research information network

NESTA

# Open to All?

## Case studies of openness in research

A joint RIN/NESTA report

September 2010

# Executive Summary


***Background***

Since the early 1990s, the open access movement has promoted the concept of openness in relation to scientific research. Focusing initially upon the records of science in the form of the text of articles in scholarly journals, interest has broadened in the last decade to include a much wider range of materials produced by researchers. At the same time, concepts of openness and access have also developed to include various kinds of use, by machines as well as humans.

Academic bodies, including funders and groups of researchers, have set out statements in support of various levels of openness in research. Such statements often focus upon two key dimensions: what is made open, and how; and to whom is it made open, and under what conditions? This study set out to consider the practice of six research groups from a range of disciplines in order to better understand how principles of openness are translated into practice.

***Method***

The study consists of interviews with 18 researchers working across 6 UK research institutions. The aim was to identify a range of practices, not to draw conclusions that could be generalised to an entire population. Research teams were therefore selected to represent not only convinced advocates of openness, but also individuals or groups which are more selective about what they share and perhaps more sceptical of the open agenda. Each team included a senior researcher at PI level along with some of their more junior colleagues. Interviews were structured to uncover researchers' levels of openness at various stages of the research lifecycle.

***Key findings***

*Benefits of openness*

Researchers identified several distinctive benefits to open behaviour in their work:

1. *increasing the efficiency of research*, for example by avoiding duplication of effort, by making research tools, protocols and examples of good practice more readily available, by reducing the costs of data collection, and by promoting the adoption of open standards.

2. *promoting scholarly rigour and enhancements to the quality of research,* for example by making information about working methods, protocols and data more readily available for peer review and scrutiny, and enhancing the scope and quality of the material published in the scholarly record, including negative results.

3. *enhancing visibility and scope for engagement,* with opportunities for wider engagements, across the research community and other, broader, communities, including new possibilities for 'citizen science' and for public engagement with the processes and results of research.

4. *enabling researchers to ask new research questions,* and to address questions in new ways through re-use of data and other material created by other researchers, supporting the development of 'data-intensive science' through the ability to aggregate and re-analyse data from a wide range of sources.

5. *enhancing collaboration and community-building,* for example by providing new opportunities for collaboration across institutional, national and disciplinary boundaries, and for the sharing of knowledge and expertise.

# Executive Summary

6. *increasing the economic and social impact of research,* innovation in business and public services, and the return on the public investment in research, by enabling individuals and organisations beyond the research community to engage with a wider range of research resources and materials.

*Barriers and constraints to openness*
Researchers also identified a set of issues which inhibit or discourage open working:

1. *lack of evidence of benefits and rewards.* Changes in practice involve costs to researchers in time and effort, and they may be unwilling to incur those costs unless they see clear benefits in terms of enhancing their ability to win more grants, secure greater recognition from their peers, and to advance their careers.

2. *lack of skills, time and other resources.* Developing, sustaining and making use of the new kinds of infrastructure required for openness demands new skills and significant effort from researchers and others, particularly at a point when standards, guidelines, conventions and services for managing and curating new kinds of material are as yet under-developed and not always easy to use.

3. *cultures of independence and competition.* Cultures vary in different disciplines, which can itself cause difficulties in cross-disciplinary work; but researchers are typically both co-operative and competitive. The key currency for securing competitive career rewards is publication of articles, conference papers and monographs; and many researchers regard the data and other resources that they create in the course of their research as their intellectual capital which they wish to exploit and mine in order to produce new publications over an extended period. Some researchers fear that openness involves a loss of control, and a risk of being scooped by others.

4. *concerns about quality.* Researchers attach great importance to peer review of research results. There are concerns, however, about how peer review operates in practice; and such concerns become more acute when researchers consider whether or how peer review or other quality assurance mechanisms might be extended to cover new kinds of material such as data. Many researchers also fear that data and other outputs may be misinterpreted or misapplied, particularly if meticulous attention is not paid to the provision of methodological and other key contextual information.

5. *ethical, legal and other restrictions on accessibility.* It may be impossible to make data and other information resources openly available because they are personal or confidential, or subject to commercial or third party licensing restrictions. Researchers themselves, or their funders, institutions or partners, may need to restrict access to their results and other materials in order to protect commercial confidentiality, or the potential for commercial exploitation. The relationships between such restrictions and the requirements of the Freedom of Information Act are not well-understood in the research community.

***Recommendations***

The key issue for policy-makers is how best to support individuals, groups and communities to work with the degree of openness which provides clear benefits to them. Single solutions will not work for all kinds and areas of research. With this in mind, research funders and institutions should work with research communities on the following areas:

1. *data management and sharing:* providing guidance and developing policies to support and promote better management and effective sharing of research data, including clear guidance on the requirements and implications of the Freedom of Information Act and the Environmental Information Regulations.

2. *research infrastructure:* supporting tools and standards which encourage open working, and providing incentives and rewards for those who contribute to the development of such resources.

3. *training and skills:* include training on data management and open working, including legal, ethical and regulatory issues, as part of doctoral programmes and continuing professional development for researchers.

4. *business models:* increasing awareness of open business models, developing business planning guidelines and toolkits to help researchers understand the risks and opportunities of working more openly.

5. *quality assurance and assessment:* providing guidance on how the peer review system might be adapted to address the growing numbers of resources which have not undergone pre-publication peer review.

6. *examples of good practice:* gathering, assembling and disseminating examples of good practice in open science, and ways in which these practices have benefitted both research projects and researchers themselves.

# Contents

# 1. Introduction

## 1.1 Background and Context

Concepts of openness have been with us for a long time. The *Oxford English Dictionary* provides examples from early medieval times onwards of the use of the word 'open' in the sense of providing unrestricted access to something, exposing it to general view or knowledge, or performing it without concealment, so that all may see or hear. Such attributes are generally seen as desirable; and for researchers and those who support their work the growth of the web and related technologies has added both impetus and new dimensions to concepts of openness as an essential underpinning of the public good that derives from research.

In 1992, the neurobiologist Steven Rose wrote of "trying to lay open my craft", and the 1990s saw the beginning of the open access movement, focusing initially on exploiting the potential of the web to promote unrestricted access – for anyone, anywhere in the world – to the records of science in the form of the texts of articles in scholarly journals. Over the past decade, the focus has broadened to encompass a much wider range of materials produced by researchers and others; and concepts of openness and access have broadened also to include various kinds of use, by machines as well as humans.

As researchers have taken up new technologies and services – with the promise of more to come – they have changed their behaviours and attitudes. Digital technologies are indeed transforming the nature of the research process itself, the types of research questions that can be asked, and the ways in which they can be addressed. As part of these developments, some groups of researchers, in the UK and across the world, have played leading roles along with information specialists in developing the ideas and principles associated with openness, and tools and services to help put them into effect. They have already had a significant effect on both policy and practice, with research funders and policy-makers showing increasing interest in openness as a means to increase the efficiency and enhance the impact of research.

But while the academy has for long been characterised – at least in part – by the exchange of ideas and findings in an essentially gift economy, we are currently at some distance from a world in which the processes and outputs of research are fully open to all. A relatively small number of individual researchers and research groups are active in promoting openness. A much larger number are sympathetic or even enthusiastic, but not always open in all their practices. Many other researchers are cautious, and see many barriers and constraints to overcome if a presumption in favour of openness is to become an everyday element of policy and practice.

Such issues have become much more urgent in the light of the recent problems at the Climate Research Unit at the University of East Anglia and the handling of requests for material under the Freedom of Information Act. For across the UK research community there is very little awareness or understanding of the Freedom of Information Act 2000 and the Environmental Information Regulations that came into effect in 2005. The *Independent Climate Change Emails Review* (Russell et al., 2010) noted that – to the extent that the research community is aware of the implications of the Act and the Regulations at all – there is much confusion and unease as to how the Act should be applied with respect to the materials developed during a research process. The Review recommends that 'all data, metadata and codes necessary to allow independent replication of results should be

provided concurrent with peer-reviewed publication'. But it goes on to state that the position with regard to supporting material such as early draft correspondence with colleagues and working documents is unclear, and recommends that the Information Commissioner's Office should hold consultations on these issues and consider what further guidance might be provided.

In this context it is important that policy should be based on a clear understanding of the working practices as well as the concerns of researchers operating in a range of contexts. This report thus presents the results of a series of case studies of research groups who are interested in concepts of openness and how they can be put into effect. It illustrates a range of views, from convinced advocates to sceptics, and a variety of practice. As a result, we are able to identify a number of key issues that will need to be addressed if openness is to become a fundamental underpinning of research policy and practice, in the UK and elsewhere.

## 1.2 Principles of openness

Some of the key principles of openness have been articulated in statements from research funders and various other bodies. Research Councils UK (RCUK) thus promulgated in 2006 the principle that

> ideas and knowledge derived from publicly-funded research must be made available and accessible for public use, interrogation, and scrutiny, as widely, rapidly and effectively as practicable (RCUK, 2006)

In seeking to define the principles more closely, the signatories to the *Berlin Declaration* (2003) prescribe that openness requires that users are able not only to access, but also to copy, distribute, and display material publicly; to print copies for private use; and to make and distribute derivative works. The *Principles for Open Science* published by Science Commons (2008) take such ideas a stage further, with specific requirements relating to publications, research tools, research data, and open cyberinfrastructure.

For data in particular, the OECD published in 2007 *Principles and Guidelines for Access to Research Data from Public Funding* (OECD, 2007), which highlights the principles that such data are a public good, produced in the public interest, and that they should be openly available to the maximum extent possible. The UK Research Councils have accepted these principles, and in some cases added to them: the BBSRC, for instance, points to the importance of the use of standards, of appropriate data quality and provenance, of timeliness, and of regulatory and ethical requirements. It also recognises that researchers have a legitimate interest in seeking to benefit from the time and effort they themselves have put into producing the data (BBSRC, 2010).

The *Panton Principles* (Murray-Rust et al., 2010) published in February 2010 draw on the still wider Open Knowledge Definition (Open Definition, n.d.) from the *Open Knowledge Foundation*, a pressure group that takes inspiration from the open source software and open access publishing movements. The key requirement here is that there should be no licence or other restriction on use of any kind, including commercial use. Like other statements of principle, however, there is a recognition that the scientific traditions of citation, attribution and acknowledgment should be respected and upheld. It is notable, however, that in these various statements and the guidance

offered by BBSRC there is no reference to the requirements of the Freedom of Information Act; nor did they arise in the course of our case studies.

## 1.3 Dimensions of openness

The various declarations and statements of principle bring with them differing assumptions as to what should be shared, with whom, when and how. It is therefore helpful to consider two key dimensions of openness.

First, there is the question of ***what kinds of material might be made open, at what stage in the research process, and how.*** The open access movement focused initially, as we have noted, on the texts of journal articles, which might be considered as a key end-product of research. More recently, policy-makers and others have focused much attention on the data collected or created by researchers during the course of their research, although it is not always clear whether it is raw data or data that have undergone some stages of refinement and analysis that are being considered. But researchers also produce many other kinds of material during the course of their research, including bibliographies, protocols, laboratory notes, software tools and so on, and many 'open science' advocates are now focusing on access to such materials. Many are also making use of social networking, blogs, wikis and other Web 2.0 tools as key means of communication during as well as at the end of the research process. Different kinds of material may be made available immediately or with varying levels of delay.

Second, there are questions as to the ***groups of people to whom the material is made open and on what terms or conditions.*** Researchers are often most comfortable about making information and other resources available to colleagues with whom they work, or whom they know, perhaps upon request. Making materials openly available to a wider research community, perhaps after requiring a registration process, represents a further stage in openness. Unrestricted access for anyone anywhere in the world takes openness to the ultimate stage. Terms and conditions may include technical issues relating to how effectively the material is exposed and described (and therefore findable); the formats in which it is presented (involving such questions as whether it may simply be read, or also used and manipulated); and the provision of contextual information (which may affect how effectively it can be understood or interpreted). Degrees of openness may also be affected by legal or licence restrictions on redistribution or re-use. Finally, legal, ethical or contractual considerations relating to confidentiality may preclude access altogether, except for tightly-controlled groups of people.

## 1.4 Benefits of openness

Apart from the virtues of openness as a value in itself, the benefits of openness in research have been described in a number of different ways (Fry et al, 2009; Lyon, 2009). It is convenient for the purposes of this study, however, to present them under six headings.

1. *increasing the efficiency of research*, for example by avoiding duplication of effort, by making research tools, protocols and examples of good practice more readily available, by reducing the costs of data collection, and by promoting the adoption of open standards.

2. *promoting scholarly rigour and enhancements to the quality of research,* for example by making information about working methods, protocols and data more readily available for peer review and scrutiny, and enhancing the scope and quality of the material published in the scholarly record, including negative results.

3. *enhancing visibility and scope for engagement,* with opportunities for wider engagements, across the research community and other, broader, communities, including new possibilities for 'citizen science' and for public engagement with the processes and results of research.

4. *enabling researchers to ask new research questions,* and to address questions in new ways through re-use of data and other material created by other researchers, supporting the development of 'data-intensive science' through the ability to aggregate and re-analyse data from a wide range of sources.

5. *enhancing collaboration and community-building,* for example by providing new opportunities for collaboration across institutional, national and disciplinary boundaries, and for the sharing of knowledge and expertise.

6. *increasing the economic and social impact of research,* innovation in business and public services, and the return on the public investment in research, by enabling individuals and organisations beyond the research community to engage with a wider range of research resources and materials.

## 1.5 Incentives, barriers and constraints

Some, but not all, of the kinds of benefits outlined above accrue to researchers themselves. Thus they may be motivated by the prospects of greater visibility for their research, or of greater opportunities for collaboration. Other motives may also come into play, such as altruism and reciprocity, or encouragement from colleagues. But for many researchers the incentives and benefits are not yet obvious or strong; and if they are asked to consider making significant changes towards openness in their practice, researchers often point to a number of barriers and constraints.

1. *lack of evidence of benefits and rewards.* Changes in practice involve costs to researchers in time and effort, and they may be unwilling to incur those costs unless they see clear benefits in terms of enhancing their ability to win more grants, secure greater recognition from their peers, and to advance their careers.

2. *lack of skills, time and other resources.* Developing, sustaining and making use of the new kinds of infrastructure required for openness demands new skills and significant effort from researchers and others, particularly at a point when standards, guidelines, conventions and

services for managing and curating new kinds of material are as yet under-developed and not always easy to use.

3. *cultures of independence and competition*. Cultures vary in different disciplines, which can itself cause difficulties in cross-disciplinary work; but researchers are typically both co-operative and competitive. The key currency for securing competitive career rewards is publication of articles, conference papers and monographs; and many researchers regard the data and other resources that they create in the course of their research as their intellectual capital which they wish to exploit and mine in order to produce new publications over an extended period. Some researchers fear that openness involves a loss of control, and a risk of being scooped by others.

4. *concerns about quality.* Researchers attach great importance to peer review of research results. There are concerns, however, about how peer review operates in practice; and such concerns become more acute when researchers consider whether or how peer review or other quality assurance mechanisms might be extended to cover new kinds of material such as data. Many researchers also fear that data and other outputs may be misinterpreted or misapplied, particularly if meticulous attention is not paid to the provision of methodological and other key contextual information.

5. *ethical, legal and other restrictions on accessibility.* It may be impossible to make data and other information resources openly available because they are personal or confidential, or subject to commercial or third party licensing restrictions. Researchers themselves, or their funders, institutions or partners, may need to restrict access to their results and other materials in order to protect commercial confidentiality, or the potential for commercial exploitation. The relationships between such restrictions and the requirements of the Freedom of Information Act are not well-understood in the research community.

# 2. Approach and methodology

This report is based on an extensive literature review and six case studies of researchers from a range of research domains. The case studies were based on semi-structured interviews with a total of 18 researchers working across six institutions: the Universities of Edinburgh, Oxford, Southampton, Leicester, and Central Lancashire; and the Science and Technology Facilities Council (STFC) Rutherford Appleton Laboratories. Each of the case studies involved at least one established researcher at Principal Investigator (PI) level, along with others working earlier in their careers.

The case studies were selected to include research teams which involve advocates and proponents of open research ('the convinced') as well as teams which are releasing data or resources on a more selective basis, with some members known to be sceptical about the open agenda. We thus sought a range of views and of practice with regard to open working in the UK research community.

The interviews were transcribed and analysed to identify commonalities and differences in the arguments and evidence that participants presented to us. We complemented this with desk research to review publications of the case study groups, including the resources each group currently makes publicly accessible online. We believe we have accurately represented our participants' views and practices. But these should not be taken as necessarily representative of the members of the research groups or projects as a whole, still less of the disciplines or research communities of which they are a part. This short exploratory study seeks only to identify a range of positions and practices, rather than to discern how widely these are shared. Nevertheless, we hope it provides insights into some current changes in practice, and their implications for researchers, institutions and funders.

## 2.1 Open working across the research lifecycle

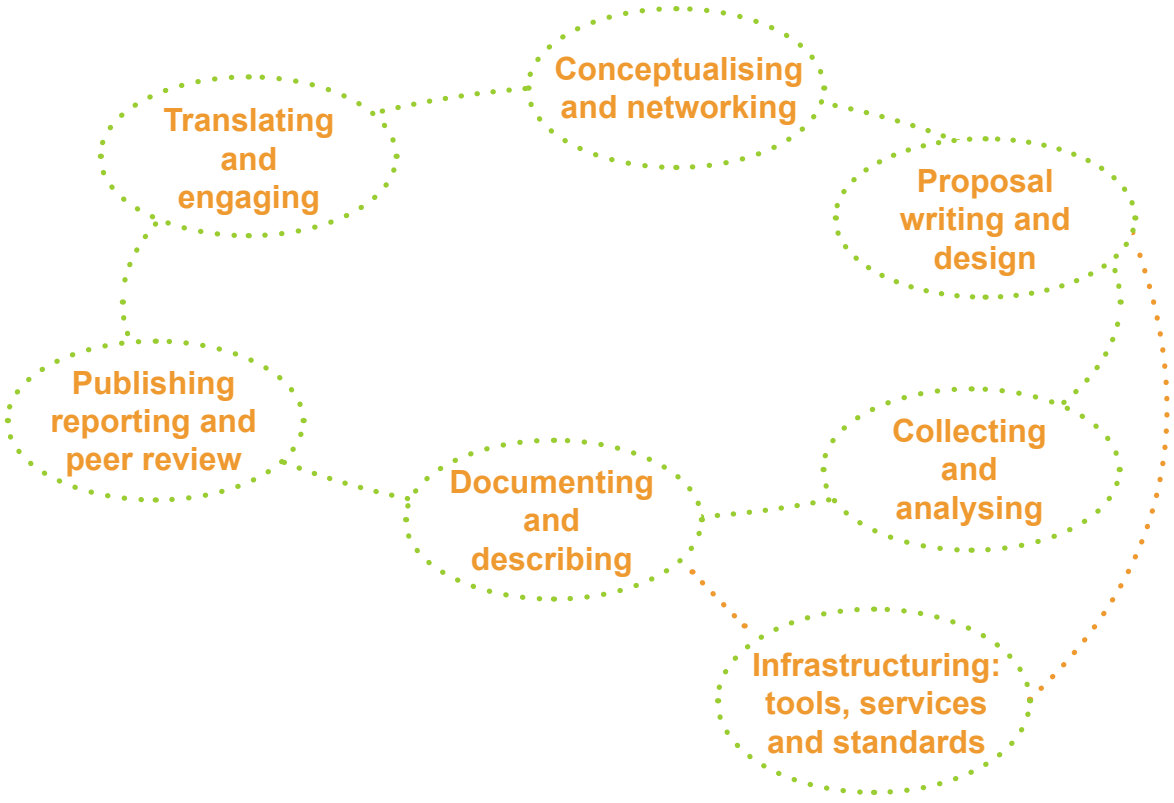
In order to investigate the range of open working in each case study, we made use of a model of the research lifecycle adapted from the JISC 'Research 3.0' lifecycle (JISC, 2009) with elements from Charles Humphrey's 'knowledge transfer lifecycle' (Humphrey, 2006). The key stages in this model are summarised in Figure 1 and comprise

- *conceptualising and networking:* sharing new research concepts and discussing possible areas for collaboration;
- *proposal writing and design:* sharing and discussing drafts of proposals and designs, including communication with funders or institutional bodies on regulatory compliance issues;
- *collecting and analysing:* performing and documenting the collection and analysis of data or other research materials;
- *infrastructuring:* contributing to development of community standards, tools and databases or other shared resources;
- *documenting and describing:* completing and reviewing final documentation or structured metadata prior to publication or submission to an archive or repository;

- *publishing, reporting and peer review:* preparing and communicating articles, reports, or other products of the research. Taking part in peer reviews of third party research outputs, or informally commenting and rating these;

- *translating and engaging:* involving the envisaged users of the research in actual or potential applications of it, in other research fields, commercialisation or policy.

Figure 1. Research lifecycle



This approach also enables us to consider some of the key dimensions of openness as summarised in section 1.3 above: what is made open, to whom, and under what conditions. The kinds of material that may be made open at the various stages in the lifecycle are summarised in Table 1.

Table 1. Research lifecycle stages and material outputs

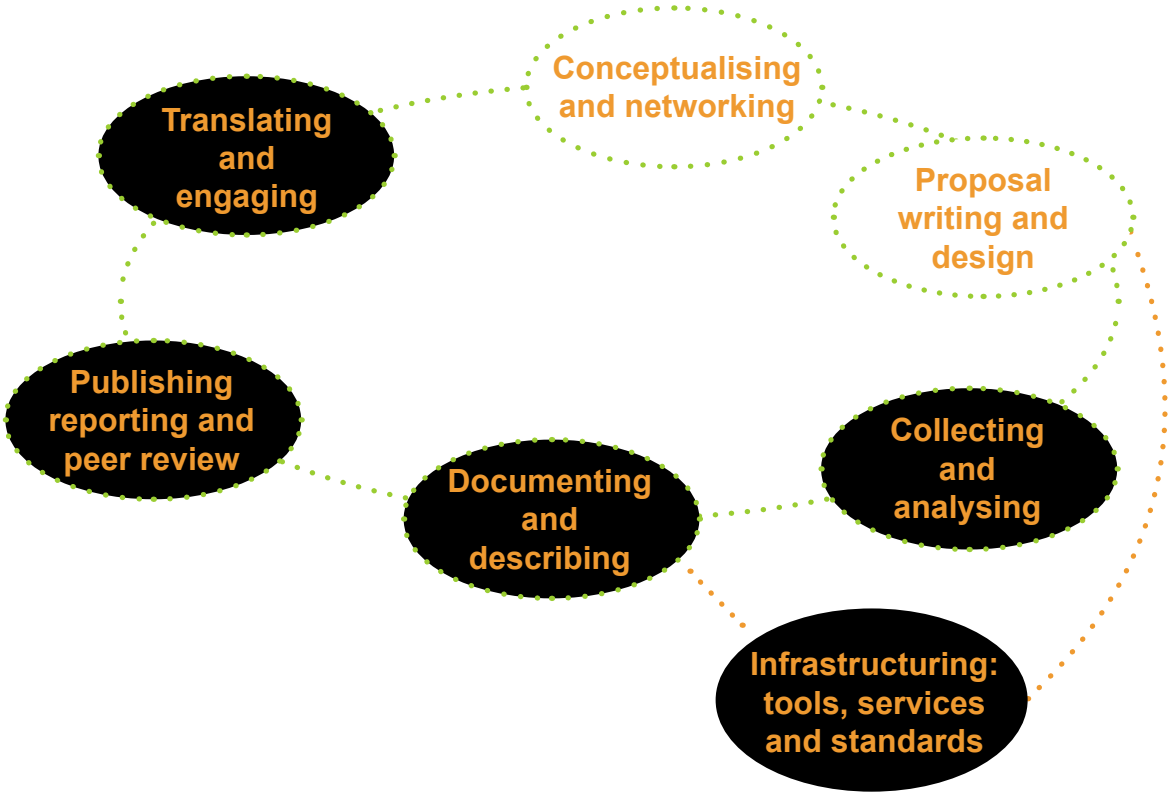
| **Research cycle stage** | **Outputs** |
|---|---|
| Conceptualising and networking | Messages, posts, user profiles, bibliographies, resumes |
| Proposal writing and design | Proposal drafts, data management plans, regulatory compliance documentation, study protocols |
| Conducting and presenting | Raw and derived data, metadata, presentations, podcasts, posters, workshop papers |
| Documenting and sharing | Lab notes, research memos, study-level meta-data, supplementary information |
| Publishing and reporting | Conference papers, journal articles, technical reports |
| Engaging and translating | General articles, web pages, briefings, public exhibits, presentations |
| Infrastructuring | Software tools, databases, repositories, web services, schemas and standards |

# 3. The case studies

Each case study highlights participants' practices at key stages of the research lifecycle, with some reference to practice in the relevant wider research community. The focus here is on behaviour as distinct from attitudes, which will be considered further in sections 4 and 5.

Each case study begins with a diagram highlighting those parts of the research lifecycle which formed the focus of that particular study.

## Case 1: Astronomy Virtual Observatory




### Participants

- Sylvia Dalla, Senior Lecturer, Jeremiah Horrocks Institute, University of Central Lancashire
- Andy Lawrence, Regius Professor, Institute for Astronomy, Edinburgh
- Jonathan Tedds, Senior Research Fellow, University of Leicester

Open data, tools and standards are well established in astronomy as building blocks of the 'virtual observatory' (VO). Participants in the case study are linked through active involvement in the Astrogrid project, the UK contribution to the International Virtual Observatory Alliance (IVOA). Virtual observatories are a means for astronomers to gain access to and analyse data from a federation of data centres.

***Collecting and analysing***

Research in astronomy is underpinned by observations of astronomical objects. Raw data are captured by scanning photographic plates or digital detectors recording objects or portions of the sky. Astronomers compete for observation time at facilities around the world. Raw data are transmitted to data centres which curate the data and make them available through web-based catalogue services. Checking, characterising and managing the increasing volumes of images collected have become major issues in astronomy.

***Infrastructuring***

Developing an infrastructure of software and standards is an essential underpinning for astronomical research, and openness has been a key characteristic of that work. Astrogrid and the IVOA support this through the development of standardised data formats, analysis tools, resources, and registries that identify where these resources are located. Thousands of standardised resources have become available through VO registries, the 'yellow pages' of VO research. But usability is critical if a resource is to be truly open, so that it is:

> "…easy to understand, with tools that go with [it]…it's all calibrated, it's documented. I think in practice those things are much more important than simply whether something is legally accessible or not…if things are set up so it's push button easy, then it really is open."

***Documenting and describing***

Metadata is standardised around the 30-year old FITS standard, and the more recent XML-based IVOA format VOTable. Both are open standards, maintained through open community efforts. Astrogrid and the IVOA use wikis to communicate the continuing development of VO infrastructure; but they do not use 'open notebook' applications and there is some resistance to complete openness:

> "They require a login just so that people within the project feel free to write everything down warts and all in progress. You know, some people would argue that everything should be open always, but certainly I know in practice that people will be encouraged to do that if they have a secure area."

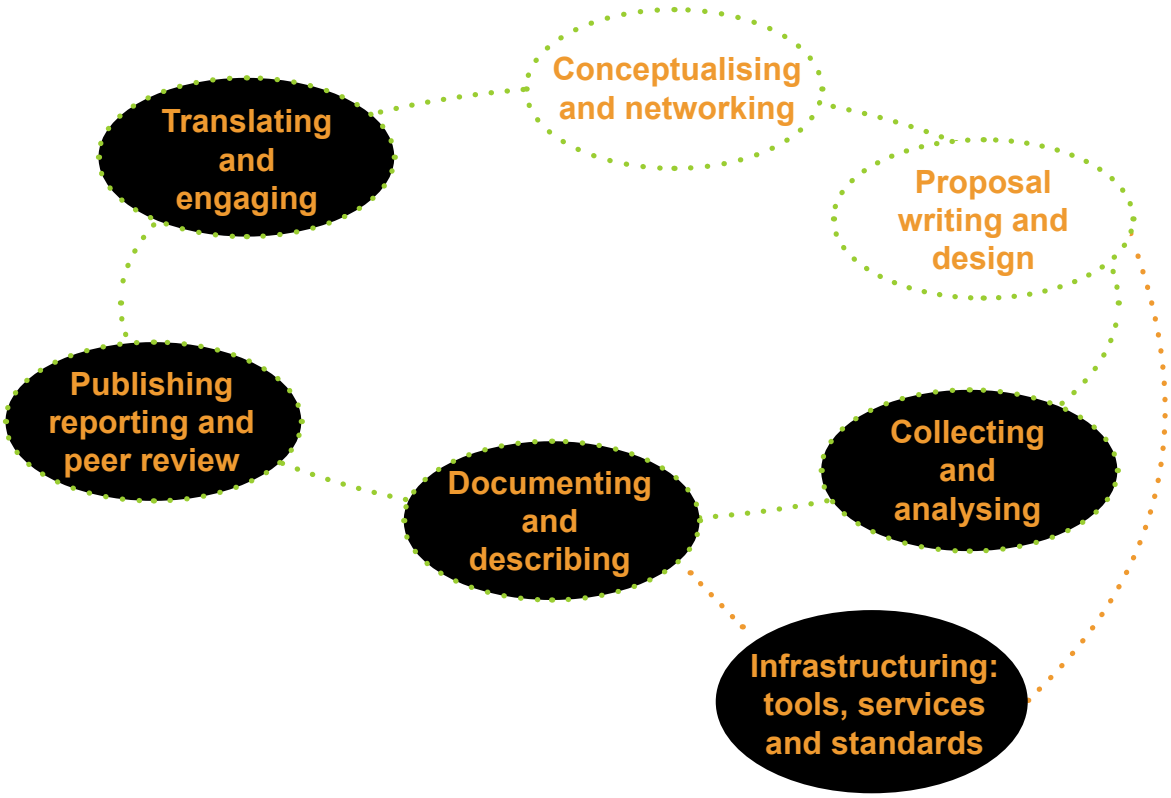
***Publishing, reporting and peer review***

It is a condition of access to large facilities that raw data should be deposited in a data centre, and the catalogue tables detailing what has been measured made available. Access to data may be embargoed, however, for up to a year. Funders and publishers do not require deposit of derived data, referred to as 'science products', and these are less commonly made publicly accessible. A recent development that aids reuse of catalogue data is the growing acceptance by journals of 'catalogue

papers'. Rather than describing the secondary use of catalogues these detail the development of catalogue resources.

***Translating and engaging***
Since astronomy research data generally has few commercial applications, the IPR constraints that affect other disciplines do not normally apply in astronomy. But skills and techniques in areas including image analysis may be applied in other disciplines and to professional and commercial practice in, for example, medical domains.

## Case 2: Image Bioinformatics Research Group




### Participants

- Graham Klyne, Research Fellow, Image Bio-informatics Research Group, Oxford University
- David Shotton, Reader in Image Bioinformatics, Dept of Zoology
- Jun Zhao, Research Assistant, Image Bio-informatics Research Group

The Image Bioinformatics Research Group's (IBRG) work focuses on image-related data, and contributes more generally to the infrastructure for open working in bioinformatics.

***Infrastructuring***

Functional genomics is the study of how an organism's genome is expressed physiologically. It uses and re-uses microarray data identifying the 'expression level' of the cellular mRNA molecules synthesised by particular genes, typically measuring many thousands of genes from a sample. The metadata standards and tools to record microarray experiments, and public databases for their deposit, curation and publication are well-established. The European Bioinformatics Institute (EBI), an outstation of the European Molecular Biology Laboratory (EMBL), is a key part of the infrastructure, providing the ArrayExpress public repository.

IBRG work has focused on the potential, alongside such large centralised resources, of a wide range of datasets emanating from small-scale projects.

> "The majority of biologists don't operate in [large-scale genomics and bioinformatics]. They are the long tail of small research groups doing observations on field voles and mink along the river Thames, and their data…tend to be kept private and not much published in the open. And those are the researchers we're hoping to work with to encourage them to publish their data."

The recent FlyData project exemplifies the Group's approach of working in collaboration with small numbers of data providers and users – in this case of data from fruit fly (Drospohila) gene expression experiments. The project used semantic web technologies and standards (RDF and SPARQL) to develop the open source OpenFlyData 'linked data' application, aiming to demonstrate how researchers may be supported to develop interfaces to search across disparate sources more cost-effectively (Zhao et al., 2009).

The IBRG approach is driven by the notion of 'sheer curation' and the provision of easily-usable web-based visualisation tools:

> "You try and make it easy for the researchers but also offer them immediate benefits for their local data management so that the archiving and publication of the selected data sets can come off the back of that with not too much additional effort."

> "If they get immediate benefits by submitting the appropriate metadata, just enough information to do the visualisation, we can capture that metadata and use it to support the subsequent publication of data."

The Group is also seeking to build open communities to develop ontologies, with a particular focus on an ontology for citations, CiTO, which enables semantic mark-up of references in research articles, characterising them by descriptors such as 'confirms' or 'refutes'. Once the citations are recorded in machine-readable form they can be used to construct citation networks and to undertake bibliographic analysis, provided that reference lists are openly accessible.

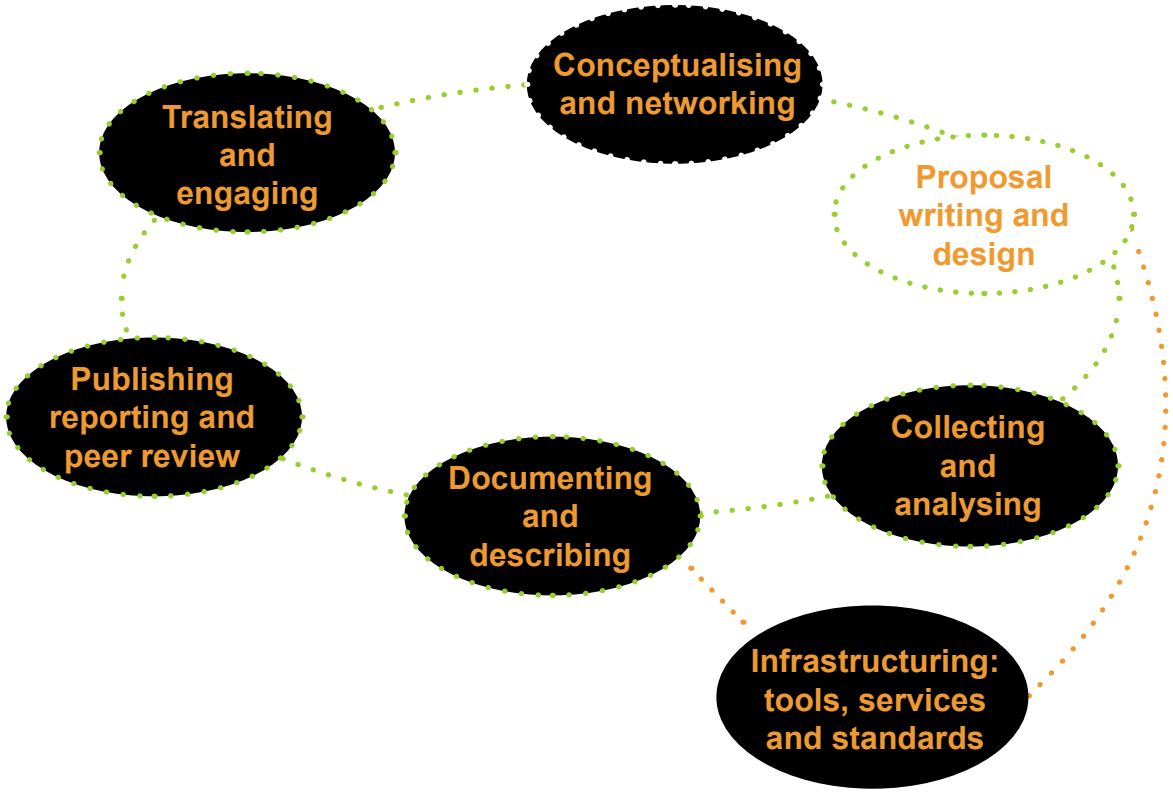
***Documenting and describing***
While IBRG collaborates on metadata documentation tools, it does not generate biological data of its own. The Group does, however, openly document and describe progress on its collaborations, via its Imageweb wiki. This is contributed to by team members, and is accessible – but not editable – without restriction.

***Translating and engaging***
The OpenFlyData application has been adapted in several cross-disciplinary collaborations beyond the functional genomics of fruit flies. One, Linking Open Drug Data, integrates published databases of drug information and traditional Chinese medicine, and won the 2009 Triplification Challenge:

> "We published a selection of medical and drug related open access databases in RDF format and…it shows there's real benefits. People can reuse their datasets and try to do some added value research using these datasets to do some drug discovery or chemical compound interaction analysis."

***Proposal writing and design***
Project documentation on the IBRG ImageWeb wiki includes details of proposal designs and of early meetings with collaborators. Editing the pages requires a login, but they provide a transparent public record of discussions before proposals are submitted to funders.

## Case 3: Open Notebook Science in Chemistry and Chemical Biology




### Participants

- Cameron Neylon, Senior Scientist, ISIS Group, Rutherford Appleton Laboratory
- Luke Clifton, Scientist, ISIS Group, Rutherford Appleton Laboratory
- Simon Coles, Manager, National Crystallography Service, Dept of Chemistry University of Southampton
- Jeremy Frey, Professor of Physical Chemistry, Dept of Chemistry University of Southampton

Participants in this case include prominent advocates of open science, and open working features at almost all stages in their work. The group works in chemical crystallography, physical chemistry, and the chemistry of biological materials; and they are based at Southampton and at the Rutherford Appleton Laboratory. The common thread linking their current work is use of the LabBlog electronic laboratory notebook, though they also use other blogging platforms and social media.

The format of online laboratory notebooks, and their varying degrees of openness, have evolved since the LabBlog first emerged from the Southampton Chemistry Department. Motivations and incentives include easy communication with colleagues across institutional boundaries, and the need for a comprehensive record of experimental processes and results, within which data may be linked and semantically enhanced, perhaps through integration of lab notebooks with open data repositories. For some, but not all, of the participants, the possibilities of better visibility are also important:

> "It really was a matter of sort of getting everything I did out, available, and the knock on advantage of that is that you get increased visibility in terms of the amount of datasets that you manage to publish."

***Conceptualising and networking***
Some of the participants have used social networking tools to seek collaboration, issuing open requests for help to addressing specific problems:

> "I found a person to solve a specific small scale problem rapidly, effectively, better than I could have done."

***Proposal writing and design***
Members of the group have also been involved in open discussion of possible proposals in response to funders' calls, using the FriendFeed platform. The blog-oriented lab book approach is used more at the post-funding experimental design stage, including measures to comply with the COSHH laboratory safety regulations, which are 'open to inspection' in the event of an investigation into an accident. 'Openness' here is not just about access for collaborators but also accountability to regulators.

***Collecting and analysing***
All participants use the blog form of Open Notebook as a day-to-day record of laboratory work, within which data can be linked. LabBlog applications have increasingly involved automatic collection of data and metadata from lab instruments ('instrument blogging'), and semantic enhancement of the research record for improved retrieval. This entails tagging blog entries with descriptors of the experimental process and chemical compounds in use, which the tool encodes in RDF syntax to aid later retrieval. The group continues to develop the tool to enable further semantic enhancement of the research record, motivated by the prospect of more effective exchange of data and metadata between laboratories.

***Infrastructuring***
The form of the lab blog continues to take shape as infrastructure for Open Notebook Science, the study participants acting variously as leaders in its development or as early adopters. Their 'infrastructuring' work to enable lab notebooks to integrate data across the research lifecycle relies both on generic semantic web and specific chemistry standards. Key among these have been standards for Chemical Markup (CML) of results records, and the InChI International Chemical Identifier for uniquely identifying chemical structures. These enable more effective linking between lab notebooks, and repositories of both publications and data.

These links have been explored in a range of projects involving the National Crystallography Service (NCS). Crystallography data are highly structured, with a well-established standard representation in the Chemical Information Format (CIF). There is also a well-established culture of making data available through repositories such as the Crystal Structure Database (CSD), located at the Cambridge Crystallographic Data Centre (CCDC), which also harvests from subscription-based journals. The eCrystals repository at Southampton, integrated with NCS, departs from this model by offering open access to pre-publication data (University of Southampton, n.d).

***Documenting and describing***

The wiki form of research record is potentially useful for writing up experiments at an intermediate stage between the day-to-day journal of the LabBlog, and the draft article. Some researchers conducting long-running experiments may also find a wiki a particularly useful way of tracking the changes in experiments through successive versions. Wikis can also be useful for open research protocols, providing 'how to' knowledge that early-career researchers find invaluable:

> "Really basic stuff, like doing gels and so on…I found out how to do that by looking at online protocols which were much easier…I mean papers just wouldn't tell you."

***Publishing, reporting and peer review***

The eCrystals repository extends the provenance chain recorded in the LabBlog by tracking the inputs to and outputs from the crystallography process, enabling links to be maintained between raw and processed data and publications. The concept of the LabBlog as data provenance record, effectively the back office in a production line linking the laboratory, repository and academic journal, is what the group refers to as 'publication@source'.

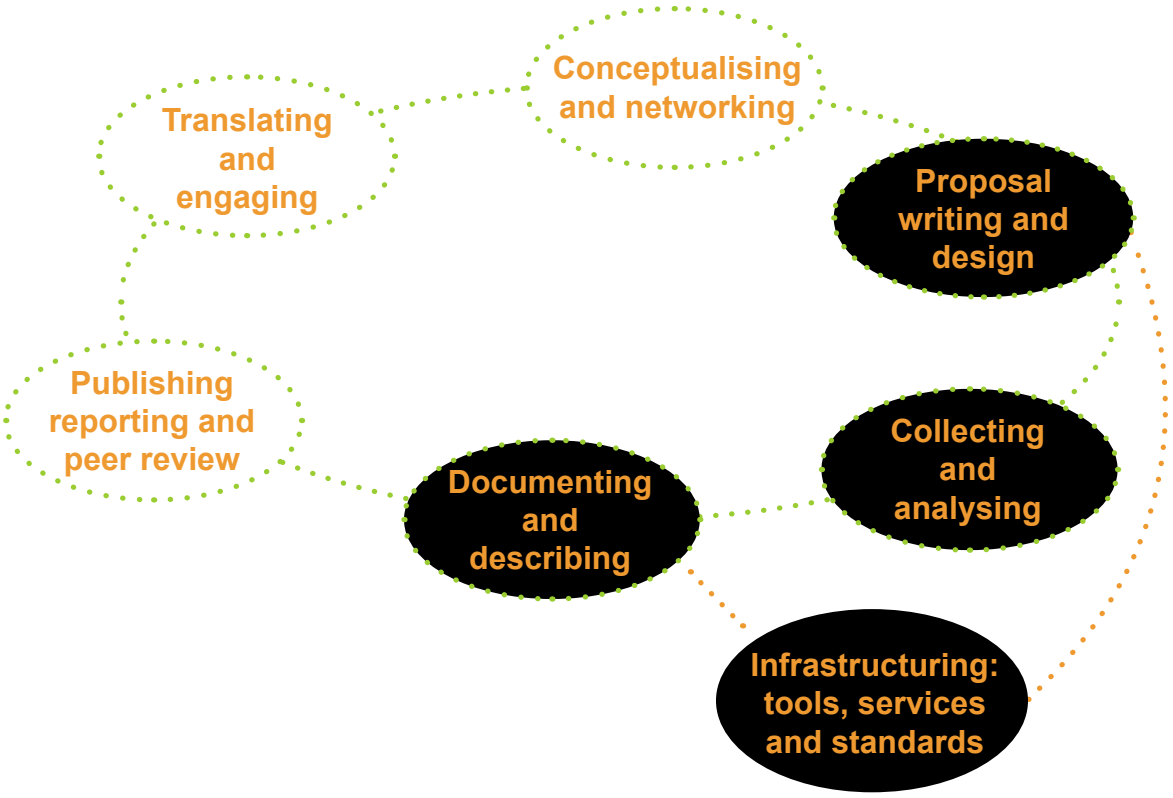
While the motivation for these developments is to maximise the re-use and analysis of data, there are shared concerns about quality assurance. The participants have direct experience of the difficulties entailed in peer review of data, and the stresses placed on traditional peer review processes when the volumes of data being generated far exceed the capacity of the peer review system to cope.

***Translating and engaging***

The LabBlog is being applied in a number of interdisciplinary projects, as well as in commercial collaborations. Both may entail limits on openness. An emphasis on the discoverability and transparency of the research record, and on integrating data in an actionable form with this record may thus not necessarily translate into a record that is open immediately and to everyone.

Degrees of openness, and enthusiasm for it, vary among the participants in this case study, who work openly with differing degrees and motivations. Some are accustomed to working in conditions of secrecy with commercial firms. Others are disdainful of the limits on disclosure involved in the patent process, though they may acknowledge the need for patent protection in order to meet the costs of a clinical trial process. All of them believe that notebook users must be able easily to control what is made open and when.

## Case 4: Clinical Neuroimaging Research Group




### Participants

- Participants preferred to be identified by pseudonyms, their job titles are as follows: Senior Research Fellow, Professor of Medical Imaging and Research Assistant

Neuroimaging research employs medical images and related datasets to investigate causes and correlates of disease, in this case psychiatric conditions. It draws on bioinformatic disciplines including genomics and neuroinformatics.

Participants in this case work together on projects using data derived from magnetic resonance imaging (MRI and fMRI) scans of the brains of psychiatric patients, their families, and control groups from the broader population. They seek to measure individuals' brain states at particular points in time, and in relation to their medical history and other factors, including genetic makeup and performance in psychiatric tests. Their work may be open to a degree; but sensitive personal data are released under strict conditions, and usually restricted to collaborators in their research community. Projects funded by pharmaceutical companies are not open at all since the data are owned by the companies concerned.

For these reasons, the group tends to be sceptical of the benefits of open working, and its members preferred to be identified by pseudonyms. The group has participated in various collaborative projects and, like others in this field, recognizes a strong case for integrating its datasets with those of other groups collecting similar data, since such integration can yield more detailed and powerful analyses. The group is involved in various multi-centre studies of this kind which pool datasets either prospectively (where subjects are recruited for a joint purpose) or retrospectively (where existing datasets are pooled and a more 'data driven' approach taken to the analysis). Such pooling requires ethical approval.

Although such studies are costly, recent examples have demonstrated their value; integrating imaging and genetic datasets studies has allowed researchers, for example, to explore specific ways that genes associated with schizophrenia are manifested clinically. Nevertheless, open publication of data or analysis software is rare, despite some US examples of neuroimaging repositories. The group does publish metadata on some of its larger datasets, however, as well as publishing its methods and results under open access terms, as is required by major funding bodies including the Wellcome Trust and the Medical Research Council.

***Proposal writing and design***
The early stages of research design tend to be closed, although research proposals are subject to review by ethics committees, which require details of how research subjects' confidentiality will be maintained, and how data will be shared and disposed.

***Collecting and analysing***
Recruiting subjects tends to be difficult, especially where these are patients suffering from uncommon conditions, and 'research fatigue' is common. Longitudinal studies, which are necessary to understand how some illnesses develop over many years, also require tightly-controlled procedures to ensure that follow-up contact has explicit consent and is renewed as necessary. A collaborative approach to recruiting subjects is therefore attractive, and the group operates an online research register which individuals can use if they are interested in participating in a study:

> "Researchers can see it's a collaborative effort and it's not just benefiting them it's benefiting other researchers as well…being able to have a wider net will give you a more representative sample of the population."

The group analyses the brain images alongside clinical and genetic data. They thus seek to identify correlations between neurological structures and functioning as portrayed in the images, and other information on subjects' physiology, behaviour, medical and social history. While the results are openly published along with description of the methods used, derived data are not made publicly available in a downloadable and actionable form.

***Infrastructuring***
Neuroimaging studies depend on effective identification and classification of the data they collect; and constant innovation in image analysis software. Comparison between studies also depends on

standardisation of terminology and tools for describing physiological, clinical, demographic and genetic changes. The group has collaborated in a number of projects to these ends.

Until very recently, for example, differences between scanners have been regarded as a major barrier to pooling datasets. But recent studies have shown these differences to be less significant than those between subjects, and this has given new impetus to the development of tools to harmonise datasets, and thus to facilitate collaboration. Code is therefore shared among group members and beyond in what is effectively the group's social network, on what they refer to as a 'warm share' basis.

> "Our best option is to do a share with another group. Several people have worked with us and seen in great detail what we do and they have taken the scripts that we use away with them. And we're continuing to work with these groups in a shared environment which is very warm because they're people we have worked with and they understand our thinking…[But] there is no way that we can get funding for open source release. We would need a software engineer here who was in effect rewriting the code."

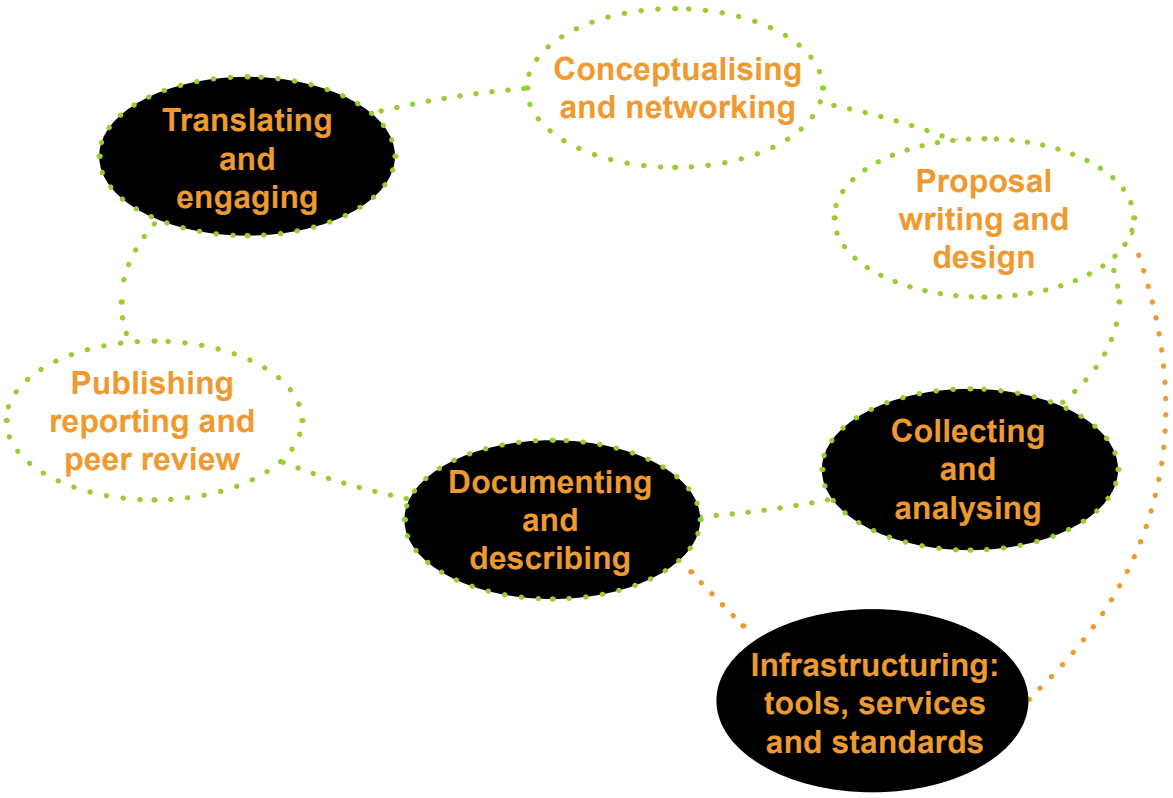
Open source development has made some inroads in the neuroimaging community generally. The SPM software for statistical parametric mapping of the brain has:

> "… been heavily backed by the Wellcome Trust and as a result it's become the global standard and the people who devised it have as a result the money .. the time and the inclination to sustain it, develop it, provide almost instantaneous answers to your questions about how to run it."

***Documenting and describing***

The group publishes metadata about its datasets, for example in a web accessible database of population studies relevant to mental health research funded by the Medical Research Council. Provenance metadata, however, is not generally shared.

## Case 5: Language Technology Group




### Participants

- Jean Carletta, Senior Research Fellow, Language Technology Group, School of Informatics, University of Edinburgh
- Theresa Wilson, Research Fellow, LTG
- Amy Isard, Research Assistant, LTG

Language technology or natural language processing (NLP) is a field encompassing computer science applications in speech technology and computational linguistics. The University of Edinburgh group is one of the largest in Europe. Open aspects of its work include publication of corpora (collections of examples of written and spoken language); contributions to standards and analytical tools; and corpus documentation and metadata. Corpora and tools have both cross-disciplinary and commercial applications.

Research questions typically involve both technological and linguistic features. The field is 'data driven' in that its focus is on automated processing of language, and its results are typically based around corpora.

***Collecting and analysing***

Open access to language resources is becoming critical to language technology. The research community strongly favours open source release of analytical tools, and is now encouraging open release of language resources of all kinds. Given the high costs of developing them, however, not all resources are free at the point of use; some bodies, such as the Linguistic Data Consortium, adopt a licensing and subscription model.

Projects are commonly undertaken by consortia working with resources – databases, corpora and tools – with different levels of access and permissions for re-use. Despite the moves towards openness, resources may not be open before results are published:

> "When people use closed things it's usually because they created it themselves and they're worried somebody else will beat them to publishing good research results if they let other people have it."

The group has led or participated in notable exceptions to this, recently including a15-member consortium which has published the *AMI Meetings Corpus*, providing streamed video recordings of meetings, along with transcripts, annotations and metadata, all of which are freely available on a Creative Commons share-alike licence.

Corpora of 'naturally occurring' human dialogue and interactions – whether in the form of text, or audio-visual and computer-based recordings – are created in order to analyse their syntax, semantics or pragmatics ('meaning in context'). Analysis involves annotating the corpora, which acquire more value as annotations are added over time. This provides a compelling motivation for maximising access to them:

> " If somebody generates data and then only they can publish…it closes who can publish to people who have the money to collect the data, which is expensive…And one of the reasons why we put out our tools and our data is so that anybody can play the game and it's not just limited to friends of…[researchers in] rich universities."

***Infrastructuring***

Language technology research depends increasingly on annotated corpora that are machine-readable and interoperable. Leading researchers thus both use and contribute to open standards for web content and structure including, for example, the XML standard, the Text Encoding Initiative and language ontologies.

Tools that are free make a huge difference to researchers:

> "You can imagine if every single group had to go and develop their own tokeniser, sentence splitter or parser, power speech taggers and what not…that would just be crazy. So there are several websites that list tools that are freely available to do these things."

> "If it's not open, you have to pay for it, and then you can't afford it. So a very large advantage is it's free."

Tools and standards may be easier to make open, however, than databases and corpora:

> "The data's not free...the curators have given it to us and we're allowed to run the software and generate text that uses the information, but we're not allowed to give it away…But then we've also built a language ontology, and that's already been given away…all just open access."

***Documenting and describing***

The group uses wikis for internal communication between consortia members, and then for public documentation purposes, but 'open notebook' publication of work-in-progress is not the norm:

> "People don't put their notebooks on a public blog or wiki for people to read until they've published the paper."

The group publishes annotations and metadata, often created using tools that they also publish openly, and with the extensive documentation required for open release. The requirement for good documentation also affects use of third-party resources:

> "There always is the question about how good is it and also how easy is it to use. I mean if there's no documentation, if it's really hard to figure out, quite often I will go hunting for another tool."

***Translating and engaging***

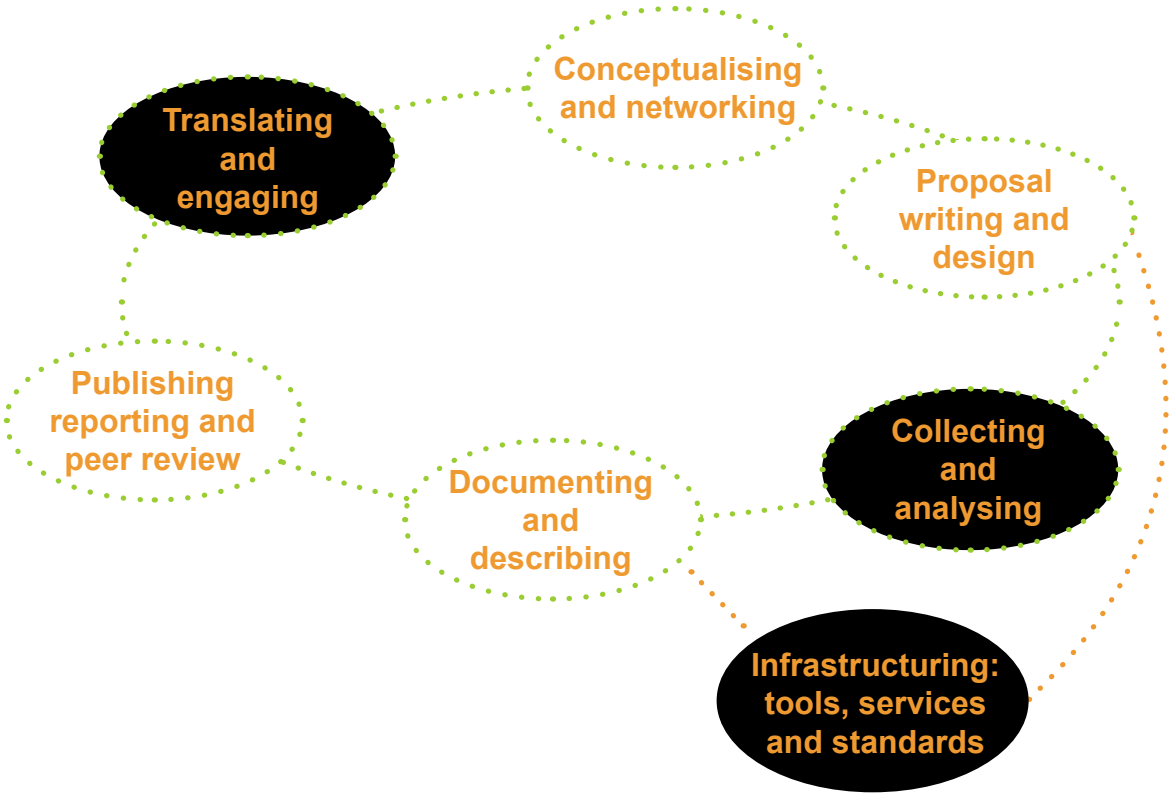
Since language technology is an interdisciplinary field, the group see open release as helpful in developing applications of their resources in other disciplines:

> "In the first instance language technologists use them but we usually try to distribute them in a way that other people can use them as well."

For example the NITE XML Toolkit, developed primarily for computational linguists, is also being used in the AMI corpus to meet the needs of organizational and social psychologists.

Language technology also has strong commercial take-up, for example in machine translation and speech recognition. The group also works in newer areas, such as 'sentiment analysis', that have high commercial potential. Open working has mixed implications here. On the one hand, collaborations involving the use of open tools can be attractive to commercial partners. On the other hand, such collaborations often involve recorded dialogue with firms' customers or relating to confidential business and so the data can never be openly published. The antipathy of commercial users to share-alike conditions sometimes leads to a 'multi-licensing' approach: the resource is made public with a Creative Commons licence, but also released under a closed licence to commercial partners.

## Case 6: Epidemiology Research Group




### Participants

- Eric Fevre, Senior Research Fellow, Epidemiology Research Group, University of Edinburgh
- Brajendra Singh, Research Fellow, Epidemiology Research Group, University of Edinburgh

The participants in this group are carrying out epidemiological studies of animal infections transmittable to humans. Open data and tools are relevant to their work in several ways; firstly as re-users of data published by government agencies, and secondly as contributors to open resources used in their own and other fields.

***Collecting and analysing***

The group's work involves collecting data and re-using data from government and other sources to analyse the prevalence of disease. There has been relatively little debate in the epidemiological research community about the scope or mechanisms for sharing data, whether in 'raw' form or as documented research products (Samet, 2009). However, there is significant overlap in policy and practice with the larger health and social science communities, since epidemiological studies typically use and contribute to large, professionally curated, cohort studies and national survey datasets (RIN, 2008).

Epidemiology is becoming more data intensive and data driven, as a result of enormous growth in the use of technology in data collection. In field studies, for example, the open source release of the EpiSurveyor tool established the use of PDAs to collect survey data electronically, dispensing with paper completely; and the rapid growth in online spatial data means that the group can draw on GIS map layers, GPS location data and images from satellites sourced from government agencies in the UK and Africa.

Data analysis involves integrating data from a variety of sources, and building simulations or visualisations to identify patterns. Increased understanding drives the need for ever-higher data resolution. For example, studies on the 2001 foot and mouth disease outbreak used datasets tracking many millions of animal movements between farms, collected by DEFRA. But many of the government-collected datasets used in the group's work are open on a subscriber-only basis, or are provided on a confidential basis to research collaborators.

For this group, openness in research is motivated by the economic and scientific benefits of enabling new questions to be asked of datasets. Field data in particular are very costly to collect:

> "I deposited a database of every village in a few districts of Uganda. It took three months to geo-reference all of them. [And] when I'd finished I bumped into someone who'd been doing a very similar exercise for a completely different reason…If we'd coordinated we could certainly have saved each other lots of time, but we didn't. And I'm sure there are millions of examples of that sort of time-wasting."

***Infrastructuring***

Much of the group's work can be considered as support for infrastructure, since it contributes to reference datasets and models. These are currently openly released through publications, and the group anticipates that software models will soon be released openly too. The UK infrastructure for depositing public health and geographic data includes the Economic and Social Data Service and web services operated by some of the larger longitudinal studies funded by the MRC and the Wellcome Trust. But this group deposits datasets in international subject-based databases, since it believes that no UK data centre provides a natural home for them.

The group also contributes to a wiki site, EpiWiki, which is used to gather together relevant guidelines, reference sources on methods and techniques, and other resources. Originally published to give access only to the informal group of colleagues who set it up, it has since been made publicly accessible in the hope of gaining wider contributions, though this has not so far materialised.

***Translating and engaging***

Epidemiology has a direct impact on health policy, and the group sees open public release of data and models as a key to more effective impact. By integrating geo-referenced epidemiological data on populations in the *Atlas of Human African Trypanosomiasis,* for example, the group has produced a mapping tool which will provide a basis for new ways to target public health interventions to those specific areas at highest risk of sleeping sickness (Cecchi et al, 2009).

# 4. Benefits

Our case study participants identified a number of potential and real benefits from open working. In some cases the evidence is patchy, and it is not always easy to see direct causal relationships between open working and the benefits claimed. Cumulatively, nevertheless, the evidence suggests that for some groups at least the perceived benefits are powerful enough to make them wish to pursue an open working agenda further, and to persuade others to do so as well. This section outlines some of those perceived benefits. The qualifications, as well as the barriers and constraints, will be considered in Section 5.

## 4.1 Efficiency of research

Openness in research clearly has the potential to improve the efficiency of the research process by, for example, avoiding duplication of effort, sharing the costs – in time and effort as well as other resources – of developing research tools, promoting examples of good practice and the adoption of open standards, and reducing the costs of data collection.

We found awareness and examples of these kinds of benefits in all our case studies. In astronomy, the high costs of collecting data and the need to compete for time on expensive instruments have engendered a culture of sharing raw data, albeit often after an embargo period while the data are analysed and results published. While the 'science products' derived from that data are not commonly shared, the sharing of curated catalogue data and virtual observatory tools is now 'incredibly more efficient' than the paper records in use until recently.

In bioinformatics, the OpenFlyData tool developed by the Image Bioinformatics Research Group has already enabled scientists to retrieve from three separate data sources a collection of genes sharing the same gene expression profile. In doing so, it has reportedly saved the effort of searching on each of 50-70 genes individually.

In crystallography, re-use of data is well-established, and one of the key benefits of services such as the National Crystallography Service (NCS) and the e-Crystals open access database. According to the NCS Director

> "It's on an exponential growth curve at the moment…the process of making things openly available so that derivative science can be done is massively speeded up, so that all in all the wheels turn quicker."

In neuroimaging, ethical and regulatory considerations preclude the open sharing of data and other resources. But researchers collaborate in recruiting research subjects in order to reduce the level of study fatigue and the effort to find willing participants. There is also increasing interest in pooling raw data or the derived statistical maps to allow 'mega-analyses' of fMRI studies (Costafreda, 2009). It is also notable that much of our group's analytical work depends on the open source Statistical Parametric Mapping (SPM) tool published by the Wellcome Trust Centre for Neuroimaging.

In language technology, the linguistic corpora which are a fundamental part of the research infrastructure are widely re-used, but completely open public access has been until recently the exception rather than the rule. The AMI Meeting Corpus, however, has been made available in this way and has been the subject of hundreds of published articles. As with all corpora, its value increases as it is augmented and new annotation schemas are made available. And the widespread adoption of common standards for mark-up of text both in the research community and in industry has brought real benefits:

> "The XML standards and things like TEI standardisation…it just makes life so much easier because if people are all thinking about the standard, then at least there's a way of getting data from one tool to another."

In language technology it is also notable that open source tools are offering researchers the means to build models they say they would otherwise have been unable to build: a researcher on a human-robot communication project, for example, uses an OpenCCG application which is "at the basis of several systems I've written and I wouldn't be able to do it without that".

Similarly, in epidemiology the use of open source tools such as EpiSurveyor have transformed the collection of research data; our group is concerned to avoid the duplication of effort as exemplified by the collection of geo-referenced data from the same Ugandan villages.

## 4.2 Research quality and scholarly rigour

The potential for more effective review and scrutiny – by making information available about working methods, protocols and data – is clearly a motivator for those engaged in open working, even though there are some concerns about how effective the scrutiny can be in practice. Open notebooks in blog or wiki form – as featured in our chemistry group and to a lesser extent the image bioinformatics research group – provide the fullest open record of how a research project is designed and undertaken.

Maintaining a comprehensive and high-quality record of what has been done, how and why, is seen as a key requirement for all our groups. And even among the chemistry and bioinformatics groups, the key motivation for their work in documenting the research process appears to be not so much openness to the public, but maintaining a research record, and an ability to exchange information with colleagues and others who may wish to collaborate with the team. Benefits have thus arisen from sharing within groups and their collaborators, rather than through open access. Recent developments of the LabBlog tool by the chemistry group have thus sought to minimise the technical and organisational costs of working within and across defined groups, more than to promote public access (Frey, 2009).

Documenting and preparing data for scrutiny and re-use encourages attention to detail and to quality assurance, although it also demands time and effort. In astronomy, for example, the quality of the data held in and available from data centres is believed to have improved dramatically over the past

decade, although the centres themselves are thought to need more capacity to ensure high-quality documentation. And in chemistry, the LabBlog tool is seen as a data provenance record, effectively linking laboratories, data centres and articles in scholarly journals.

Checking data quality remains, of course, an important issue. In chemical crystallography the data is structured and formulaic, and some fairly stringent checks – for syntactic correctness, for example – can be performed automatically. Even here, however, humans cannot be entirely removed from the process; and lightweight data checking panels are being considered as one mechanism to give deposited data a seal of approval.

## 4.3 Visibility and scope for engagement

Several of our groups have noted a positive impact on the visibility and recognition of their work, particularly in areas – such as astronomy, crystallography and language technology – where forms of data publication are becoming widely accepted. In language technology, for instance, one researcher noted that

> "I've key-noted conferences on the back of having produced data resources and tools, and I get invited to more of the kinds of closed door meetings that might end up in generating funding than I would do otherwise."

Visibility within one's institution is also mentioned as a benefit. One astronomy researcher noted that

> "I think…it's an aid to communication within an institute as well if people can see what kinds of projects are happening in which area."

Examples of openness leading to engagement with the wider public were harder to find, although the astronomers in our case study were interested in the approach adopted by the US-based Sloan Digital Sky Survey which enables interested amateurs to classify images of galaxies online, through the Galaxy Zoo project. They see the potential of this 'citizen science' approach as a means of sharing workload. It may not necessarily, however, lead to a more open approach to putting information about the research process and findings into the public domain.

## 4.4 New research questions

The case studies show a number of examples of 'open working' making it feasible to address research questions or to adopt approaches that would otherwise be impracticable. In astronomy, for instance, the infrastructure created through Astrogrid and other virtual observatory projects allows a much more ambitious approach to the analysis of data and the search for patterns.

> "Instead of looking at one [solar] flare you look at thousands of flares and then you can find correlations…and]you can have a much clearer picture of what's going on than if you're just studying one in detail."

> "Astronomers are now expected to be what we call multi-wavelength [and] to reference what else we know about a particular object or class of objects from other wavelengths…You can do much better science if you're spanning a wider energy range and taking more than one window on a science problem."

The image bioinformatics group's work in data linking through its Linking Open Drug Data application also demonstrates new capabilities to find correlations by integrating and semantically marking up data from open sources and making them searchable in new ways.

In chemistry, data mining from open repositories in crystallography is reported to be having a significant impact in areas such as the predictive modelling of drug efficacy and of the behaviour of materials. And our neuroimaging group believes there is significant potential for improved diagnosis by investigating the associations between imaging and genomic data. More specifically, our epidemiology group aims through the integration of open geospatial and public health data sources to visualise in the Atlas of Human African Trypanosomiasis patterns in the geographic incidence of sleeping sickness.

## 4.5 Collaboration and community-building

Several groups pointed to how open working, through the sharing of data, software and other tools, reduced the barriers to working across institutional or disciplinary boundaries. Our astronomy group noted that institutional boundaries sometimes remained difficult to cross at an international level. Nevertheless, they point to how techniques developed through astronomy's open infrastructure have had a huge impact in various areas of medical image analysis.

Similarly, the image bioinformatics group's work has led to cross-disciplinary initiatives with clinical researchers and with art historians interested in building visualisation capabilities around image data gathered from different sources; the chemistry group is collaborating with neuroscientists; and the language technology group with psychologists.

Perhaps most interestingly, the potential for the pooling of genomic data with neuroimage data appears to be exerting a pressure from the relatively open genomics community towards more openness in the neuroimaging community, which up to now has been relatively less open in its workings.

There is up to now rather less evidence of openness facilitating the building of new research communities, perhaps because open working has not yet become sufficiently widespread to achieve the kinds of network effects that are necessary to build and sustain communities, as distinct from relatively small-scale or ad hoc collaborations.

## 4.6 Social and economic impact

Evidence of open working leading to enhancements in impact, or of synergies with open innovation, is patchy. The most potent example comes again from astronomy, where open working has played a part in leading to the collaboration with medical image analysts, which has led in turn to automated analysis of tissue microarray images; and new techniques of image analysis for detecting breast cancer and for analysing MRI brain scans. The MOPED software for the latter has been patented and commercialised through a spin-out company.

In language technology, the open sharing of corpora and of tools has played a part in the development of improvements in speech recognition software. Several of our groups are aware of the moves towards more open approaches to research and open innovation in the commercial sector:

> "It's very interesting that the pharmaceutical industry now is starting to lower its firewalls. They realise that the cost of doing all the research in-house is becoming prohibitive. They're outsourcing more and more of their research but they're also opening up access to some of their data."

But more generally, the experience is that working with commercial partners serves to limit openness. As one principal investigator put it:

> "Pharmaceutical companies have some strict restrictions on what can be made available… sometimes they effectively contract research in universities to answer a problem, in which case I'm bound and gagged every night I leave the building."

More positively, some of our groups noted that moves towards open access to government data were encouraging more data sharing in the research community; and epidemiologists have substantial experience of re-using government data in ways that have had a direct influence on health policy.

# 5. Barriers and constraints

There are of course qualifications to many of the benefits outlined above, as well as constraints which may prevent researchers from adopting the kinds of approaches advocated by the open science advocates. Researchers operate in a complex set of cultures in which it is important not only to communicate and disseminate their work, but to register their claim to what they have done, and to gain recognition and esteem from their peers. It is thus that they secure rewards in terms of a flow of funds to support their work, and advancement in their careers. In such a culture, the benefits and rewards that flow from open working – particularly at this early stage when adoption of such approaches is not widespread – may not always be clear.

## 5.1 Lack of evidence of benefits

Our groups acknowledge that large-scale improvements in the efficiency of research depend on achieving a critical mass in the adoption of common standards. In astronomy, some significant progress is being made:

> "If they feel it's getting to the point where that's the only game in town, that's what you have to do or you have to do or you suddenly become invisible, then…you'll see a viral spread."

But fundamental differences between measurement approaches in different sub-domains have to be accommodated, and the benefits of open working are thus at present unevenly spread:

> "There's still a lot more work to do….because things are done completely differently from one wavelength of astronomy to another."

Standards are similarly seen as essential in the image bioinformatics group:

> "Community standards are central for data publication and sharing…the key to success for linked open data is that…everything is described in vocabulary that is expressed in standard format."

In crystallography the International Union of Crystallography (IUCr) has sought to ensure that data sets accompanying journal articles are freely available on its website as supplementary material, but:

> "…now it's about changing the hearts and minds of everyone out there to be able to accept and adopt."

Aside from the issue of standards, there are some questions as to whether the tools adopted for open working, as currently configured, can actually deliver some of the desired benefits. Thus there are difficulties for our chemistry group in linking together data descriptions at the granular log level and narrative descriptions at the level of research protocols:

> "We have a real tool gap that prevents us from linking the two effectively together. We need those narrative descriptions because otherwise a human can't get at the data effectively."

## 5.2 Lack of incentives, rewards and support

Even where other benefits are clear, career rewards and incentives can be weak. There is a widespread perception that 'infrastructuring' work is not regarded highly in the scientific community, and that it is judged in accordance with inappropriate criteria. Many of our groups made a distinction between scientific and engineering cultures:

> "If you're an engineer, you're hired to do a particular job [such as] producing a piece of astronomical software. And you get your money and [get] promoted if you deliver what you were asked to do. But if you're an academic scientist, then you score your points by writing papers, thinking of things that other people didn't think of, getting the papers out first."

The relative invisibility of infrastructure is a barrier to demonstrating its impact:

> "Much of the work in the UK Virtual Observatory Project was into infrastructure which may not appear to the researcher….and in so many cases people will be making use of the derived resources without actually knowing that."

And in bioinformatics, problems can arise in securing recognition for applications that, while adding value to other researchers' work, may not be novel in computer science terms, or offer new forms of data, and may thus be difficult to publish as original work.

In order to address these kinds of problems, several of our groups pointed to the need for mechanisms to facilitate the award of credit for use of infrastructure, and especially of data. The chemistry group thus pointed to the potential significance of the DataCite project which aims to make datasets citable by assigning persistent Digital Object Identifiers (DOIs) to them. This makes possible the generation of metrics as evidence of the use and impact of datasets alongside the metrics available for journal articles. The hope is that such metrics will then be incorporated into institutional assessment regimes – and thus be considered in staff appraisals and promotions – as well as national assessments such as the REF.

But there is also a need for active and sustained support from funders:

> "The SPM software … has been heavily backed by the Wellcome Trust and as a result it's become the global standard and the people who devised it… have the money to sustain it, develop it, [and to] give you almost instantaneous answers to your questions about how to run it. That, I guess, is the best model."

## 5.3 Lack of time, skills and other resources

Open working requires time and effort, and researchers sometimes lack the skills required:

> "We certainly need the same kind of training that I think all researchers need in thinking more carefully about how they curate their own data… it's very ad hoc at the moment."

Even if they have the skills, they often seek to calibrate the efforts they are prepared to make and the benefits they expect to receive. The returns on those efforts are not always immediately obvious. One of the leading advocates of open science noted that benefits do not flow automatically from open working, and that community-building takes time:

> "You've got to put in the effort, you've got to show willing to help other people both to build up the community and to create your own credibility…One of the complaints people make… is that it's 98 percent noise. That noise is community building, which is critical for getting any value out of the system."

Aside from the efforts required for community building, the process of developing standards can be slow and tortuous, with many frustrations along the way, as a member of our astronomy group noted:

> "There are different historical legacies…between countries and between international projects, even within the same wave band…and so it's a long process to get international agreement. I think a big frustration since 2001 has been just quite how slow that process has been."

Similarly, considerable efforts may be required to produce the kinds of detailed documentation and semantically-expressive metadata that are required in order to realise the full potential of open release of data, code and tools. Researchers do not always see providing the resources for this kind of work as a high priority.

Some of the strategies for dealing with such issues include automatic streaming from instruments, embedding the creation and editing of metadata in other tasks, and improving the usability of metadata editing tools. Our chemistry group is testing automatic 'instrument blogging' using the LabBlog tool; but in epidemiology, an open notebook approach would require data to be linked to the questionnaire or other research instrument used to generate them. The image bioinformatics group is using the allied approach of 'sheer curation', seeking to reduce data creators' efforts by generating metadata as part of other research tasks such as visualisation. In astronomy, the use of standards-compliant metadata editing tools is reducing effort. But the key thing is still to have the resources to employ specialist staff:

> "For larger projects it's easy, we'll have one or two people dedicated to this aspect of the data and so they can become trained in how to do this. I think it's much slower for smaller research teams; it will take a long time before they routinely put their data into the virtual observatory."

Preparing code for open release can also be demanding and time-consuming, with an onus placed on the creator to document it sufficiently for re-use, and to provide continuing support. Such effort tends to be seen as an end-of-project activity that can easily be squeezed out, especially if it is not seen as a priority by the funders. Hence our neuroimaging group have a preference for 'warm sharing'. Rather than sharing code openly, they make it available to former colleagues familiar with their approach:

> "They've seen in great detail what we do [and] not only do they have the scripts but they have the sort of thinking behind them. And we're continuing to work with these groups to in effect have a shared environment, which is very warm share because they're people [who] understand what our thinking is."

Such an approach is shared by the chemistry group in their open notebooks. Providing a comprehensive record of their projects as they proceed is important both for communication with collaborators and to comply with safety regulations; and the LabBlog is seen as more effective than paper records. But at least in some areas they recognise that it is 'impossible to write up everything'. Moreover, group members are aware that their blog entries are most readily understood by their close collaborators. And some of them at least regard the effort that would be required to produce more generally-understandable entries as equivalent to that required to produce supplementary material for articles. Hence they prefer to see the blog as a post-publication supplement.

## 5.4 Cultures of independence and competition

Research is both collaborative and competitive, and researchers are strongly aware of the importance of the 'intellectual capital' they build up during the course of their work: both of the effort involved in creating it, and of the value it has as they seek to develop their careers.

> "Scientific fields can be very competitive…there is not so much sharing because there is competition instead."

Our groups thus recognise the reservations that many researchers express about sharing data in particular:

> "There are long term studies…which have gone on for 35, 40 years now, where the value of the data set grows every year as more observations are accumulated, and the research group, while taking great pains to preserve that data set, is very reluctant to make it open because that is the intellectual capital that many of their researchers build their careers on and they can't just give it away."

Similarly, our neuroimaging group is aware of the exchange value of their data and its attractiveness to potential collaborators:

> "We are seeking to collaborate and jostling for positioning in terms of having large enough collected samples of data that merit inclusion in these kind of consortia, because you're really talking about having several hundred patients to justify being included in these kind of things."

Hence we found general support among our groups for the proposition that researchers should be able to maintain a competitive edge by withholding data for a reasonable period to allow them to publish their results and conclusions, and to gain credit for that. The contrasting view is summed up in the oft-quoted remark of Robert Merton, that "in science, private property is established by having its substance freely given to others who might want to make use of it" (1988). The chemistry group also pointed out that time-stamping of blog and wiki posts may also offer some protection:

> "The key concern with putting things online is often that someone's going to take that and then publish it. And it's protecting against that where the time stamp's important. You can say this person submitted to this journal on this date. You can then write back to the editor and say no, well they may be the first to publish…but we did it first."

Perhaps paradoxically, in areas of work where researchers feel that rewards and credit are harder to obtain, a culture of open sharing is more evident. Thus in the 'infrastructuring' aspects of astronomy, chemistry and many other fields – the work in developing software, tools, standards, ontologies and so on – researchers are more willing to collaborate and to share their work openly than they are in the scientific work such infrastructuring is intended to serve.

## 5.5 Concerns about quality and usability

Researchers are concerned, as both creators and users of open data and other resources, that effective quality checks are in place. In physical chemistry, there is a long tradition of expert review of canonical datasets. But the ever-increasing volumes of data and other outputs are generating new problems:

> "It's very difficult to get a group of experts together in the way that you used to, and the quantities of information and the difficulty of getting it all together mean that you need a different approach to providing them…with the necessary information to enable them to come to an opinion on the validity of some data."

There is thus an increasingly widespread view that while peer review is essential, it is becoming unsustainable in its current forms. Some open science advocates therefore see no long-term future for pre-publication peer review. Rather, it will be replaced by tracking of use, comments and ratings, not only of publications, but of all kinds of 'research objects' including data, software, protocols, reagents and so on.

> "We'll start to see feeds of commentary of ratings aggregated from multiple sources, and that can scale because there the attention is focused on the things that are getting used, whereas the pre-publication period puts everything through a bottle neck…peer review of everything is important…we just need better ways of doing it."

But even where data or tools are of high quality, the exigencies and demands of individual disciplines may limit their use in other fields. Cross-disciplinary transfer of annotation tools from language technology, for instance, depends on mutual understanding of the terminology used in annotation schemas, and even then the potential may be limited by the resources that can be devoted to usability:

> "It's quite difficult for somebody with a humanities or arts background to be able to run the things because we don't have the money to package them nicely enough for them."

## 5.6 Ethical, legal and other restrictions on openness

Ethical and regulatory restrictions on openness apply in many areas of the life sciences and social sciences, and among our case studies particularly in neuroimaging. The data that can be shared openly is restricted to that which can be effectively anonymised. That may be of relatively little use, since analysis depends on identifying relationships between the neurological structures and functioning shown in the images on the one hand, and information about subjects' physiology, behaviour, and medical and social history on the other:

> "We have to make sure that there's no way that anyone can…get back to the people who have been involved in the study…There's a number of groups already putting raw scan data on the web. But the raw scan data is of minimal use because they will withhold subject details."

Working with data created or gathered by others – whether from the research community or from other organisations – may present a number of problems and restrictions:

> "A lot of astronomy data is free and open, but there's also a lot that's not. Some of the technical problems we try to solve are to try [to ensure] that some things could be completely open, and others completely private, and others open for a consortium of people that all want to be in the same club. And every model should be equally served."

Our astronomy group also expressed the fear that completely open notebooks might lead to a lack of candour in recording the progress of work, warts and all. Hence their preference for a secure area for project participants. Similar issues have arisen with our chemistry group's LabBlog system. Identifying who owns data and acquiring permission to use it complicates the requirements of openness still further, along with the development and use of linked data tools and infrastructure. Other frustrations can arise from the non-availability of government data or the software and models used by scientists in government restrict progress elsewhere:

> "It would be helpful if software tools could be made publicly available…If we are not able to replicate results then you're got a kind of, back of an envelope calculation and its not good science."

Restrictions may also arise, as we have noted (Section 5.6) when working with commercial partners. Commercial considerations may also arise from the work of researchers themselves, and there was some confusion and unease among our groups on how to deal with such issues. Most believe that decisions on whether and when to pursue commercialisation or open release of tools and techniques are best left to the groups themselves:

> "If you get to keep your IP it encourages you to be more open, to look for commercial opportunities, all those kind of things."

There was also broad support for a layered approach to openness, limiting access according to types of user or levels of data, for example. Thus metadata might be openly released for discovery purposes, providing enough machine-readable information to allow access conditions to be negotiated and determined. In similar vein, the language technology group has a differentiated approach to licensing for commercial and non-commercial purposes.

Some groups, however, are calling for more detailed advice from funding bodies on the balance between openness and commercialisation, and for:

> "a clearer idea there of whether we should be doing things openly, even if it's just to be accountable to the tax payer...it's very fuzzy. I find in the chemical and physical sciences everyone's very, very confused about what this means."

# 6. Conclusion

## 6.1 Degrees of openness

Many open scientists have drawn inspiration from the open source software movement, and many lay stress on the public good view of research and its fruits. But the motivations that drive them are not different in kind from other scientists. Rather, they have made choices which make sense to them in the precise contexts in which they are operating.

Our case studies thus illustrate how researchers vary in why they may wish to work openly, and in how far it is practical to do so at different stages in the research process. The different groups thus work with different degrees of openness with regard to the two key dimensions we set out in the introduction to this report:

- what kinds of information and other material they make available, at what stage in the research process, and how; and
- the groups of people to whom the material is made open, and on what terms and conditions.

The image bioinformatics group makes research proposals and designs openly accessible at an early stage; and at least some members of the chemistry group are committed in addition to providing access to comprehensive documentation of the research process in the form of lab notebooks, blogs and wikis. Others, such as the astronomy and language technology groups, however, are more cautious about providing access to research proposals, or open documentation of the research process, at least until results have been published in conventional form.

Providing access to data generated during the research process is rather more common, though again there are significant differences between the groups, particularly with regard to raw and refined data. Thus the astronomy group operates under disciplinary norms in providing access to raw data (particularly catalogue data) but not to refined data. Others, such as the chemistry, language technology and epidemiology groups, provide access to refined and derived data, and to reference datasets. But for the neuroimaging group, the confidential nature of the data they use imposes severe restrictions on accessibility.

All of the groups see themselves as making valuable contributions to the intellectual infrastructure for research in their fields. It is thus in relation to standards and protocols, software, analysis tools, methods and techniques that openness is most prevalent across all groups. They see making work of this kind freely available to others as a key contribution to the research community, not only in their own disciplines, but in others too.

As to the groups of people to whom the material is made available, the default position of most groups is not that everything should be open freely and without restriction to everyone. In some cases, such as with the research proposals and designs of the image bioinformatics group, the restrictions are minimal, in the form of a requirement to register and log in. In other cases, such as language technology, restrictions are imposed by third parties in the form of a requirement to pay a subscription.

Most groups, however, display a preference for openness in working with known collaborators. This is articulated most explicitly in the neuroimaging group's 'warm sharing' of tools with researchers who have worked with the group in the past. And while the members of the chemistry group work with differing degrees of openness, they share the view that notebook users should be able to control what is made open and when.

Imposing a delay on access to data and other materials is thus a feature of the practice of many of the groups. There is also a common realisation across the groups that making materials openly accessible requires effort to ensure that they are properly presented, with adequate metadata, proper calibration, high-quality documentation, and appropriate tools. Only if they are easy to understand and to use can materials be said to be truly open. But all this requires effort and resources, and has to find a place among competing priorities: "So it's both a good thing and also a kind of trap".

Restrictive conditions are fewest in relation to the kinds of tools, protocols and techniques made accessible as a service to the academic community, and when materials are made open to colleagues and collaborators, as distinct from the world at large. Nevertheless, access and use of materials may be restricted by third party licences, or where there is the potential for commercial exploitation (tools and techniques with commercial applications may be patented), or when dealing with sensitive personal or commercially-restricted data.

## 6.2 Incentives, benefits and constraints

Researchers' choices as to what to open, when and how are thus not straightforward, and they have to address a range of technical, cultural, disciplinary, ethical, legal and policy challenges. They do not make simple choices between open and closed ways of working. Rather, they work with different degrees of openness, which may vary within the group or team. Moreover, technological, cultural and policy developments are changing the environment in which they work, and attitudes and behaviours are changing as a result.

Our case studies suggest that the benefits arising from open working can include:

- *improved efficiency* in the research process, by reducing the costs of data collection; sharing the costs in time and effort of developing research tools; and promoting examples of good practice and the development of open standards. The sharing of data in fields as diverse as astronomy, neuroimaging and language technology is reported to have brought huge improvements; and the use of open source tools has enabled researchers to gather and analyse data in ways that would otherwise have been impossible.

- *improvements in research quality* and rigour, through more effective review and scrutiny. Openness gives added impetus to the requirement to maintain comprehensive and high-quality records of what has been done, how, when and why. The requirement may derive, however, as much from the needs of collaborators as of wider communities.

- *enhanced visibility* in the research community and in researchers's institutions, although the evidence for improvements in visibility and engagement beyond the research community is less strong, at least in the UK.

- *ability to ask more ambitious research questions* by aggregating or linking data for analysis from several sources. This means that researchers in fields ranging from epidemiology to chemistry have enhanced abilities to investigate associations and correlations, and to search for patterns, across a wide range of phenomena.

- *easier communication* with collaborators across disciplinary and institutional boundaries, though the evidence that openness facilitates the building of new communities is again less strong.

The benefits may be less strong in some areas than the open science advocates suggest, however; and the evidence that open working in itself enhances the social and economic impact of research is at best equivocal.

On the other side of the coin, even where the benefits to the research community that arise from open working may be clear, the career rewards and incentives for individuals and groups of researchers can be weak, especially when set against the time and effort required, and other barriers:

- *invisibility and lack of credit*. Researchers gain rewards in their careers through writing papers presented in high-status journals. Work in developing the infrastructure is generally not so highly regarded, and it may even be invisible. Hence there is interest in developing usage and citation metrics for data and other elements of the infrastructure.

- *lack of time and other resources*. Preparing data, software, standards, code or notebooks so that they can be understood and readily used by others beyond the team or groups of close collaborators takes considerable time and effort, and these tasks may not be given high priority unless the return on that investment is clear. Hence there is interest in finding ways to reduce the time and effort involved. But the response in some areas is to restrict availability to collaborators and colleagues who do not need detailed documentation and support.

- *competitive advantage*. Research is both collaborative and competitive, and researchers are strongly aware of the importance of the 'intellectual capital' they build up during the course of their work: both the effort involved in creating it, and the value it has as they seek to develop their careers. Hence our groups support the proposition that researchers should be able to maintain a competitive edge by withholding data for a reasonable period to enable them to publish their findings and conclusions, and to gain credit for that.

- *quality and usability*. Researchers are concerned, as both creators and users of 'research objects' of many different kinds, that effective quality checks are in place. They also retain a strong belief in peer review. But there is an increasingly widespread view that in its current

form it cannot cope with the vast volume and range of materials that are being produced. Hence the growing interest in 'post-publication' and machine-readable forms of peer review and tests of usability.

- *ethical, legal and other restrictions.* In the life sciences and social sciences in particular, ethical and legal restrictions may mean that only data of minimal value for research can be made openly available. In other cases, data and other resources created or owned by third parties – particularly commercial partners – may imply limitations on what can be made openly available. Researchers are also aware of the need to achieve a balance between openness on the one hand and the opportunities for commercialisation on the other, and some would welcome further guidance on this.

# 7. Recommendations

Our study suggests that the key issue for policy-makers is not so much how to maximise openness, but how to support individuals, groups, communities and institutions in working with the degrees of openness that provide clear benefits to them. That requires a clear understanding of what works for different groups and communities; and better policies and strategies to incentivise openness to the degree that it is appropriate in different contexts. Single solutions will not work for all kinds and areas of research.

## 7.1 Data management and sharing

There is an urgent need for research funders and institutions to work together to provide guidance and develop their policies to support and promote better management and effective sharing of research data. This is particularly urgent in the light of the recent difficulties at the Climate Research Unit at the University of East Anglia, and the need to provide clear guidance on the requirements and implications of the Freedom of Information Act and the Environmental Information Regulations. It is critically important that this work should be pursued in close consultation with the Information Commissioner's Office as well as with research communities working in different fields and with different kinds of materials; and that policy should be developed based on a close understanding of the working practices, needs and concerns of those communities.

## 7.2 Research infrastructure

Many researchers feel that their work in developing research infrastructures is relatively unsupported and unrewarded. Research funders and institutions should work with their communities:

- to support and promote the sustained development and use of resources, tools and standards, especially those – such as user-friendly metadata tools – that facilitate and encourage open working;
- to provide incentives and rewards to those who contribute to the development of such resources.

## 7.3 Training and skills

Researchers often lack the skills necessary for effective data management and for open working, including the legal, ethical and regulatory issues that they may encounter. Research funders and institutions should ensure that such issues are included as part of doctoral training as well as of continuing professional development, taking full account of the specific contexts in which researchers are working.

## 7.4 Business models

Many researchers are uncertain as to the balance they should seek between open working on the one hand and commercial imperatives on the other. Research funders and institutions should work together and with the research community to increase awareness of open business models, and

develop pilot business planning guidelines and toolkits to help researchers assess the opportunities and risks if they seek to work more openly.

## 7.5 Quality assurance and assessment

Researchers are concerned that current peer review practices cannot meet the need for assurance and assessment of the increasing volumes of resources that are being made available to the research community and others. Research funders and institutions should work together with the research community and publishers to provide guidance on how the peer review system might be adapted to address the issues raised by the release of the broad and growing range of resources that have not undergone pre-publication peer review.

## 7.6 Examples of good practice

Research funders and institutions should gather, assess and disseminate examples of good practice in open science, and ways in which it has brought improvements in the efficiency and quality of research; facilitated collaboration with other researchers; enabled researchers to address new or more ambitious research questions; and enhanced the visibility and impact of research, and engagement beyond the research community.

## Annex A: Statements of Principles on Open Science

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities [Accessed 31 August 2010 from http://oa.mpg.de/openaccess-berlin/berlindeclaration.html]

Research Councils UK Position Statement on Access to Research Outputs [Accessed 31 August 2010 from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf]

OECD Principles and Guidelines for Access to Research Data from Public Funding [Accessed 31 August 2010 from http://www.oecd.org/dataoecd/9/61/38500813.pdf]

Science Commons Principles for Open Science [Accessed 31 August 2010 from http://sciencecommons.org/resources/readingroom/principles-for-open-science/]

Open Knowledge Definition [Accessed 31 August 2010 from http://www.opendefinition.org/okd/]

Panton Principles for Open Data in Science [Accessed 31 August 2010 from http://pantonprinciples.org/]

## Bibliography

BBSRC (2010) BBSRC Data Sharing Policy [Accessed 17 August 2010 from http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf]

Cecchi, G., Paone, M., Franco, J. R., Fèvre, E. M., Diarra, A., Ruiz, J. A., Mattioli, R. C. and Simarro, P.P. (2009). 'Towards the Atlas of Human African Trypanosomiasis' International Journal of Health Geographics, 8:15

Costafreda, S. (2009) 'Pooling fMRI data: meta-analysis, mega-analysis and multi-center studies.' Frontiers in Neuroinformatics, 3:33

Frey, J.G. (2009) 'The value of the Semantic Web in the laboratory'. Drug Discovery Today 14(11-12), 552-561

Fry, J, Lockyer, S., Oppenheim, C., Houghton, J. and Rasmussen, B. (2009) Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes. [Accessed 17 August 2010 from http://ie-repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf]

Humphrey, C. (2006) e-Science and the life cycle of research [Accessed 17 August 2010 from http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc]

JISC (2009) Research Lifecycle diagram [Retrieved 17 August 2010 from http://www.jisc.ac.uk/what-wedo/campaigns/res3/jischelp.aspx]

Lyon, L. (2009) Open Science at Web-Scale: Optimising Participation and Predictive Potential

[Accessed 17 August 2010 from http://www.jisc.ac.uk/media/documents/publications/research/2009/open-science-report-6nov09-final-sentojisc.pdf]

Max Planck Society (2003) Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities [Accessed 17 August 2010 from http://oa.mpg.de/openaccess-berlin/berlin_declaration.pdf]

Merton, R.K. (1988), 'The Matthew Effect in Science .2. Cumulative advantage and the symbolism of intellectual property'. ISIS 70 (299), 606–623

Murray-Rust, P., Neylon, C., Pollock, R. and Wilbanks, J. (2010) Panton Principles: Principles for Open Data in Science [Accessed 17 August 2010 from http://pantonprinciples.org/]

OECD (2007) OECD Principles and Guidelines for Access to Research Data from Public Funding [Accessed 17 August 2010 from http://www.oecd.org/dataoecd/9/61/38500813.pdf]

Open Definition (n.d.) Open Knowledge Definition [Accessed 17 August 2010 from http://www.opendefinition.org/okd/]

Russell, M., Boulton, G., Clarke, P., Eyton, D. and Norton, J. (2010) The Independent Climate Change E-mails Review. [Accessed 17 August 2010 from http://www.cce-review.org/pdf/FINAL%20REPORT.pdf]

Research Councils UK (2006) Research Councils UK' updated position statement on access to research outputs. [Accessed 17 August 2010 from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf]

Research Information Network (2008) To share or not to share: Publication and quality assurance of research data outputs. Research Information Network

Samet, J.M. (2009) 'Data: To Share or Not to Share?' Epidemiology, 20(2), 172

Science Commons (2008) Principles for open science [Accessed 17 August 2010 from http://sciencecommons.org/resources/readingroom/principles-for-open-science/]

University of Southampton (n.d.) Welcome to eCrystals – University of Southampton [Accessed 17 August 2010 from http://ecrystals.chem.soton.ac.uk/]

Zhao, J., Miles, A., Klyne, G. and Shotton, D. (2009) 'OpenFlyData: The Way to Go for Biological Data Integration' Data Integration in the Life Sciences 5647/2009. 47-54

## Acknowledgments

This report is based on research undertaken by the Digital Curation Centre. We are grateful to them for the work they have done.



**The Research Information Network** has been established by the higher education funding councils, the research councils, and the national libraries in the UK. We investigate how efficient and effective the information services provided for the UK research community are, how they are changing, and how they might be improved for the future. We help to ensure that researchers in the UK benefit from world-leading information services, so that they can sustain their position as among the most successful and productive researchers in the world.

The Research Information Network
96 Euston Road
London NW1 2DB
contact@rin.ac.uk

www.rin.ac.uk

**NESTA** is the National Endowment for Science, Technology and the Arts - an independent body with a mission to make the UK more innovative. Our endowment status means we operate at no cost to the UK taxpayer.

We invest in early-stage companies, inform policy, and deliver practical programmes that inspire others to solve the big challenges of the future.

NESTA does not work alone. Our success depends on the strength of the partnerships we form with innovators, policymakers, community organisations, educators and other investors. We bring the best ideas, new flows of capital and talented people together, and encourage others to develop them further.

NESTA
1 Plough Place
London EC4A 1DE
research@nesta.org.uk

www.nesta.org.uk