.

28. Persistent identifiers using open standards and bi-directional linking are keys to discoverability and the linking of data ontologies.

29. Knowledge engineers need to work closely with researchers to help develop ontology and other standards, especially where regional vocabulary does not match global vocabulary.

30. Standards should be applied for discoverability, interpretability and subsequent reuse of the data.

31. Validating and promoting the openness and quality of data is increasingly important. There are various existing metrics for this (such as 5 stars, open data button or mark) and new ones are being developed (for example, the Open Data Institute offers an "Open Data Certificate"). The use of this as a standard to describe the quality of the dataset should be considered.

32. Providing sufficient metadata is context and user specific. For this reason almost 100 communities have developed over 500 standards (minimal information requirements, terminologies/ontologies and exchange formats) exist in the life, natural and biomedical sciences (source: BioSharing.org registry). Metrics are needed to identify the quality, usability and adoption of these standards, for example, linking them to databases and tools implementing them.

33. There is sometimes a gap between data and machine-readable data, and we should aim for the latter. This is key to linking datasets.

34. The legal framework for developing and maintaining open community standards is embryonic; more is required if we are to have both academic and commercial (e.g. pharmaceutical companies and publishers) to contribute to their evolution and use them.

35. There needs to be clarity and awareness raising by institutions to researchers about what the requirements of their grants are, and who owns the data so that they know how to make it open.

36. It should be recognised that professional expectations of best practice can be just as valid as formal standards (e.g. the Protein Data Bank).

C.5 Tools

37. Development of standards and procedures should be accompanied by the development of the tools that enable the application of such procedures in an economically and scientifically viable manner. Even for well-accepted procedures, the lack of appropriate tools can represent a major bottleneck.

38. Funding tool development can be difficult, since tools do not always involve development of new concepts but rather implementation of concrete solutions with existing technologies, but it is important that this is done.

39. Requisite tools are those for authoring, deposition, generating metadata, data reuse, and quality control. Some may be cross-domain, others domain-specific.

40. It is important in evaluating priorities for tool development that we focus on areas where adequate tools do not yet exist for important functions, rather than do needlessly replacing adequate existing tools. At the same time it is important to we are locked into the use of inflexible tools and processes in ways that inhibit exploitation of novel, more effective solutions.

C.6 The boundaries of openness

41. Although openness should be the default position for publicly and charitably funded research, there are legitimate boundaries to openness. They are for commercial exploitation, where the public interest in such exploitation outweighs that of openness; where use is made of personal data, where the right of privacy must be respected and where ethically robust data governance is required; and in relation to safety and security (see: http://royalsociety.org/policy/projects/science-public-enterprise/report/). These boundaries are however complex, and require a more detailed analysis than is possible here. The issue of data privacy for example is one of very considerable current debate within the UK and in Europe.

C.7 Skills and careers

42. There are now many training materials & programmes, for example those produced by the Digital Curation Centre, the National Centre for Text Mining, those developed by JISC, and those connected with the UK Elixir node that are designed to serve both researchers and technologists (such as curators, developers etc.) in the academia and industrial sector.

43. Training and capacity building are absolute priorities for universities – the focus should be on:

- Data literacy for all (undergraduate programmes).
- The development of generic data specialists, who might best be located as part of a library function.
- Specialists with domain expertise, who should be located in operational units (depts., centres)

44. There is currently no career pathway for data scientists/specialists. It will be difficult to generate the necessary cohort of specialists until this is done. There is currently a major skills gap. Some training has been incorporated into PhD programmes through CDTs but there should be a recognised distinction between basic data literacy for all researchers, and the supporting roles of data scientists.

45. There is a role for professional bodies in providing CPD for data scientists/specialists.

46. The above are major issues for university and institute libraries. Leadership is needed from the library community to address the issue.

D. The roles of institutions

47. A concordat on open data (analogous to the Concordat on Research Integrity) could be a powerful tool in helping to accelerate culture change. To be successful it would need to be taken forward in partnership between funders (research councils, funding councils, charities and bodies such as JISC), research institutions and the broader research community.

D.1 Funders of research

48. Deposition of appropriately open data should be a requirement of in allocating research grants. This includes both data that provides the evidence for a published scientific article and all other data generated during a funded project. Compliance should be monitored and the funder should certify whether appropriate deposition has taken place. Fulfilment of this condition should be a prerequisite for further funding. Ensuring compliance should be a responsibility of the institution, the researcher and the funder. Approaches and standards need to be harmonised as far as possible by research funders and in particular in Research Councils.

49. Data repositories should be deemed acceptable as locations for data deposition when they apply the criterion for of intelligent openness for the data that they accept, can demonstrate reasonable sustainability and preservation of datasets in their published form, provide stable identifiers for submitted datasets, use unique and persistent identifiers, are reasonably interoperable, provide expert curation, are broadly supported and recognized within their scientific community, implement relevant community endorsed reporting requirements, provide for confidential review of submitted datasets and allow public access to data without excessive restrictions.

50. It is important that the creation, maintenance, evolution and integration of community standards are funded, a priority that the deposition of "intelligently open" data is made as simple as possible for all those responsible and that data curation and sharing tools are made readily available if we are to empower researcher to provide rich descriptions of their data. It is important to be aware of current and emerging technical solutions, be open to international collaboration and seeks ways to accommodate principles for open data in public-private partnerships for funding research.

51. Funder should ensure that evaluation systems employed by them give weight gives weight to data publication so that it can be appropriately balanced with the credit given to conventional publication.

D.2 *Research performing institutions, particularly universities*

52. Data citation, if properly recognised by universities and widely understood by researchers, is a powerful lever and important incentive for open data. It is an issue that universities should take up with enthusiasm.

53. It is vital to develop, recognise and advocate efficient and effective practice in data management and to stimulate and adopt the results of studies of how this is best achieved. Institutions should be encouraged to collaborate in implementing such data management systems to common standards and protocols rather than going it alone. The current EPSRC deadline for data management risks inefficient hoc responses from institutions. We should make haste at a rate that ensures the best outcomes.

54. The broad base of technical knowledge and the institutional roles represented within the Forum can be a means of articulating best practice in supporting developments within universities and other research institutions, including drawing on earlier work and the work of bodies such as the Digital Curation Centre.

55. Phased implementation of open data regime through pilot projects could be an effective way of testing best practice. These could at include data-rich areas of natural science (chemistry is suggested, partly because of its large industrial hinterland & the involvement of the Royal Society of Chemistry in developing open data) and in the life and bioinformatics sciences (where drivers such as ELIXIR, in the UK and across Europe, already bring together key stakeholders together). There are excellent examples in fields of biology and chemistry (e.g. the Protein DataBank) where such principles are routinely adhered to. It would also be of value to include an area of the humanities where use of digital data brings benefit and where approaches to non-digital data could also be explored.

56. University libraries are currently exploring and adapting to the data needs of researchers. It would be of particular value if the Forum could articulate clearly what it believes to be the function of institutional libraries and their staff in the 21st century. Consideration needs to be given to the dynamic between appropriately skilled library staff, existing data curation staff and researchers, with suggestions for mechanisms to maximise this benefit.

57. Exploring the role of "data scientists" in institutions and the extent to which adequate numbers are trained to satisfy commercial and research needs to be a priority. Providing a career pathway for such scientists is an important priority. Good example of the effective deployment of such skills is important in developing new habits and processes of working with data.

58. It is important that universities consider how awareness and competence in data management and data handling can be embedded in the training of students, particularly but not exclusively in the STEM subjects. Researchers need to be made aware of the principles and practices that should underlie effective data management before embarking on a research project, rather than having to "reverse engineer" data management for data deposition at the end of a project.

D.3 *Learned Societies*

59. Learned societies and academies articulate the principles and priorities of researchers in the disciplines that they represent to a greater degree than do the institutions that employ researchers. They thus have the potential to fulfil a central role in stimulating responses to the challenges and opportunities of open data and data sharing.

60. They are means of:

- Spreading good practice and community standards, if existent in the area, and if not, fostering their collaborative development.
- Recognising and maintaining standards of best practice in research excellence through links to CPD.

- Defining the diversity of data types that could help develop an operational data taxonomy that would help in planning implementation processes
- Articulating a coherent industry – academia – research institute perspective in their discipline
- Taking part in and leading pilot schemes, especially where learned societies can provide connections between academia and industry
- Implementing open data policies in society-owned journals
- In publishing and giving grants in ways that help to their communities.
- In working with professional bodies to maintain domain-specific standards and best practices of research excellence.

D.4 Publishers

61. Research publishers play a fundamental role in enabling access to data. They should expedite the move to a position where are data are deposited concurrently with the articles based on them in community-endorsed discipline specific repositories or generalist ones (such as Data Dryad and FigShare). The responsibility for maintaining the data in repositories and providing access and support is best located in the research communities that create the data. It is important that publishers and the editors and reviewers - who come from the research community - should routinely require this as a condition of publication. To do otherwise should come to be regarded as malpractice.

62. Submission to a community-endorsed, public repository is already mandatory for many journals in the fields of DNA and protein sequences, macromolecular structures, microarray data and crystallographic data for small molecules. In these cases referees view the datasets and ensure that the data will be accessible upon publication via a unique accession number.

63. Many publishers (including Nature Publishing Group, PLOS, EMBO Press and Elsevier) strongly recommend deposition of other types of data sets into appropriate public repositories. In data- and computation-intensive fields (eg in molecular systems biology), some journals have, for several years, mandated deposition of all datasets (and software codes) that are central and integral to a published study, in community databases or, when no such database exists, required their inclusion in the published paper as associated dataset. PLOS will require deposition, posting, or appropriate means of access (for sensitive data) from March 1, 2014.

64. Helping to ensure sustainable, long-term, intelligently open archiving of the scientific record should be a core mission of scientific publishing. Their policies on data deposition should continue to evolve towards a state where deposition in a community endorsed public repository is a requirement of publication. PLOS is already moving to this position. Other publishers (eg EMBO Press) require sharing of published data even where no community repository exists, for example by using its own infrastructure to host the data. Progress will vary by discipline, and is dependent on:

    - Community norms of data sharing and use and support of community data repositories.
    - The establishment and use of data standards for particular data types.
    - Whether repositories are public and accessible to everyone and have permissive terms of use. Or for private data, that a mechanism exists for granting access on a need to know basis.
    - Mechanisms for journals to host and expose intelligently open curated datasets as an integral part of published articles.
    - Long term funding of a repository to ensure it is a persistent resource.
    - The level of technical support, capacity and sophistication of the data deposition process to avoid placing undue burden on authors.
    - Provisions datasets to be peer reviewed and robust linking infrastructure to allow persistent and unique linking between the paper and the dataset.

65. Publishers are encouraged to follow the example of those working to:

- increase the utility of the data contained in the articles they publish, including publishing the data points behind figures (source data), and reducing the data held within supplementary information.
- implement data citations within their articles and provide guidance to authors on the citation of datasets.
- Increase the utility and discoverability of data presented in published figures by providing the underlying source data and structured descriptive metadata.

66. Again we stress the importance of developing common procedures to minimise the burden of treating vital metadata must be a priority, although the major bottleneck is the lack of efficient tools that would enable authors, curators, editors to apply such procedures. An important development was initiated in the US in a White House paper (http://www.whitehouse.gov/mgi) about developing an open platform for innovation based on the theme of data – tools – skills (see also paragraph 6)..

67. It is important that models for data deposition ensure that barriers, including financial barriers, to database access and the integration of data from many sources are minimised, so that they can be exploited using common tools rather than by a plethora of proprietary tools that create barriers to effective data integration.

E. Low-hanging fruit: issues that should be acted upon promptly

68. We have identified several actions that could be carried forward with immediate effect:

a) The creation of a concordat between the principle representative bodies involved in funding, prosecuting and supporting research to support and apply the principles summarised in this paper.

b) The adoption of data-citation as a routine process in reporting research. We advocate use of the DataCite system that is already gaining ground as the norm in many research-performing institutions.

c) To promote agreement on standards with international partners.

d) To identify the need for the tools for effective data management and deposition where they do not exist or are inadequate and to create awareness of effective existing tools.

F. Recommendations to UK Government

69. Government should endorse the principles enshrined in this report, and that open data as defined here are the expected default position for research data funded from public sources.

70. Government should support the Forum in attempts to align International approaches to the challenge of Open Research Data

71. Government should work with the Research sector to support and help resource appropriate skills development

72. Government should encourage funding councils, research institutions and learned societies to develop credible and attractive career pathways for data scientists within the public sector – this applies also to the Government's own research capability

73. Government should work with the Research sector to ensure that EU legislation does not have a negative impact on the Open Research Data agenda.

Appendix 1 Introduction: The Forum

Purpose

The purpose of the Forum is to identify practical steps that could be taken by UK institutions to develop an efficient and operational open data regime; to identify barriers to its implementation and how they might be resolved; to make recommendations to institutions and bodies that represent institutions that play important roles in the scientific enterprise; and to advocate to Government any steps that it could take to support necessary changes. The principles that are suggested as the rationale for the work of the Forum are set out in appendix 4.

Background

The initiative to create the Forum derives from on the Royal Society's 2012 report on Science as an Open Enterprise[1] and on UK Government (Research Sector Transparency Board) and international moves (e.g. G8 – June 2013) to promote an open data regime. It has been conceived as a relatively small group to deliberate intensively on key issues in relation to open data for relevant UK institutions. It comprises members of research councils, funding councils and charities that fund research; universities and institutes that undertake publicly funded research; learned societies that reflect and influence the principles and priorities of their communities; those that publish the results of research; "intermediary bodies" such as libraries, and groups expert in the open data field; and private bodies, such as industry, that collaborate with publicly funded research institutions.

The focus of the Society's report was on the natural sciences, but its recommendations apply equally to many areas of social and medical sciences, where data is most frequently in digital form. However, digital technologies are not restricted to these areas, but create challenges for the whole range of research and scholarship and to "data" that is not easily rendered into digital form. Although the initial focus of work should be on digital data, the prospect of broadening its perspective to other forms of research information should be considered.

Next steps

It is recognized that individual institutions present in the Forum cannot necessarily speak for the communities of which they are part or which they represent without wider consultation and discussion within those communities. It is hoped that those present will stimulate the debate within their communities about the adoption of the principles and processes needed to underpin a vigorous, creative and cost-efficient open data regime. The meeting will be conducted under Chatham House Rules with a note of the meeting, once agreed, made publicly available.

The Royal Society is currently planning only for the January 2013 Forum meeting. It will be a question for participants whether they wish the Forum to have a longer life, and if so, how such a longer lifespan should be organised and supported.

The Minister of State for Universities and Science, David Willetts MP, has written to indicate his desire, as chair of the Government's Research Sector Transparency Board (RSTB), to use output from the Forum as key input to the work of the RSTB and in ensuring that the UK can take a lead in this area.

[1] The Royal Society: *Science as an Open Enterprise*. Report 02/12. June 2012. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

Appendix 2 Agenda

21 January 2014, Council Room, Royal Society, 6-9 Carlton House Terrace, London, SW1Y 5AG

12:30-13:00	Registration opens
13:00-13:15	Prof Geoffrey Boulton FRS Welcome and Introduction
13:15-14:00	David Willetts MP, *Minister of State for Universities and Science, "Why should the UK care about open data, and what might be the role of Government?"* Followed by discussion
14:00-15:00	Session 1 - B1 Agreement on Principles, chaired by Prof Geoffrey Boulton FRS (See Appendix 4)
15:00-15:30	Tea and coffee
15:30-16:50	Session 2 – B2 Issues for Institutions and funders These discussions are envisaged to be fluid and can extend beyond the sessions scheduled.
16:50-17:10	Tea and coffee
17:10-18:30	Session 3 – B2 Issues for Institutions and funders
18:30-19:00	Drinks reception, Library Events Room
19:00-19:45	Dinner - Library Events Room
19:45-20:30	Sir Mark Walport FRS – "*Why open data is an important priority for UK science"* Followed by discussion
20:30	Dinner ends

22 January 2014, Council Room, Royal Society, 6-9 Carlton House Terrace, London, SW1Y 5AG

8:30-9:00	Tea and Coffee and pastries
9:00-10:30	Session 4 – B3 Contextual Issues, led by Sir Nigel Shadbolt
10:30-11:00	Tea and coffee
11:00-12:00	Session 5 – B3 Contextual Issues
12:00-12:30	Wrap-up and next steps, led by Prof Geoffrey Boulton FRS
12:30	Forum ends

Session structure

B1 - Agreement on principles

Are the principles set out in appendix 4 appropriate, sufficient and acceptable?

B2 - Issues for institutions

In considering how an open data regime could operate in the UK in ways that observes the principles set out in appendix 4, the following are prompts for issues that arise for the institutions that would need to be involved:

a) *Funders of Research (research councils, charities, funding councils).*

- How effective is the increasing requirement for data management plans? Are such plans being implemented? Is implementation being monitored and are sanctions for non-compliance in place or being considered?
- Are the criteria for "intelligent openness" as advocated in the Royal Society report appropriate as a basis for setting standards for data deposition?
- Who should be responsible for setting criteria, monitoring compliance and implementing sanctions (research performing institutions or funders)?
- How can the assessment of research evolve so as to reward open data on an analogous scale to journal articles? Are funding councils considering such a change on an appropriate timescale?

b) *Research performing organisations, particularly universities.*

- How can systems of progression, promotion and reward be adapted to incentivise adoption of open data principles?
- Data management systems - are consortia the most efficient way of utilizing available resources?
- What principles should apply to the potential conflict between the scientific imperative for open data and any requirements for confidentiality in public/private funding or in realizing commercial benefit, and how should the interface be managed?
- Is the management of IP sufficiently flexible in ways such as suggested by the Hargreaves Review?
- How can demands on researchers, consequent of an open data regime, be minimized (e.g. equivalent of the US SHARE approach – appendix 6, par 13)?
- How should the library function evolve, what support should an institution offer to its researchers in data analysis and management and how should this be structured?
- Is the current level of training in data use appropriate to the needs of students and researchers, and is the output of specialist "data scientists" adequate for the actual and latent demand?

It is understood that the Russell Group of universities has a working group on open data. It would be helpful to have comments on its thoughts and progress.

c) *Learned societies*

- How are they able to play a major role in advocating the benefits of open science in their scientific communities?
- Could they be agents in developing and spreading good practices that are well-adapted to their specific areas of science?
- How can they promote collaboration to exploit the opportunities offered by effective data sharing, including awareness of relevant tools and training opportunities for members?
- Are there other roles that they could adopt?

d) Publishers of scientific journals

- How, and how quickly, can publishers move towards a requirement for concurrent publication of "intelligently open data" that provides the evidence for a published concept?
- What criteria should be used to determine the acceptability of a particular database for the above purpose?
- What is their view of the standards and protocols that should apply to the publication of such data and of best practice?
- Are there implications for the roles of referees?
- Is "publication bias" a serious issue in this context?
- Are they engaging with open and persistent researcher identification initiatives to ensure connectivity and accurate attribution of researchers and data?

e) "Intermediary bodies"

This phrase is used to describe bodies with an enabling, supporting and advocacy roles in data management, analysis and use. It would be helpful to have comments on the roles that these bodies might have and the extent to which integration of effort and processes of communication and support should be developed to which they might contribute.

f) Private performers of research and its users

Although the Forum meeting is primarily concerned with the "supply side" of research, the private/public interface and implications that an open data regime might have for the users of research are important issues. It is important to categorise the issues that need to be explored in these areas.

B3 - Contextual issues

There are a series of overarching issues that provide a context for the work of individual institutions. The following are suggested:

a) What should be the principles determining a national data ecology: which data, located where, managed by whom, and the international setting?

b) Definitions and standards

- Protocols for release and access
- Taxonomies and types of data
- Technical and metadata standards

c) *Legal, regulatory and ethical issues.* Managing the boundaries: business; privacy; safety/security Skills, training and career development

- For researchers
- For data and information specialists

d) Data for use

- Data publication and citation
- Intelligent openness as the criterion
- Environments for use of data
- A research agenda?

e) Priorities and quick wins

f) Recommendations for any actions by Government

Appendix 3 Participants

Those with a * attended for one day.

Organisation	**Name, Position**
Academy of Medical Sciences	Dr Naho Yamazaki, Head of Policy
Advisory Panel on Public Sector Information	Prof David Rhind FRS, Chair
British Library	Mr Richard Boulderstone, Chief Digital Officer
CODATA	Dr Simon Hodson, Executive Director
Department for Business, Innovation and Skills	Rt Hon David Willetts MP*, Minister of State for Universities and Science
Digital Curation Centre	Mr Kevin Ashley, Director
University of Oxford, representing Dryad	Dr Susanna Sansone, Associate Director and PI
Elsevier	Ms Anita De Waard, VP, Research Data Collaborations
EMBO	Prof Thomas Lemberger, Deputy Head of Scientific Publications
Figshare	Dr Mark Hahnel, Founder
Geological Society	Mr Neil Marriott, Director of Publishing, Library and Information Services,
GlaxoSmithKline	Mr Robert Frost, Policy Director, Medical Policy
Government Chief Scientific Advisor	Sir Mark Walport FRS*
HEFCE	Mr Mario Ferelli, Head of the Analytical Services Group
Intellectual Property Office	Dr Joanna Huddleston, Head of Copyright and Research
JISC	Ms Rachel Bruce, Innovation Director, Digital Infrastructure
Medical Research Council	Baroness Onora O'Neill FRS*
NaCTeM	Dr Sophia Ananiadou, Director
Nature Publishing	Ms Ruth Wilson, Publisher, Scientific Data
Open Data Institute	Sir Nigel Shadbolt, Chairman and Co-founder
Open Knowledge Foundation	Mr James Casbon, Board Member

PLOS	Mr Cameron Neylon, Director of Advocacy
Research Councils UK	Prof Rick Rylance*, Group Chair of Research Council UK Executive
Research Councils UK	Mr Mark Thorley*, Head of Science Information and Data Management Coordinator
Research Information Network	Dr Michael Jubb, Director
Research Libraries UK	Mr David Prosser, Executive Director
Royal Society	Prof Geoffrey Boulton, Council Member and Chair of Science Policy Advisory Group
Royal Society Of Chemistry	Mr David James, Executive Director, Strategic Innovation
Russell Group	Prof Nick Wright, Pro-Vice Chancellor (Research and Innovation), Newcastle University
STFC/RDA	Dr Juan Bicarregui, Head of Scientific Applications Support Division
Universities UK	Mr Paul Clark, Director of Policy
University of Warwick	Prof Peter Elias*, Institute for Employment Research
W3C	Mr Phil Archer*, eGov Consultant
Wellcome Trust	Mr Dave Carr*, Policy Adviser
Wellcome Trust	Ms Nicola Perrin*, Head of Policy

Observers

Organisation	**Name, position**
Department for Business, Innovation and Skills	Mr Ron Egginton*, Head BBSRC and ESRC Team
Government Office of Science	Ms Clara Davies*, Assistant Private Secretary to Sir Mark Walport
Royal Society	Ms Caroline Dynes, Policy Adviser
Department for Business, Innovation and Skills	Dr Anna Macey*, Policy Advisor, Research Funding Unit
Royal Society	Mr Tony McBride, Director of the Science Policy Centre
Russell Group	Ms Alison Torrens, Research Fellow

Appendix 4 Principles for open data, data sharing and data management

10. Four sets of principles need to be addressed if the above opportunities and challenges are to be grasped:
    a) The data that provide the evidence for the concepts in a published paper together with the relevant metadata and computer code must be concurrently available for scrutiny and must be consistent with the following criteria of "intelligent openness". To do otherwise should come to be regarded as scientific malpractice. The data must be:

        i. ***Discoverable*** **– readily found to exist by online search**
        ii. ***Accessible*** **– when discovered they can be interrogated**
        iii. ***Intelligible*** **– they can be understood**
        iv. ***Assessable*** **– the reliability of their source can be evaluated**
        v. ***Useable*** **– they can be re-used**

    b) The data generated by publicly-funded research that is not used as evidence for a published scientific concept should also be made intelligently open after a pre-specified period in which originators have exclusive access.

    c) Existing processes, reward structures and norms of behaviour that inhibit or prevent data sharing and new forms of open collaboration should, wherever possible, be reformed so that data sharing and collaboration are encouraged, facilitated and rewarded.

    d) Although the default position for data generated by publicly funded research should be one of openness as defined above, there are justifiable limits to openness. They are: where commercial exploitation is in the public interest and requires limitations on openness; to preserve the privacy of individuals whose personal information is contained in databases; where data release would endanger safety (unintended accidents) or security (deliberate attack). However, these instances do not provide justification for blanket exceptions to the default position, and should be argued on a case-by-case basis.

Appendix 5 The data explosion and the changing the landscape of science and research

11. New means of acquiring, storing and analyzing data have created an unprecedented explosion of digital data in recent decades. Coupled with ubiquitous means of instantaneous communication, they are fundamentally changing the nature of the scientific enterprise. They permit analysis of large and complex datasets to reveal relationships in phenomena that have hitherto been beyond our capacity to resolve. They facilitate new modes of collaboration that increase the creativity of the scientific enterprise through interaction of many brains and many communities unbounded by institutional walls. They enable scientific concepts and the evidence that underlies them to be more effectively disseminated through society and in education, in ways that could change the social dynamics of science, making science a public enterprise rather than one conducted behind closed laboratory doors.

12. There is also a negative impact of the data explosion. It challenges the principle of "scientific self-correction" that requires both concept and evidence (data) to be concurrently published so that the logic of an argument and the integrity of data can be independently scrutinized and supported or refuted through attempts at replication. The widespread failure to ensure that large datasets which provide the evidence for scientific concepts, together with the meta-data (& the software codes) that are needed for replication, are concurrently published undermines self correction. The apparently increasing incidence of published results that are not reproducible is not only because of logical errors in an argument, but also because of the absence or partial absence of the underpinning data, metadata or relevant computer code, making replication impossible. There should be an absolute requirement for concurrent publication of the data underlying a scientific article. Such a requirement would also be a deterrent to what seems to be an increasing trend for the "cherry-picking" or its fraudulent invention.

13. The impact of digital technologies is not restricted to science, but creates challenges for the whole range of research and scholarship. In the "digital humanities" for example, research often entails new methodologies

and intellectual strategies that are nonetheless grounded in traditional humanistic areas of focus (the nature of authorship, continuity of concepts over time, the social context of artistic expression). The challenges not only apply to data that are born digital, but also across large corpora of text, as well as visual, aural, audiovisual, sensory, neurological and even kinesthetic forms of information.

Appendix 6 International and national trends and actions

14. Science is an international enterprise, and needs highly permeable national boundaries. With this in mind, G8 science ministers in 2013 published commitments to an international open data regime, which is being followed up by the G8+5+1 group.[2] CODATA is promoting the concept of open data, several research intensive countries have developed or are developing open data policies and open data pilots are being developed by the EU as part of the Horizon2020 programme.

15. The development of enabling solutions for open data is inevitably an international exercise. For example the Research Data Alliance (RDA), an international collaboration launched early in 2013 and rapidly acquiring new members, implements the technology, practice, and connections that make data work across barriers and aims to accelerate and facilitate research data sharing and exchange. At a procedural level, a US initiative, SHARE, involving the Association of American Universities and the Association of Research Libraries, addresses the issue of the burden of compliance for individual researchers where there are many funding agencies with distinct policies and procedures. It has developed a single-deposit mechanism that allows principal investigators to focus less on process and more on research, and for researchers to access, reuse, and mine their colleagues' research results and data. In recognition of the vital international dimension, SHARE is being designed to facilitate international collaboration.

16. Within this international context, there are in the UK, three strands in an evolving process:
    a) The government, as a proxy for the public and national interest, is promoting an open government-data regime and has set up the Research Sector Transparency Board to do so for publicly-funded research.

    b) Recognising that the involvement of institutions and communities that are engaged in the public research process is vital to the successful development of an open data regime, the Royal Society has convened this Open Data Forum. It recognizes that an efficient regime will be one that is well-adapted to the characteristics of the UK research system.

    c) Those that are technically adept in the data–science field need to provide solutions to specific problems and procedures that minimize cost and maximize benefit.

[2] Britain, Russia, France, Germany, Italy, Canada, Japan, US + China, Brazil, India, Mexico, South Africa + Egypt.