European
Commission

Impact of Open Science methods and practices on the economics of research and science

Case Studies from Life, Mathematical and Social Sciences

Professor Jonathan Adams

December 2015

EUR 27398 EN

Research and Innovation

EUROPEAN COMMISSION

Directorate-General for Research and Innovation
Directorate A – Policy Development and coordination
Unit A6 – Data, Open Access and Foresight

Contact: Katarzyna Bitka

E-mail: katarzyna.bitka@ec.europa.eu

RTD-RISE@ec.europa.eu

RTD-PUBLICATIONS@ec.europa.eu

European Commission
B-1049 Brussels

EUROPEAN COMMISSION

Impact of Open Science methods and practices on the economics of research and science

Case Studies from Life, Mathematical and Social Sciences

Professor Jonathan Adams

Directorate-General for Research and Innovation

2015

EUR 27398 EN

LEGAL NOTICE

This document has been prepared for the European Commission however it reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

More information on the European Union is available on the internet (http://europa.eu).

Luxembourg: Publications Office of the European Union, 2015.

doi:10.2777/14

ISBN 978-92-79-50444-0

ISSN 1831-9424

Contents

CHAPTER 1: SUMMARY

This report is about Open Science, how it changes research methods and practice, and the effect of this on the economics of research. 'Economics' should not be understood as narrow accountancy. Research objectives become ever more ambitious, building on discovery and facilitated by innovation. The 'cost' of an activity – in time and consumables - is continually reduced by technology so the scope of the research enterprise can expand. The technological and sociological elements of an 'open science' economy affects research by enabling us to tackle entirely new challenges and its impact is not through reducing costs but through accelerating and amplifying benefits.

Summary questions

No simple definition of 'open science' was produced by the interviews, meetings and background reading for this report. The consensus is that it refers both to technological advancement, particularly in networks and in data management, and to social adaptation among researchers, particularly in collaboration and transparency. The latter is accelerated and made more pervasive but not initiated by the former.

The questions considered in this report focus on three practical issues relating to the emergence of this 'open science'.

- Is there a 'step' change in the research landscape that requires specific action now?
- How pervasive and diverse are the implicit changes in the landscape?
- What steps could any agency reasonably take in response to change?

Summary outcome

- The research base is responding well to open science because it was pre-adapted by many earlier changes. There is no step change: new technology and open data facilitate evolution rather than triggering it.
- The important policy issues will be to enable researchers to continue bottom-up to adopt and adapt. No top-down action in regard to assessment, funding or management is required beyond those already made by the European Commission, but additional facilitative actions would enable better community response.
- First, recognition via assessable, transitive credit for data providers whose data is re-used could be implemented to address valid concerns.
- Second, the costs of making data both open and useful – particularly the management of vast numbers of small datasets rather than a few big data aggregations – could be more clearly recognised and managed.
- Open research is collaborative and research opportunities cross national borders. There is a risk that national, domestic innovation policies may turn inwards. It will fall particularly to trans-national European Commission programmes to maintain open science across borders.
- The effect of open science on research has primarily been via enhanced benefits and expanded opportunities, but a possible approach to model the effect on specific costs can be outlined.

Summary discussion

The wider economic and societal benefits of open research and open data are likely to be significant and pervasive but are not part of this report. However, the pre-Web shift towards 'Mode 2' knowledge transfer recognised that researchers were increasingly working with economic and social beneficiaries. Direct and early user-engagement has become a well-established part of research.

Open science and open data are linked aspects of a continuing beneficial cultural shift towards more open research practice (OECD, 2007; RDA Europe, 2014). Open science has been associated in part with the potential for collective intelligence that is apparent in the development of the Internet (e.g. O'Reilly, 2005; O'Reilly and Battelle, 2009), and some commentaries particularly link

‘Web 2.0’ with a ‘Science 2.0’. This has been promoted by those agents who advocate not only more transparent public accountability, since much research is funded from general taxation, but also greater openness in research design, project management, data release and publication practice.

Open science has made research challenges more feasible. It may make a unit of research activity less costly, but that ‘unit’ is overtaken by new opportunities. We can sequence genomes, not just parts of genes. We can compare massive biomedical databanks, not just search for individual cases. To say that open science has reduced research costs is meaningless: open science has made unaffordable research become economically realistic.

Open science is adaptation and evolution, not disruptive revolution. It affects the business of science insofar as it grows out of new technologies adopted by researchers because they further existing objectives. Some research expectations and needs helped to drive the emergence of the technology. Many of the characteristics of open science existed before Web 2.0 and some pre-date personal computing.

Open science is pervasive. This report looks at research practice in environmental sciences/ecology and economics in some detail and also at some examples in social sciences, plant sciences, and mathematics. Interviews cover changing practice over several decades, in some instances referring back to 1945.

- Environmental and economic research sub-disciplines have for decades been engaged in multi-disciplinary teams, engagement with users, very big data, and cross-border collaboration. They are part of the demand-side that helped drive the development of communication and computing technologies (e.g. international trade; climate change).

- Social and economic sciences have well established data archiving, aggregation and sharing. Some social sciences already insist on posting of data and code to support publications (e.g. Journal of Peace Studies).

- Plant sciences have created excellent shared databases to support international work (e.g. DFG’s BrassiBase).

- Mathematics remains an innately individual practice, and it is not an experimental science so it does not generate data, but it has a fundamentally open research culture and rapidly adopted new technologies to further that culture (e.g. the Polymath project).

Not all sub-disciplines and researchers engage so strongly with open research. Micro-economists develop detailed analysis and theory that they prefer to bring to maturity before publishing, in order to achieve maximum impact. Field ecologists worry that, if the data that they collect is immediately made open, then analysts might exploit it before they have time to develop their own publications. Past experiences appear to support the concerns of both groups.

Not all ‘open’ elements are consciously endorsed – or even understood - across all research disciplines and institutions (Van den Eynden and Bishop, 2014). Senior and well-informed researchers question whether ‘open science’ has a clear or common meaning. Radical voices are keen to be heard, and are more easily distinguished; a majority more tacitly accept the ethos of open science because it is already ‘business as usual’; while an uncertain body of ‘shy dissenters’ are passively resistant.

If change is already in progress, however, then direct intervention and significant modification of the research process seem unwarranted. Research funders and managers should instead look more to mechanisms that facilitate further change, which an innately entrepreneurial research base can then manage on its own.

- There needs to be a better credit mechanism for high-quality, usable data.

Data should be seen as a valid primary output of research but due credit for this is inadequate at present. Analysts are closer to publication than fieldworkers and are felt to receive disproportionate esteem, and sometimes to exploit data without due acknowledgment. A better system for recognizing and giving credit is absolutely required to ensure open data equity. This will need to be equivalent to existing citation credit for journal articles. To do this, archived data must be identifiable and tractable (via DOIs for every dataset) so citations to papers using the data can be transitively aggregated back to the data source.

- The transaction costs of open data have not been fully considered.

‘Big data’ sources have an unquestionable value and enable quite new analytical developments. By contrast, many fields produce very large numbers of relatively small and specific datasets. Aggregating these smaller datasets requires sufficient metadata to support re-use, which implies additional work for original researchers, additional curation and new archiving systems and processes. Good archiving curation encourages data deposit. It will be a problem to search masses of small data. The cost-benefit of this data proliferation has not been assessed. The costs are real, however, and unless additional resources are found and allocated for this purpose they will be incurred at the expense of other research.

- Open science is international science.

Open science is highly collaborative in project inception, development and validation. Internationally collaborative co-authorship has grown hugely over the last 30 years. Data show that cross-border co-authorship is associated with researchers from leading institutions and with the most highly-cited part of publication databases such as *Web of Science*. International collaboration is thus beneficial to good science, but it conflicts with an emerging focus on harvesting economic and social impact for domestic beneficiaries. It is important that funds continue to be available to address cross-border research opportunities. The Horizon 2020 programmes are an ideal complement to domestic national funding if evaluation thresholds and criteria are properly applied.

CHAPTER 2: CASE STUDIES: LIFE AND SOCIAL SCIENCES BEFORE AND AFTER OPEN SCIENCE

The evidence in this chapter is organized by discipline. The original EC specification for this project indicated a requirement for findings from two case studies - one in life sciences and one in social sciences – so as to estimate the impact of open science by comparing practice before and after its appearance so as to assess its impact. Such an approach might allow comparison and contrast that would provide detail of change, and its constraints, by exploring and addressing two distinct disciplinary domains. The specification also argued that the general uptake of open science practices by life sciences was more advanced than in social sciences.

The report's title also refers to the economics of research, part of the specification. This can refer to both the cost-accountancy of research and the wider research economy. This is best addressed by looking at technologies' impact both on cost-reduction and on benefit-gain by enhancing affordability and feasibility.

2A. Methodology and studies

Consultation included groups and individual researchers from the environmental sciences, economics and mathematics, to discuss their experiences of and reactions to open research concepts and practice. This included individual interviews (senior and junior researchers, librarians, research managers, policy consultants), 'open data' meetings (ecology), a panel for a Q&A conference session on open science (research managers), and an informal expert group (mathematics).

The work started with discussions with senior researchers, policy/funding managers and policy analysts, to identify case study topics and correspondents. This covered both the generalities of open science and specific issues around the initial target disciplines (ecology and economics). It then proceeded through:

- A general review of recent reports and background literature, which reflected early discussions in being almost entirely positivist.
- Attending scientific meetings at which there was an opportunity to discuss open science.
- Structured interviews (outline below) with researchers and managers, emphasizing questions around 'business change' before and after the emergence of innovative technology and desirable future changes to research funding and management.

2A (i). The business of research

The economics of research includes what can be done and how it can be done as well as how much it costs. For example, scientific communication by letter (1960s) is succeeded by telephones (1970s) and jet flight enables regular attendance at international conferences (1980s). These developments are succeeded by widespread personal computing and the Internet (1990s). So, after 2000, the ability of Web 2.0 to (vastly) accelerate communication is a facilitator of what was already an expanding cultural drive to increase interaction.

Functionality before and after change can be compared so as to ask how change has affected what can be done. This may be facilitative (an increase in volume or a reduction in time) or creative (where a function enables a wholly new opportunity).

The rapid acquisition and transfer of massive amounts of information made possible by Web 2.0 is having significant effects, but it is not the only significant technological influence. While it is expected to accelerate some processes, this may not always be true where human factors, e.g. in data acquisition, remain critical. Communication has been vastly improved, in frequency and coverage, but there are again human limits and exceptions may exist, e.g. because of language factors. Similarly, collaboration can occur over a much greater distance, but feedback indicates that more than technology is required to make these successful and the technology itself needs to be used appropriately to add value.

Accessibility to research results may be much better, or at least potentially better, but access alone is not sufficient to deliver valuable outcomes. Again, there are more complex factors so, for example, while vast quantities and types of data are now available, the accessibility or utility of some data is constrained because the data are poorly curated.

These are generic characteristics, not specific to any field, so questions may be asked irrespective of disciplinary context. This provides reasonable comparability between fields for the purposes of this study.

For open science:

- What is now easier than it was before?
- What is now feasible that you wanted but were previously unable to do?
- What can you now do that you had never imagined?

Again, these are all generic characteristics for comparisons.

A third area of consideration is that of constraint. New technologies may become available that would evidently enable more, better research but they cannot be accessed for cost reasons. Or there may be cultural constraints, such as language or personal safety barriers. The 'rate-limiting factor' may – as indicated earlier - be human (e.g. interpretation of outcomes) rather than information transfer and processing.

An essential counter-factual part of any economic analysis is to ask not only the question about 'what has changed' but also 'what additional net benefit can be demonstrated?' For open science the benefit is much clearer than the cost reduction.

2A (ii). The economy of research

Can we calculate the cost-benefit of changes in research enabled by open science? Yes, but it is not informative to work from old research and calculate cost reductions. Instead, we could deconstruct, for example, the sequencing and comparison of genomes across Eurasia, which has recently given us new insights into human relationships and opens new possibilities for health research. It also exemplifies the massive acceleration of technology, collaboration and data processing (see The 1000 Genomes Project Consortium, 2015).

This work started with the Human Genome Project in 1990. The first complete genome was mapped by 2003. The recent report is for the 1000 Genomes Project (started in 2007), which actually covers 2,500 individuals in 26 distinct populations. This work required massive collaborative activity for many teams, adopting comparable methodologies, sharing results and processing vast data through a common analytical system. Ten years ago the laboratory work was beyond the resources of any research agency or group of agencies. Twenty years ago the data analysis would have been a fantasy. Technology and Open Science practice have made this feasible, affordable and deliverable. The data outcomes are huge and provide pervasive value to many other research domains.

A valid accounting analysis could be done by comparing the costs of sequencing one genome for 1000 Genomes (2007-2015) with an estimate of how many work hours were required for Human Genome (1990-2003) prior to the introduction of new technologies, including collaborative networking. This is a non-trivial task, but it is possible to identify the elements such as networked collaboration, data sharing and data analysis that relate to 'open science' and consider how these would have functioned.

At a specific level we can point to technology with identifiable cost impacts. For molecular biology, the auto-cycling PCR machine enabled one researcher rapidly to perform multiple operations that previously would have taken days to do. The cost per test was vastly reduced, but the science then rapidly changed so that a single test was no longer meaningful. The lesson is that when evolving infrastructure changes the nature of research then it is only fleetingly possible to compare like-for-like costs.

This is captured in comments from researchers:

> "Before computers, we had field notes and laboratory books and then we published in journals. Now data is digital, what we can do has expanded into large volumes, rapid analysis, complex models and it can be readily shared." (Professor, Water Resources)
>
> "What is 'possible' is what developed. Researchers can adopt new technology for their own use pretty quickly. We share possibilities as well as data." (Professor, Ecology)

Open Science is made possible by Web 2.0 technologies that have changed the ways in which people can do science, so the comparison of before and after can be couched in terms of methodologies and benefits but not meaningfully in terms of itemized cost. Furthermore, since publicly funded research is not an innately profitable business, there is no savings or surplus to show. Instead, change allows new, faster and better things to be done with the resources that are available.

2A (iii). Constraints on this analysis

CONTINUITY. Defining a threshold to compare 'before open science' and 'after open science' is a problem because no single event, such as the availability or adoption of a specific technology, marks where science becomes demonstrably more open. Discussions around this are reminiscent of those around the emergence of Mode 2 knowledge production (Gibbons et al, 1994). A new mode of working may already be so much a part of normal practice that recognizing its novelty or identifying the point at which it was novel is almost meaningless. An exact definition of 'before and after' is therefore frustrated by evolution of research practice that has built on cultural change, increasingly sophisticated research challenges and the rapid exploitation of improvements in computing and electronic communication of ideas, progress and outcomes, so that discoveries and learning capabilities became more rapid and more effective over decades.

POSITIVISM. Information and opinion is skewed towards a positivist view of the benefits of open data and the transformative nature of open science. It is difficult to deliver a balanced view of costs, as well as benefits, because informants almost universally endorse the need for and value of open science. A diverse expert group of scientists and policy analysts across Europe were asked to identify researchers who had a more critical view of open science: no significant criticism emerged from this. Instead, informants identified researchers and commentators who had a well-informed view, usually because they had been advocates of open science before the emergence of Web 2.0.

Analysis and conclusions may thus be affected by a deficit of evidence regarding people who are unaffected by or even opposed to (elements of) open research. First, as a good outcome, a long evolutionary shift that has made the practice of open research an implicit part of contemporary research culture means that there are few researchers who are antagonistic. Second, because of the predisposition of many research-funding agencies towards open research practice (for data and publications) there is a positivist assumption and a consequent 'shy dissenter' problem so that concern about open research is not expressed in public.

2A (iv). Specifying the entities to be studied

Early discussions suggested that a variant to the original EC specification would be more informative.

- Disciplinary variation in the changing economy of research within life sciences is considerable with major differences between Organismal and Molecular science. For example, the trajectories of development in 'ecology' and 'plant sciences' provide useful comparative information.

- Ecology is only one aspect of developments of policy relevance in environmental sciences which include e.g. physical geography, atmospheric science and economics; these all offer valuable insights on the evolution of research cultures.

- Some social science uptake may lag behind, but this needs to be examined at a fine-grained level. For example: in macroeconomics, the drive to address large datasets and longitudinal studies is very marked; in microeconomics, the research culture can be rather different.

- Data are core to open science so there is pervasive links to the research culture of mathematics, yet mathematics as a research economy has a different relationship to source data for its own evolution.

For these reasons, a view taken from two case studies, one in life sciences and one in social sciences, would be unduly arbitrary, partial and limited. The range of informants was therefore broadened around 'ecology' to include both plant sciences and environmental sciences and around economics to include more diverse evidence. The report also includes mathematics because it was evidently a core enabling technology for open science in all fields yet rather different in its own culture.

Detailed examples form the 'case studies', but experience is very diverse and a single case study cannot be representative. They are therefore a backbone to which other findings are attached. However, case studies described in other projects do not greatly differ in their conclusions.

2B. Environmental science

Environmental sciences, including ecology, show the long-term development of 'open science' prior to the Internet. They also depend on extensive cross-national collaboration and data sharing, but interviews reveal a problem of attribution and credit associated with open data access and sharing.

SHARING DATA. The early development of international programmes in environmental science means that **collaboration** and **data sharing** started with large-scale geosphere-biosphere programmes in the 1970s and 1980s. Ecologists may work on particular locations such as large lakes where the development of long-term, shared data sets is important. Other research is based on comparative studies with significant geographical or temporal separation, so comprehensive **datasets** must be available to **enable follow-up** work. For example, 'transects' across latitudinal series' of sites in Europe are used to support local studies, to feed into studies of climate change effects across a long North-South baseline, and to build into detailed climate change monitoring over decades. This requires a high degree of **sustained collaboration**, openness and sharing regarding methodology and a preparedness to contribute to group outcomes not all of which carry every contributor's name.

DATA CREDIT. A culture in some areas distinguishes, and gives recognition to, those researchers who are extremely effective at data collection through careful fieldwork and those researchers who are particularly strong at analysis. For example, since the late 1970s, a number of studies of animal population dynamics driven by theoretical input and modelling from a group at the University of Oxford under Professor Lord May PRS have drawn on detailed fieldwork and laboratory studies by researchers at Imperial College London. Other Oxford studies have modelled fish-stocks using data from the UK Fisheries Laboratory at Lowestoft, the International Whaling Commission, and others. New Mathematical Ecology models by these groups were driven forward by the adoption and application of existing technology from maths and physics.

ESTABLISHED GOOD PRACTICE. A group of younger ecologists and plant scientists who contributed to discussions were clear that **data sharing is essential** to any progress. One PhD plant researcher expressed this in terms of making data available for other people to analyse so as to accelerate problem-solving:

> "If someone else can get the work done then we can get on with other things."
> (Researcher, Biotechnology)

The DFG programme on 'Evolutionary plant solutions to ecological challenges' (http://www.ruhr-uni-bochum.de/dfg-spp1529/Seiten/index.html) is an exemplar on open and accessible data. This has been used as a case study of **incentives for data sharing** and is described in an excellent report from Van den Eyden and Bishop (2014). In this network it is expected that supplementary data are made available as 'proof' supporting analytical results. The work is also linked to the BrassiBase database of gene data.

CASE STUDY – WATER RESOURCES. A group of researchers at the University of Leeds (UK) were working with Yorkshire Water (a regional public water utility) in the 1980s, on water quality, resource management and consumer behaviour. They have since then developed complex data models, created a consultancy organization, assisted in the development of accessible data resources, and extended research activity overseas.

The core research platform on water quality and chemistry, catchment management and economics, and a reputation for competent and expert data analysis and modeling, enabled receptive early engagement with (public sector) industrial partners and multi-disciplinary, multi-agency teams. This enabled rapid exchange of information at the outset and during the implementation of projects, leading to strong work specifications that were informed by the end-users as well as the researchers, and very effective interpretation of outcomes because of direct user-agency involvement in project development.

This was **prior to the Internet**. Initial discussions revealed that water utilities held significant data resources but needed not only help with analyzing and using that data but also in scoping the challenges they faced and defining solvable problems.

> "At first [1980s] they would send us data on floppy disks. If it was the wrong data then I had to phone up and explain what we really needed. Then they would extract that and put it in the post. In the 1990s, I was allowed a 'dongle'

> on my computer which enabled me to access their IT system directly." (Professor, Water Resource Management)

Computing power led to better analyses, and networks enabled rapid data transfer. But this was not entirely beneficial because the interaction with the user changed.

> "In the 1990s we were modeling GIS [geographic information systems] on Sun workstations. Clients would bring teams in so we could take them through the data, the analysis and what it meant. What we did then was as complex as anything we did later on PCs. The problem later was that many more client staff could access the results via the Internet and they didn't understand what they were looking at." (Former academic; now senior consultant)

> "The industry understands the openness agenda. They can see the value of peer-reviewed publications, which benefit them in arguments to government. But they still have issues about core information they don't want published." (Researcher, Freshwater ecology)

Internet **technology built on this experience** and extended water research and environmental management consultancy into the Middle East. A project in Abu Dhabi was instanced as an example of what is now feasible. Data collected in the Gulf can be made available on an essentially continuous basis and analysis in Europe can tune fieldwork. Desk-to-desk Skype conference calls mean that discussions about fieldwork, data, interpretation and policy implications take place frequently and at marginal cost.

The technology acceleration from 'data by post' to Skype calls does not create a complete solution to faster research because human factors are central to effective cooperation.

> "Skype doesn't overcome the value of face to face contact when you start any collaboration. You still need feet on the ground at the start of the project and to keep regular checks on what is actually being done. And the data can come in partially-digested attachments so you have to go back for a lot of additional information to enable analysis." (Professor, Geography)

International collaboration emerged with the growth of the Web and a 2011 European project on climate change effects on the regulatory regime for sea-bathing with teams in Barcelona (Mediterranean coast), Germany (Baltic) and Leeds (North Sea). Technology enabled rapid **data exchange and modeling** of climate projections and effects. In the UK this was followed by the inception of an Environmental Virtual Observatory (EVO) to explore how technology could be used in environmental management across three universities (Prof Bridget Emmett, Bangor Centre for Ecology & Hydrology; Prof Adrian MacDonald, Leeds; Prof Robert Gurney, Reading) and drew in other partners. It was designed to make environmental data more visible and **accessible** to a wide range of potential users and provide tools to facilitate the integrated analysis of data and **visualisation of data assets**. This was a two-year pilot project supported by the Natural Environment Research Council (NERC) with a focus on the "Sustainable use of Soils and Water" as an area with potential to illustrate the benefits of cloud computing to climate change impact models driven by projections from Global Circulation Models (GCMs: twenty-two GCMs with differing climate projections are used in the 4th IPCC assessment).

A **data-sharing problem** arose in one multi-party project.[1] The project linked specialists who were planner/managers, field-workers, technical and data/analytical. As projects became very large and multi-site, the team cohesion was less easily maintained. An analytical group went so far as to publish without acknowledging the project management, field-workers or other groups because they had not recognised the obligations implied by the project structure. This is an early 'echo' of a generic problem about **acknowledging specialist roles** and **appropriately citing data** as a critical resource.

Concluding points. This case study shows that technology enables a mature research program to cover huge geographical distances, many partners, a diversity of disciplines, a mass of data and complex modeling. This has not been achieved suddenly but through **progressive development** of working methodology alongside evolving technology, and human factors remain critical to good cooperation. As complexity has grown, **data-attribution** – and recognition of other specialist roles - has emerged as a challenge to good working relationships. There are well-curated and managed

[1] The specific project is intentionally not named.

databases emerging in specialist areas that provide **good practice**. Younger researchers are supportive of open research.

2C. Economic and social sciences

Economics has a long tradition of data development, collaboration and the use of technology. Economists have long been involved in large-scale, data-intensive international projects and have global (IMF, OECD, World Bank) and national (ESRC, Leibniz Institute[2]) institutions that aggregate and curate core datasets accessible for general use. The general principle of open publication and data sharing is well-established in macro-economic research but microeconomics offers some contrasts.

CASE STUDY – UK AND GERMAN DATA ARCHIVES. The UK's Economic and Social Research Council (ESRC) has been a lead organization in developing the quantitative and methodological research environment in the UK, supporting economic and social science **data storage and access** since 1967. The UK only established a Research Council for social sciences in 1965 but it was preceded in 1963 by the Social and Economic Archive Committee (SEAC), established to investigate and propose solutions to the problem of **sharing information** about social surveys and their **data**. SEAC was particularly concerned that costly survey work was being replicated through poor communication between researchers and that data were being 'lost'. To meet this concern SEAC compiled an inventory of survey data that could be made available for secondary research.

That led to a Social Science Research Council (SSRC) 'data bank' which became a 'survey archive' in 1970 and then the 'ESRC data archive' in 1980. It held such a wide and comprehensive range of data that it was referred to in a House of Lords debate (1984) as the "Rutherford Appleton Laboratory of the social sciences world". It provided methodological advice as well as holding core government data drawn from censuses and surveys such as the UK Family Expenditure Survey. Its position as a neutral agent (funded by a public agency, providing support and training for the HE sector) is seen as a factor in **encouraging its widespread use for data deposits** as well as data sourcing.

A main initial drawback was a lack of SSRC computing capacity, but there were also human factors. Some comments on the early days are reminiscent of current concerns about open data:

> "There was convincing evidence that there was a large amount of data just ready and waiting for an archive to hold them. But, when the archive was established, most of the potential depositors identified in the study found **reasons not to hand over their data**. Some claimed to be still working on them, other people had developed **an almost parental attachment** to them, and there appeared to be serious legal obstacles to acquiring the large datasets collected in government social surveys." (Research Manager, Economic science)

In Germany also, there are major social-economic datasets **aggregated and curated by key centres**. For example, GESIS is an archive for socio-economic data managed by the Leibniz Institute for Social Sciences. The centre curates (by standardizing variables in order to facilitate comparisons across time or regional units) and manages surveys that are required to **comply with defined methodology** and technical requirements. These are archived and processed according to **recognized standards** and then made accessible. Reference studies for which there is a large demand are combined into 'study collections' and the centre directs the development of software for data processing, archiving, and analysis as well as online-search.

The UK archive has now become the UK Data Service but remains an ESRC-funded comprehensive source of economic and social data and a **single point of access to validated and curated** census data, government-funded surveys, longitudinal studies, international macrodata, qualitative data and business microdata. It supports training and guidance for data creators and users and promotes data sharing to encourage data reuse as well as developing **good practice for data management**.

ESRC remains concerned about the future development of the UK research landscape and recently (March 2015) organised a workshop on 'Data for Discovery'. The headline concerns from this meeting were: the need for incentives to show that the research system **values open science** and rewards those who follow good practice; alternative forms of output to offset traditional, peer reviewed journal articles (referring to the need for **credit for data sources**); and the importance of methodology and of analytical tools as forms of **research contribution**. The workshop

[2] http://www.gesis.org/en/institute/gesis-scientific-departments/data-archive-for-the-social-sciences/

provoked agreement on priority issues but was less clear about what practical actions were required to support open research in economic and social science.

SHARING DATA. Macroeconomics and studies of international trade led themselves more readily to **open data sharing** and collaboration because informed outcomes require both multiple data sources and equal insights on the policy constraints for each national economy. At the University of Cambridge, the economic demands of rebuilding the post-war European landscape led Richard Stone and James Meade to develop the basis of modern national accounting. With John Maynard Keynes, Stone established the Department of Applied Economics (DAE) in 1945 as a research focussed organisation. The DAE rapidly became a key data centre, collating large-scale datasets, **leading in the use of technology** for analysis and modelling (using suites of comptometers in the 1950s before computers became readily available).

COLLABORATION. There is a natural **relationship with economic research users** in government and industry and relatively open flow of people between academic and government departments. The Cambridge DAE attracted visitors from elsewhere in Europe and from North America, even in the 1970s. Despite limited computing power and communications, international economic analytics were enabled by mailing boxes of computer punch-cards back and forth across the Atlantic.

OPEN DATA. In other areas of the social sciences open research is more recent than in economics but has been the case for more than a decade. One interviewee drew attention to the area of international relations and an essay by Gleditsch (2003) describing the practice in the *Journal of Peace Research* of requiring the posting of data and computer code for all quantitative articles. Dafoe (2014), reporting via the Social Science Research Network, comments that:

> "… for the majority of published statistical analyses, readers have to trust that the scholars correctly implemented the many stages of analysis between primary data collection and the presentation of results — including data cleaning, merging, recoding and transforming, analysis, and output"

Despite this, he found more than two-thirds of studies made supplementary data available.

DISCIPLINARY CONTRASTS. Microeconomics does not necessarily follow the same model of openness as macroeconomics. For example, a centre may develop a unique line of interpretation and understanding of taxation that would lead to changes in current policy. To achieve maximum impact the researchers are **unlikely to publish anything but a fully developed and validated model**. As an instance of this, Jean Tirole (Toulouse) won the 2014 Nobel Prize in economics for work on game theory and the organization of industry, specifying conditions under which competition or collaboration might be favored. Three other groups had been working competitively in the same area, racing to reach the critical theoretical insights. It is extremely unlikely that any of these groups would have been willing to share let alone openly publish their current thinking.

Concluding points. Social sciences generally, and economics particularly, have embraced open science, collaborative international **data sharing and data archiving since the 1970s**. However, some **sub-disciplines are less disposed** culturally to adopt open research. **Incentives to encourage sharing** are essential and mediated by effective archive management and support, which requires funding. Formal **credit for multiple research roles**, including data production, is seen as essential.

2D. Mathematics

Open science is intimately bound up with open data and collaboration and, in this context; mathematics poses something of a strange position. People who will make a major contribution to solving the problems of handling big data will have a mathematics background: data analysis is now an important technical specialization in its own right. However, first, mathematics is not an experimental science. Mathematicians develop methodology to apply to multiple data problems rather than generating data themselves. Second, mathematics is not a collaborative science. The single author paper is still prevalent in mathematics. (These points are, of course, not equally true of applied mathematicians let alone statisticians, who may be both highly collaborative and significant data producers.)

PROBLEM SHARING. Mathematicians, if neither experimental nor collaborative, have nonetheless enjoyed an open problem-solving culture, and this has been accelerated by Web technologies. This may explain the contradiction regarding comparative attitudes to open data in Youngseek's findings (Youngseek and Stanton, 2012), because the sharing is **problem-driven not data-driven**. A core part of this is the approach to developing solutions to published conjectures and problems. If other mathematicians deem the problem to be ‘interesting’ then they work on it as individuals, and then share and contribute to an iterative series of partial but improving solutions through publication.

POLYMATH FORUM. In January 2007, Professor Sir Tim Gowers FRS (University of Cambridge) posted a blog asking “Is massively collaborative mathematics possible?” in which he proposed a new format for solving these complex research problems, based on **a large number of small contributions** from many mathematicians in a public online forum, as opposed to the traditional model of a small number of collaborating privately and intensively. From this blog he launched the first “Polymath” project.

This approach has been credited with the rapid development of solutions to the ‘twin primes’ problem building on the work of Yitang Zhang (Nature, 2013; Zhang, 2014). The (very old and unsolved) conjecture was that primes came in pairs (twin primes) no more than 2 numbers apart; Zhang published a solution that set the maximum distance between twin primes as no more than 70 million; others saw this, seized on the analysis and successive efforts rapidly brought that boundary down.

A point made in interview by several mathematicians was that the Polymath approach was successful because it provided an **accelerator** to what their culture had previously sought to do rather than actually changing the culture: “this is what we do anyway”. While they might seem to work on their own, in an un-collaborative way, in practice they succeed through their contributions to wider networks.

The Polymath approach also highlights the more ‘continuous’ nature of open research, unlike traditional episodic methodology based on funding-project-analysis-publication. No formally agreed, peer reviewed publications were involved in the iterative reduction of the twin prime boundary. This is a feature of applied research projects where feedback from end-users can be incorporated into current work and thus modify objectives as the project develops, often without the need for a series of intermediate publications. It points towards the idea that, in the future, researchers in all disciplines may publish only a handful of ‘original’ papers, and that most work will go into the validation, development and modification of these seed-concepts by themselves and others.

Concluding points. The Polymath project demonstrates the way innovative technology can enable the acceleration of group solutions through rapid feedback in an open forum that allows due credit to participants.

CHAPTER 3: LESSONS LEARNED

This chapter develops the disciplinary evidence (Chapter 2) into themes supported by literature review and broader, cross-disciplinary interviews.

No sharp, disruptive transition into open science was identified in interviews with researchers or in case study examples. Web 2.0 is technology that facilitates or accelerates the **continuing evolution** of research culture. Some characteristics associated with open research came before low-cost travel and personal computing. The factor most frequently mentioned in the context of recent change is the greater **general accessibility** and volume of data and easier personal and **data communication**.

The research base is naturally entrepreneurial, acquiring new technologies and applications and exploiting them to advantage. Discovery is not always about new data but also about methodologies and analysis that allow better interpretation or understanding. With Web 2.0, the emphasis is on the speed with which research can **progress because of more rapid access** to information and access to a more comprehensive range of information. The practice of research has changed less for some researchers and for others there are new possibilities because of data access.

Changes made possible by open science will **appear progressively** not disruptively and **none demand pervasive top-down restructuring** of research funding, management and evaluation. No one spoken to as part of this report saw a need for change to assessment of proposals, or evaluation, nor for significant change to funding mechanisms (although nuanced changes were highlighted).

Resources are already being used more effectively than in the past because links are being made, sometimes between contemporaneous data sets and sometimes between new and older data. One key part of the way forward is through **better management of data resources** as well as more open availability of data.

3A. Time-track of evolving research culture

> "This is a process of co-evolution, not revolution. Researchers exploit new technology very quickly and sometimes their expectations help the technology to develop." (Senior university manager)

Open science is part of a very broad change in **the relationship between science and society**, working at different rates according to culture, discipline and methodology. A summary timeline would need to go back to the 1940s so as recognise that technological change and social change cannot be sensibly separated. The most recent technology appears in a social structure (a research economy) that is already undergoing change and is thus able to adopt and exploit the opportunities that the technology provides.

	Post-1945 recognition of need for international collaboration
1950s	International research facilities and programs (1951 - European Coal and Steel Community programme links research base and industry; 1954 – CERN founded by international subscription)
	Data archives and centralised data collation (1957 – ICSU founds World Data Centre to archive/distribute data from the International Geophysical Year)
1960s	Public policy expectations of research and innovation
	European Union sectoral research programmes (1984 – subsumed in EU First Framework Program)
	Rapid decline in travel costs and growth in air travel
	Growing engagement between research producers and users
1970s	Development of accessible computing power

	Central main-frame computers in universities; access by distributed terminals
	Development of data management software
	BASIC enables widespread custom software development amongst experts
1980s	Personal computers (early versions in the late 1970s)
	Organisational intranets
	Packages for text and data management
	Emergent internet (1988 – first direct US-Europe IP connection)
1990s	Increasingly usable and accessible Internet
	Emergence of WorldWideWeb and first web browsers; accessible on-line text and data resources
	1991- ArXiv preprint repository founded
	1996 – Page rank algorithm massively improves web search opportunities
2000	Improvements in text and data management and website design and accessibility; proliferation of data from research; logistical, commercial and political issues about data access
	2004 – Beta version of Google Scholar provides free access to cited research literature
	2006 – European Research Advisory Board report 'Scientific Publication – Policy on open access'; National Science Foundation (USA) endorses open access to data 'Cyberinfrastructure Vision For 21st Century Discovery'
	2009 – US government launches open data portal
	2010 – EC High Level Expert Group on Scientific Data report 'How Europe Can Gain From The Rising Tide of Scientific Data'; OpenAIRE launched
	2012 – LinkedUp project launched; EU publishes data protection regulations
	2014 – EC public consultation on 'Science 2.0'

The evidence of past change suggests that facilitation of **adaptability** can achieve more than potentially disruptive managerial, structural change. The structure of research, particularly science, has moved from an individual activity through organisational shifts to a national enterprise and international networks (Adams, 2013). Both researchers and research organisations have adapted to this continuing process, which has internal and external influences.

Internationally coordinated research programmes and data management have been part of economics and environmental science for many decades. The case study in economics started with the post-1945 need to rebuild the European economy, which led to cross-national economic theorisation, modelling and data sharing. This was exemplified by the ICSU launch in 1957 of a coordinated data archive for the International Geophysical Year. For ecology, the 1972 UN conference on the environment built on prior international research and assisted the further collation of common methodologies and data sets for environmental benchmarks and monitoring, leading to the 1980s International Geosphere-Biosphere Programme.

Fine-grained evolution is seen in the case study of water management. Groups of environmental research producers (academics) and users (staff from public-sector bodies) discussed problems associated with uncontrolled run-off: researchers were looking at changing land use, vegetation cover and drainage rates while the research users needed to control river flows to mitigate erratic flooding in natural and urban environments. Meetings improved information exchange and refined research objectives and application.

These are (unconscious) responses, at international and regional level, to what has been described as a shift to a risk-conscious society, where risks go beyond the individual and require coordinated societal action (for the environment specifically back before 'The Limits to Growth' [1972] and conceptualised more generally by Beck [1992]). The management and outcomes of research no longer have esoteric status as a 'desirable' but become central and relevant to common objectives. Thus, the prioritisation of research choices and design of research activity is of wider significance and the knowledge outcomes of research and their applications are more directly and immediately relevant to society and the economy.

The environmental science case study reflects **changes in the relationship between research/knowledge producer and user** that were recognised in the concept of 'Mode 2' knowledge (Gibbons et al, 1994). A team crossing disciplinary and functional boundaries meets for a project or longer program to work on specific problems, linking knowledge domains and linking problem, solution and application. Traditional (Mode 1: academic) processes of research-analysis-publication-read/use are replaced by more continuous interaction and cyclical knowledge exchange between producer and user with flexible objectives and rapid dissemination of outcomes. Methods, data and conclusions are more transparent and accessible because the user is involved.

Gibbons et al (and see Nowotny et al, 2001) trace the growth of Mode 2 back to the post-1945 period but Loet Leydesdorff (Etzkowitz and Leydesdorff, 2000) sees Mode 2 as the original C18th formulation of the research enterprise from which Mode 1 was a derived C20th construct while others (e.g. Fuller, 2000) argue that both Modes were evidently present in the C19th origins of the modern German university.

Web 2.0 enables more rapid and high-quality international collaboration, but **for researchers this was already growing**. In 2014, more than 50% of the journal articles indexed on Thomson Reuters *Web of Science* and with an author from France, Germany, the UK or indeed any western European nation were co-authored with an author from at least one other country.[3] This change was the result of **steady and continuous internationalisation** since the early 1980s, when less than 5% of papers had international co-authorship (Adams, 2013).

The economics/social science case study describes the early development of mass data archives: for economics from the late 1940s, and social sciences more generally from the 1960s. For ecology this develops from the 1970s. We can thus trace a history of **extensive international data cooperation** to the 1970s and earlier, a growth trend in international co-authorship to the early 1980s, and a growth in producer-user cooperation and collaboration (labelled as Mode 2 knowledge sharing) to the 1980s. This suggests that the appearance of **the Internet is a facilitating event** as a means of rapid communication of data, information and ideas (removing the need for sending card decks and floppy disks by post) rather than an immediate causal factor. The extent to which it is transformative is dependent on interpretation. It engages with transformation in specific fields (such as environmental science) and other social change (such as producer/user engagement) and structural change (international collaboration) which long precede it and to which **the system was already adapted** or adapting.

Concluding points. Web 2.0 has enabled an acceleration and intensification of internationally collaborative research and activities such as data exchange, data aggregation and storage, and knowledge transfer that are part of collaboration but it does not create any disjunction. If there is no 'step' change then there is no argument for a top-down structural response in funding or organisation. Facilitating the adaptive response of the research base is likely to be better support for the best research.

3B. How pervasive and diverse are the changes in the landscape?

Attitudes towards open data sharing are **diverse and inconsistent between and within disciplines** (Van den Eynden and Bishop, 2014). Research reveals disciplinary differences in data sharing practice and more granular differences between research groups, because different disciplines (macro- and micro-economics) perceive data as culturally different or play a different research role. At a somewhat more simplistic level, Youngseek and Stanton (2012) suggested that

[3] Eastern Europe is catching up: about one-third of Poland's 2014 papers had international co-authors: see Supplementary Information at go.nature.com/nszeck

data sharing was important among biologists, chemists and ecologists, but less so for computer scientists, engineers and mathematicians, but this may be an issue of interpretation about the status of source data for basic and applied science. Significant variations in attitudes and practices are found within each discipline as well (RECODE, 2014).

Sustained international collaboration, openness and data sharing have 'always' been a part of culture for some researchers. Comparative studies with geographical or temporal separation require comprehensive datasets to support follow-up work. Ecological transects from Norway through Spain enable climate change research and monitoring over decades. Macro-economic datasets are made up of many individual contributions through national agencies.

However, because of the fine-grained disciplinary differences, there is no single response. In some areas, researchers will continue to 'guard' their data and theories while research is in play. It would be of no advantage to impose the need for immediate open data publication or international collaboration on such groups.

The clear **benefits of Web technologies** to such well-established work is, first, that databanks can be set-up and fully curated at known locations (generically via international repositories such as GenBank or ENA for gene sequence data; or specifically via DFG's BrassiBase, an online knowledge and database system for Arabidopsis researchers) and provide rapid access to all participants. Second, close collaboration can occur immediately across Europe and daily over much greater distances e.g. between Sweden and New Zealand or the UK and Abu Dhabi.

A **problem with existing funding** is that much is national, and national priorities are now influenced by a parochial impact agenda. The EC Framework programmes have no national boundary, so Horizon 2020 can play a vital role in by-passing the double jeopardy experienced when two national research groups in a joint project seek funding from their respective national agencies. This is ideal complementary facilitation of work better done at cross-national level than within a single country.

Concluding points. Change is widespread, but there is no single model of disciplinary reaction to the opportunities of open and collaborative research and there are positive and negative contrasts at a fine-grained level in natural and social sciences. Generally, observing the benefits of open research practice will encourage 'bottom up' change. Regional rather than national agencies will have an essential role in sustaining international collaboration.

3C. Data protection and credit for individual researchers

A repeated concern amongst researchers is that making source data openly available creates an undesirable lack of sight regarding subsequent re-use. This is primarily about a fear that no 'due credit' credit given to the source. The economics case study showed that similar issues of data ownership arose in creating the UK economic and social data archive. Unless this is solved, it will restrict commitment to data archiving.

Data is **not given the same esteem** as other research outputs. This point was made at the ESRC workshop and by ecologists. An indicator is the frequency with which data is submitted as an 'assessable output' in the UK's cyclical university research assessment process (the RAE, called the REF in 2014). In each cycle, each academic submits four outputs to a peer panel for assessment. Outputs can be publications (books, articles, proceedings), or other forms of output (patents, videos, performances, software and databases). In the RAE2008 cycle, there were 1,011 instances of data/database as a submitted output in a total of 214,287 submitted items – around 0.5%. In the REF2014 cycle there were just 136 instances of data out of 190,962 items – less than 0.1%. The frequency of patents was unchanged – around 220 - in each cycle.

The **lack of indicators of utility**, or any connection to analyses based on the data, may be an explanation as to why the creation of important data sets is not seen as an academically excellent activity. By contrast, journal articles for which citation counts are available are now more frequent outputs in the UK research assessment system.

The ecology example showed that diverse research contributions can be recognised, but due reward is not universal. There are also projects where the analysts 'run away' with the data and publish, which one argued was valid:

> "If you took the data, did the analysis, did the interpretation and wrote it up then why would you not publish? What contribution did the data guy make? Did they show what it all meant?" (Junior researcher, Mathematics)

Mathematicians were relatively unconcerned about this issue, possibly because they are more concerned with methodology. Contemporary macro-economists are familiar with large-scale datasets over which there is no personal 'ownership'. Ecologists have a wide range of views. A particular data set may support an extended body of work (e.g. data collected during early fieldwork, possibly overseas). In such cases there is significant resistance to full disclosure of original material that forms the focus of theorisation, modelling and analysis; the same situation arises for micro-economists. Clashes between experimentalists and modellers about credit for ecology concept development in the 1980s were cited as an example.

To offset these concerns, it is essential that efforts are directed towards **a system that gives credit for using data** that is at least **equivalent to any citation** to the contents of a research publication.

A better system of acknowledging all **specific researcher contributions** to research outcomes may be required (Allen et al, 2014). Authorship is a standard signal of IP ownership, so citation of the paper is a route to accumulating esteem. However, first, arbitrary decisions may be made about who is a named author and, second, authorship is no indicator of the scale or mode of contribution. Some researchers are naturally gifted in laboratory practice or the use of particular technologies ('green fingers' in the language of plant science research) while others are proficient at planning, coding or analysis. The 'authorship' of a paper should be more complete and more specific.

Data sharing metrics would also be important. Daniel Katz (University of Chicago) has proposed a system of transitive credit.[4] In summary, archived data is DOI-tagged; papers using data reference the DOI; citations to those papers are then transitively assigned to the data DOI. Through this route an **index of data utility** is rapidly established because datasets frequently used in papers that are then cited will be seen to acquire significant citation counts.

Concluding points. Insufficient credit and esteem is given to data as a research output and to the researchers who create valuable datasets. A better credit and tracking system would encourage open data sharing. The EC may provide leadership to establish an appropriate system for such research data protection and credit, discussed in the next chapter.

3D. Big data vs lots of small data: compliance and curation costs

Open data means **more data**: that is more data to collate, clean, curate, archive, validate and more data to monitor. There may be a major problem from **many small data sets** as a few massively large datasets.

The cost of compliance with open data mandates from research funding agencies is a cause of real concern for many researchers. Tracking all data during acquisition, and assessing, harvesting and storing associated metadata, is a clear cost. There is a cost in setting up systems to ensure that all these elements are acquired and then in curating the material. The cost to researchers is reduced if there is a good institutional system in place, but data collation would be more effective if such systems were coherent across institutions. Publication is a further cost, and the total researcher time absorbed in these activities may be non-trivial. Some part of the system – the research group, institution or funder - needs to absorb these costs and that can only be **at the expense of other research**. There are also costs for institutions in establishing a culture of data sharing and providing appropriate training to incorporate this in the development pathway for all young researchers.

There is a need for meta-data to enable data to be properly re-used. This is recognised when seeking to use historical data in longitudinal studies. Ecologists and economists have experienced problems with datasets that lack sufficient additional information to be certain about collection methods, associated conditional variables, or other factors affecting interpretation. This must to be solved for contemporary data

There are implications regarding monitoring and validation by other expert researchers. The **curation costs of processing and posting 'approved' datasets** is potentially very large for agents such as the Leibniz Institute and the UK Data Archive. It is essential, however, that good practice in data management is followed because this makes researchers **confident that data deposits are properly managed**. If significant effort is devoted to open publication of data and supporting metadata, then it seems reasonable to infer additional costs in validating those data – if not systematically then at the point of re-use, or otherwise there is little point in publishing the

[4] http://www.slideshare.net/danielskatz/transitive-credit

additional material. There seems to be no published economic analysis of these costs, although they are discussed in recent reports that arrive at different conclusions as to who should pay.

There is a completely separate set of issues about the costs of massive data management for economic benefit. Several people pointed out the need to distinguish between the benefits of ‘big data’ and **the costs of ‘lots of small data’**. If the benefits of open data publication are to be fully realised then there are downstream implications for drawing these small datasets together, mapping it and making it accessible for wider re-use. These costs are associated with the creation of data centres and repositories.

Is all data in existing data centres equally useful? Almost certainly not. There is an unproven value in archiving many small and disparate datasets with uncertain metadata and that do not obviously join up with anything else. It is unclear what benefits or new insights might be derived from investing in linking all these datasets.

<u>Concluding points</u>. The costs of managing and maintain standards in an open data environment have not yet been fully recognised. Data management costs to individual research teams are not trivial. High standards are required to make aggregate data searchable and useful and to inspire researcher confidence. There is a risk of massive numbers of small datasets swamping the system.

CHAPTER 4: RECOMMENDATIONS FOR POLICY OPTIONS

4A. Steering policy development

Open science is not revolution but acceleration, made possible by information technologies, of changes that have been evolving over a long period. The research base responds entrepreneurially to change and will continue to adopt and adapt 'bottom up'. The existing management structures of research do not require significant 'top down' modification. Adaptation is patchy but individual sub-disciplines will respond flexibly to enable the best science to exploit available technology.

Although the content and objectives of research proposals are likely to become more strongly data driven, peer referees will remain competent to judge all components of these proposals in terms of values such as challenge, credibility, innovation and excellence. The evaluation process does not fundamentally change where data analysis rather than experimentation becomes the main focus.

There are, however, a number of areas where the European Commission could guide and facilitate further evolutionary change with benefit to the research environment. This is not a prescription for action but an outline of opportunities for leadership from the Commission that may help steer policy development elsewhere.

- CREDIT: Play a lead role in addressing the need to ensure that data credit is given where it is due, a key problem around open data and research (action within 2 years).

- COMPLIANCE: Recognise and assess the potentially significant compliance costs – in time and effort – of onerous mandates on open data and publication (2-5 years).

- CURATION: Explore the disparate growth of huge volumes of small, poorly documented datasets (within 2 years). Assess the management costs and benefits associated with these compared with well-curated, properly aggregated big data (within 5 years).

- COLLABORATION: Focus on adequate support for cross-border research projects complementary to the domestic focus of national research bodies (on-going).

CREDIT (action within 2 years): A system to give assessable credit to those who create useful data and make it open is required very soon. Data science has already emerged as a distinct discipline and data scientists will be the equivalent of the super-technicians who powered up leading laboratories in the past. That role needs to be recognised and appropriate training and support should be given at institutional level.

While data scientists usually work at the analytical end of the research process, close to interpretation and publication, data creation is associated with researchers in the laboratory and field. In multi-disciplinary, cross-border projects recognition in regard to data creation may be weakened. Lack of adequate recognition, credit and esteem undermines the emergence of open research. New systems are required that record and track data storage and re-use (e.g. via DOIs). Those who use archive data must cite those DOIs accurately. Instances of use and citations to user-publications must be collated and aggregated to enable the development of 'transitive credit' or similar indicators of data value and excellence. Standard practices should therefore include:

- Archive data in an open data system with a trackable DOI.

- Cite DOIs consistently when researchers use the data archive.

- Gather and report DOI data citations.

COMPLIANCE (2-5 years as good practice emerges): The real cost of open data needs to be audited as soon as possible 'Open data' may involve significant additional work on the part of the team producing the data. Wherever the costs of this are absorbed (project funding, institutional management, researcher time) that will reduce the resources available for original research.

To justify such work, it would seem appropriate then to pay attention to the information given, which implies monitoring and validation by other expert researchers. Consequently, in an open data research 'ecosystem', both compliance and validation invoke additional resources. The number of steps required to make the open sharing of data effective are considerable. Even an outline deconstruction of the process of data development also suggests that the cost of compliance may be much higher than advocates suggest. A very simple set of steps should include:

- Open publication of initial ideas
- Publication of project/grant proposals
- Data management plan
- Data collection and provision of associated meta-data
- Data curation (cleaning, transforming, analysis)
- Replicability and repeatability

A data management plan could be part of a pre-proposal and thus subject to discussion and revision, but in some niche disciplines this will create significant problems in revealing IP around developing concepts, theory and models.

Publishing data has a value, but effective re-use requires that additional information is available. Only a full description of how data were acquired, including any relevant meta-data, would enable later analysts to determine how they would process the data, possibly in a comparative analysis or via incorporation into a larger database. A complete report on data curation and how the original data were processed for analysis may be seen as obligatory as part of research validation.

Standard methodologies in many laboratory-based fields mean that data management plans are reducible to codification, but the collection of original field data for ecologists introduces a range of complexities and exceptions. For macro-economists, the problem is usually addressed by adherence to standardised data definitions and collection methodologies (e.g. those promoted by the OECD). However, a researcher on economic migration described a lengthy analysis where even basic national data were clearly subject to reinterpretation and valuation.

Replicability and repeatability traditionally refer to basic experimental work and carry meanings according to discipline. In laboratory-based research it infers that the experiment can be repeated elsewhere with the same result. Some attempts to do this between laboratories highlight the very specific conditions required and this adds significant value to the outcome. In field-based subjects, economic and environmental conditions carry so many variables that replicability may be problematic, so appropriate replication is usually built into initial research. In a data-driven age, repeatability refers as much to the analysis as to the original research. Sufficient information on data handling, statistical methodology, model assumptions and so on must be given that another researcher could deliver the same conclusions by repeating the analysis on the same dataset. This is widely discussed in current literature, particularly in reference to the need to examine basic assumptions about data distributions and statistical tests rather than to consider only the statistical probabilities of outcomes.

CURATION (rapid assessment, then response over 2-5 years): The challenge of massive amounts of small data needs to be acknowledged and evaluated and a management plan developed. While massive, complex datasets attract much attention and use, and well managed data centres undoubtedly play a key role, there will be a vast number of instances of open data in small, very specific datasets that do not immediately link to anything else and are poorly curated and labelled. The cost of linking is unknown, the responsibility for this is unclear and the benefits uncertain. The extent to which available open data are being re-used with identifiable value has not been audited. There may be a new information problem for researchers in trawling for nuggets of high value in this flood tide.

COLLABORATION (ongoing): Funding for high quality, cross-border research need to be safeguarded. If Europe is to continue to take advantage of the international research developments seen in ecology and economics then Horizon 2020 will have a critical role in supporting international activity that complements domestic activity. Open research can be cross-border and international research co-authorship continues to expand. Multinational papers tend to be more highly cited, perhaps because it is only worth paying the cost of collaboration for tangible benefits. But national research budgets may work against this collaboration by focussing on retainable domestic benefits (especially economic and social impact). The UK's recent Research Excellence Framework (REF) exercise emphasised impact as well as excellence and is likely to be emulated elsewhere. This focuses policy attention on the tangible, non-academic returns on investment. It risks an environment that turns against research projects for which the benefits cannot all be harvested locally. Added to the 'double jeopardy' encountered by research groups applying to two agencies in different jurisdictions, this may undermine excellent international projects. It needs to be redressed.

4B. Cost analysis

Finally, an important issue that it was not possible to address fully within the format of this report is that of the economic cost-benefit of open science. Informants repeatedly emphasised that the principal economic impact of open science is on benefits, not costs. However, it is reasonable to raise the question of how a cost analysis might be applied.

The financial data available on historical research activity are likely to be poor and the project may not be readily related to current activity. Instead, it would be more informative to examine a recent project for which more reliable cost data are available and deconstruct this in terms of comparable past activity.

It would also be necessary to apply estimates to benefits. Key impacts of open science relate to rapid data sharing and analysis. We can argue that the data would have been shared anyway, that integrated databases would have been compiled and that analyses would have been progressively developed. But what is the benefit of immediate sharing, rapid integration and powerful analysis? What is the benefit of making the outcomes rapidly available to a wider range of potential users?

Open science enables rapid and effective collaboration over distance. It would therefore be appropriate to analyse an example project that reflected this. A standard 'laboratory project' would be easier to deconstruct and cost but would show little of the relevant enhanced benefits.

Examples of steps to analysis could include:

1. Project preparation

Costs: are there any improvements that reduce costs to the principal investigators?

Benefits: are project proposals more readily subject to criticism and amendment so that the final project plan is better refined (background, methods, planning, risk avoidance)?

2. Project management

Costs: what are the specific activities and their unit costs? What is the person time, machine time, consumable cost for each unit of research? What would those costs have been historically?

Benefits: how much faster can data be exchanged and collated between subsidiary groups? How can groups learn from one another in real time within the project and thereby refine what they are doing, improving methodology and sharpening objectives.

3. Project outcomes

Costs: what is the staff cost in collating and analysing results, including accessing and using reference databanks? What would the costs have been at a benchmark point in the past?

Benefits: was reference data available previously? How does the availability of reference data enhance the outcomes? How much more rapid is analysis? What additional forms of analysis can now be performed? How are the results made available to potential users?

REFERENCES

Adams J (2013). The fourth age of research. *Nature*, 497, 557-560. http://www.nature.com/nature/journal/v497/n7451/abs/497557a.html

Allen L, Brand A, Scott J Altman M and Hlava M. (2014). Credit where credit is due. *Nature*, 508, 312-313. doi:10.1038/508312a

Beck U. (1992). *Risk Society: towards a new modernity*. Sage Publications. ISBN 9780803983465

Clarke M, McDonald A and Boden P. (2013). Examining the relationship between debt and deprivation in the UK water industry. *Applied Spatial Analysis and Policy*, 6, 47-68.

Dafoe A. (2014). Science deserves better: the imperative to share complete replication files. Political Science and Politics, 47, 60-66.

Etzkowitz H and Leydesdorff L. (2000). The dynamics of innovation: from National Systems and ‘‘Mode 2’’ to a Triple Helix of university–industry–government relations. Research Policy, 29, 109–123.

Fuller S. (2000). The Governance of Science. (2000). Open University Press. Buckingham. ISBN 0335202349

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P and Trow M. (1994). *The New Production of Knowledge: the dynamics of science and research in contemporary societies*. Sage Publications. ISBN 1446265871

Gleditsch N P. (2003) Symposium on Replication in International Relations (editor). *International Studies Perspectives,* 4, 72–107.

Gowers W T and Nielsen M. (2009). Massively collaborative mathematics. *Nature*, 461, 879-881.

Van den Eynden V and Bishop L. (2014). Incentives and motivations for sharing research data, a researcher’s perspective. A Knowledge Exchange Report, available from knowledge-exchange.info/Default.aspx?ID=733

Meadows D H, Meadows D L, Randers J and Behrens W W. (1972). *The Limits to Growth*. Potomac Associate Books, ISBN 0876631650

Nature (2013). http://www.nature.com/news/first-proof-that-infinitely-many-prime-numbers-come-in-pairs-1.12989

Nowotny H, Scott P and Gibbons M. (2001). *Rethinking science: knowledge and the public in an age of uncertainty*. Polity Books, Cambridge. ISBN 0745626086

OECD. (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding, pp 1-23. OECD, Paris.

O’Reilly T. (2005). What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software. http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

O’Reilly T and Battelle J. (2009). Web Squared: Web 2.0 five years on. http://www.web2summit.com/web2009/public/schedule/detail/10194

RDA Europe (2014). The Data Harvest: how sharing research data can yield knowledge, jobs and growth, pp 1-38. https://europe.rd-alliance.org/documents/publications-reports/data-harvest-how-sharing-research-data-can-yield-knowledge-jobs-and

RECODE (2014). Institutional barriers and good practice solutions. http://recodeproject.eu/wp-content/uploads/2014/09/RECODE-D4.1-Institutional-barriers-FINAL.pdf

The 1000 Human Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526, 68-74.

Youngseek K and Stanton J M. (2012). Institutional and Individual Influences on Scientists’ Data Sharing Practices. *Journal of Computational Science Education*, 3, 47-56.

Zhang Y. (2014). Bounded gaps between primes. *Annals of Mathematics*, 179, 1121-1174.

- **Sources**

Name	Location	Email	Telephone
Dr Liz Allen, Head of Evaluation	Wellcome Trust, Gibbs Building, 215 Euston Road, LONDON NW1 2BE, UK	l.allen@wellcome.ac.uk	+44 20 7611 8426
Dr Fiona Armstrong, Deputy Director Policy, Resources and Communications	Economic & Social Research Council, Polaris House, SWINDON SN2 1UJ, UK	fiona.armstrong@esrc.ac.uk	+44 1793 413 065
Professor John Beath FRSE, Secretary-General, Royal Economic Society	School of Economics & Finance, University of St. Andrews, ST ANDREWS KY16 9AR	jab@st-andrews.ac.uk	+44 1334 462 421
Dr Lutz Bornmann, Analyst	Division for Science and Innovation Studies, Administrative Headquarters, Max Planck Society, MUNICH, Germany	bornmann@gv.mpg.de	+49 89 2108 1265
Professor Peter Buckle, President, Society of Ergonomics	Helen Hamlyn Centre for Design, Royal College of Art, Kensington Gore, LONDON SW7 2EU, UK	peter.buckle@rca.ac.uk	+44 20 7590 4242
Dr Tim Evans	Department of Physics, Imperial College, LONDON SW7 2AZ, UK	t.evans@imperial.ac.uk	+44 20 7594 7837
Professor Ghislaine Filliatreau, Director	Observatoire des Science et des Technique, HCERES, 20 rue Vivienne, 75002 PARIS, France	ghislaine.filliatreau@obs-ost.fr	+33 1 4439 0680
Professor Nils Petter Gleditsch	Peace Research Institute, OSLO, and Norwegian University of Science and Technology	nilspg@prio.no	+47 2254 7721
Dr Mark Hahnel, Founder and Director, Figshare	Figshare Ltd, 4 Crinan St., LONDON N1 9XW, UK	mark@figshare.com	
Dr Phil Heads, Associate Director of Strategy and Impact	Natural Environmental Research Council, SWINDON, UK	phhe@nerc.ac.uk	+44 1793 411699
Professor Daniel W Hook	Department of Physics, Washington University, St Louis, USA	dwh@wuphys.wustl.edu	+1 314 935 6216
Dr Wolfram Horstmann, Director	State and University Library, University of	horstmann@sub.uni-	+49 551 39 14494

and University Librarian	Göttingen, Platz der Göttinger Sieben 1, D-37070 GOTTINGEN, Germany	goettingen.de	
Professor Peter Hudson	School of Epidemiology & Infectious Disease, State University of Pennsylvania, HARRISBURG, USA	pjh18@psu.eu	+ 1 814 865 6057
Dr Daniel S Katz, Senior Fellow	Computation Institute, University of Chicago, 5735 S Ellis, CHICAGO I-60637 USA	d.katz@ieee.org	+ 1 773 834 7186
Professor Adrian McDonald, Director CREH	School of Geography, University of Leeds, LEEDS LS2 9JT UK	a.t.mcdonald@leeds.ac.uk	+44 113 34 33344
Mr David Mount, Director	Countryside Training Partnership, Ruskin Villa, EDALE, Derbyshire S33 2EZ, UK	david@countrysidetraining.co.uk	+44 1433 670300
Dr Ismael Rafols, Technology policy analyst	InGenio, CSIC-UPV, Universitat Politecnica de Valencia, Camino de Vera, 46022 VALENCIA, Spain	i.rafols@ingenio.upv.es	
Professor Gunnar Sivertsen	Nordisk institutt for studier av innovasjon, forskning og utdanning, Wergelandsveien 7, N-0167 OSLO, Norway	gunnar.sivertsen@nifu.no	+47 22 59 51 00
Dr Mark Whitton, former Director	GMAP, 7 Thornfield Road, LEEDS LS16 5AR, UK		
Professor Paul Wouters, Director	CWTS, Willem Einthoven Building, Wassenaarseweg 62A, University of Leiden, 2333 AL Leiden, The Netherlands	luijt@cwts.leidenuniv.nl	+31 71 527 3909

KI-NA-27-398-EN-N

This report focusses on practical issues relating to the emergence of ‘open science’ and researchers’ response to the opportunities created for collaboration and data development through innovative technology. It has been readily adopted by researchers because of pre-adaptation via many earlier changes. Future policy may be best aimed at further enabling such response rather than top-down action on assessment or funding.

Additional facilitative actions by the European Commission that would enable such community response include: recognition via assessable, transitive credit for data providers; recognition of the costs of making data both open and useful, particularly the vast numbers of small datasets; and balancing domestic innovation policies with trans-national programmes that maintain collaborative, open science across borders.

Studies and reports

Publications Office

ISBN 978-92-79-50444-0