GOVERNANCE OF DATA ACCESS: ANNEXES

June 2015

CANCER RESEARCH UK
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
MRC Medical Research Council
wellcome trust

ANNEX 1: DATA ACCESS STAKEHOLDER RIGHTS, RESPONSIBILITIES, INTERESTS & RISKS

Stakeholder	Rights	Responsibilities	Interests	Risks
Data requester/ user	Clear structure for appeals; transparent access mechanisms; appropriate level of support in making access request; easy to locate and navigate relevant information about data and access process.	Should comply with requests for information; should truthfully state reasons for access requests; should supply accurate credentials; should sign access agreements in good faith.	Accessing good-quality data in a timely manner and without incurring excessive costs or delays.	Time consuming access processes; professional disagreements over access and use of requested data.
Data producer	Right to an embargo period if data are released early; manageable time period and resources to be allocated for supporting data access; credit and recognition for making data accessible.	Provide advice to DACs or those requesting data in a fair manner; ensure compliance with data protection legislation if in possession of personal data.	Ensuring data is used appropriately; maintaining career interests and reputation of self and study; protecting confidentiality of study participants.	Breach of trust by data users (e.g., 'scooping' on publications, breaches of access agreements) onerous workload to format data; competition from other researchers using datasets.
Data Access Committee member	Manageable workload; support in making particular decisions; assistance in managing appeals in an appropriate manner.	To fairly consider access requests in accordance with ethical guidance and any other relevant criteria; openness and transparency about reasons for access decisions. Ensure compliance of data-sharing practices with relevant legislation.	Ensuring good process and accountability for data access processes and decisions; supporting data community; complying with funder policies.	Onerous burden of work; potential conflicts of interest with other researchers over use of data; potentially risking confidentiality of data.
Funder	Clear flagging of issues from others in the system (DACs/PIs/data users); access to information on data us for monitoring and audit purposes.	Clear advice to DACs and PIs as to their responsibilities and funders' expectations; providing support for DAC members and PIs to achieve the level of access expected by funders.	Advancing biomedical or health-related research; maintaining good and transparent governance procedures.	Undermining public trust in research if data security is breached; poor implementation of policies leading to research being stifled; excess bureaucracy limiting value of data investments.
Research participant	Clear data access mechanisms and processes explained from outset of study; terms of consent to be adhered to; rights and interests protected by the study.	Familiarise themselves with how data may be used and to ask questions of the data producers if unsure prior to giving consent.	Ensuring data is securely accessed by qualified, appropriate researchers; privacy; providing a benefit to society as a whole; advances in research/ treatment.	Potential re-identification; stigmatisation/ discrimination; feeling of loss of control over personal data.
Institution	Transparent, proportionate and cost effective access and sharing mechanisms for researchers.	Ensure researchers are accessing data legitimately and complying with Data Access Agreements	Enabling data sharing and usage to enhance research capacity and reputation.	Potential loss of access and withdrawal of funding if DAAs are breached; reputational damage; loss of competitive advantage if data openly shared.

ANNEX 2: THE FUNDING AND RESEARCH ENVIRONMENT

1. Interest in improving the administration of access for data and linking between datasets is increasing in a range of contexts, in industry, government and academic research:
   - Several partners from the pharmaceutical industry have joined to create an online platform for allowing access to clinical trials data.[1]
   - Yale University has developed a model to facilitate access to patient-level clinical trial data, named the Yale University Open Data Access (YODA) Project.
   - The UK Government is continuing to support increasing the availability of large amounts of routinely collected administrative data for linkage in research, through the ESRC's development of the Administrative Data Research Network.[2]
   - The Global Alliance for Genomics and Health is seeking ways to reduce the barriers to responsibly sharing genomic and clinical data internationally.[3]
   - MRC has developed a Research Data Gateway as a platform to improve the discoverability of MRC cohort datasets.[4]
   - The UK Government is planning to roll out its delayed *care.data* scheme in the NHS, extracting patient records from GP surgeries to centralise in the Health and Social Care Information Centre database. There are tentative plans to allow accredited researchers limited access to linked anonymised primary and secondary care datasets at some point in the future.

2. The ESRC, MRC, CRUK and the Wellcome Trust have all had policies in place for several years requiring that data that could be a useful resource for the research community is made accessible to secondary users where feasible (see **Annex 7**). However, up until now there has been little scope to ascertain to which grants such policies are applicable (with the exception of ESRC), assess the effectiveness of these policies across funders and establish whether data access is improving in practice.[5] It is difficult even to audit how many studies supported by EAGDA funders have formal data access mechanisms in place.

3. Within the research community, EAGDA recognises that there is a lack of clarity over what the requirement to share data where possible means in practice: what data should be accessed, under what circumstances, how should access be governed and what properly constitutes "access" to data? In some cases there are good reasons to withhold or restrict access, and the controls on the use of data depend fundamentally on the type of data, who is using or proposing to use it, and for what purposes.

4. Additionally, most studies are project-driven and focused on the outcomes required of the specifically funded project. The majority of grants are awarded competitively on short-term funding cycles, on the basis of proposals to conduct original, high-quality science. It is therefore unsurprising that, as enabling data access for secondary users has increased in prominence, individual studies have put in place access mechanisms that have suited their

[1] https://www.clinicalstudydatarequest.com/
[2] http://adrn.ac.uk/
[3] http://genomicsandhealth.org/
[4] https://www.datagateway.mrc.ac.uk/
[5] Some initial work has previously been undertaken: in 2009 the Wellcome Trust commissioned research to identify the access requirements and governance models developed for GWAS and cohort studies in the UK and USA.

own resources, staff capacity and culture, generally within the costs of their core grants, with data access activities undertaken whilst primarily focusing on the core business of that grant.

5. This has led to a "cottage industry" landscape of data access, with differing approaches across different fields and study types. Inevitably, this variation generates difficulties for potential secondary users of data, not only in terms of discovering datasets but also in attempting to navigate a number of different access requirements, processes and policies in order to seek access to different datasets. Particularly for users requiring access to a number of datasets (for example, in genetic or genomic studies), negotiating these different requirements can severely impede and delay research.

6. Whilst it is important that the local characteristics of individual studies are recognised and not stifled by prescriptive regulation, EAGDA recognises that some degree of co-ordination in the approaches to data access studies use may be of benefit, both to studies themselves and to data users. Repositories have their own standards and protocols for the depositing of data, but it is important to recognise that the whole life cycle of data needs to be considered when planning, setting up and maintaining data access mechanisms.

7. Many studies now use a Data Access Committee (DAC) or equivalent body such as a project Steering Committee to make decisions on access requests and to oversee the management and administration of data access. DACs are one potential mechanism designed to ensure the protection of study participants and in some cases, ensure high quality science is conducted using a study's datasets. They are, however, only one part of the complex system of data access. These mechanisms have evolved in response to the needs of the community but with little overarching strategic guidance. They tend to be attached to a specific study, and as a result each has been developed in response to the context, needs and politics of a particular data community, often overseen by the study lead.

8. Unlike the US, where many of the studies in question are funded under the umbrella of the NIH, there is a piecemeal approach to the way data access operates in the UK. With little top down oversight (which has not in the past been needed) a variety of different mechanisms has been developed, for logistical, institutional, administrative, cultural and technical reasons. However, with increasing possibilities for linking across datasets and a burgeoning culture of data-driven research, it is not clear whether these systems of governance and management can continue to be fit for purpose.

9. There is some movement towards consolidating DACs across similar studies, for example the Wellcome Trust Sanger Institute now manages requests for data from the Wellcome Trust Case Control Consortium as well as several other projects based at the Institute, and the International Cancer Genomics Consortium (ICGC) has common access procedures for a range of component studies. Consolidation may be appropriate in some circumstances but has not previously been strategically considered at a high level by funders.

10. In light of the wide range of research, industry and government contexts in which data potentially can be accessed and linked, there is a pressing need for a high level policy response to ensure that:
    - the right balance can be struck between maximising the use and value of data for research and protecting participant confidentiality;
    - administrative and practical barriers to data access can be reduced where possible.

ANNEX 3: RESEARCH ON GOVERNANCE OF DATA ACCESS

AIMS AND OBJECTIVES

1. This project aimed to examine the data access landscape for UK genetic, epidemiological and health-focused social science cohort studies supported by the EAGDA funders.[6] The goal was to identify issues on which funders could take action to support their research communities in promoting the development and maintenance of robust, proportionate data access mechanisms.

2. The project had three primary aims relating to the management and governance of data access:
   - To assess the number and range of Data Access Committees (DACs) and other governance mechanisms for data access in the UK, together with international studies and consortia for comparison.
   - To ascertain whether current mechanisms are operating effectively and determine the key barriers to increasing the accessibility of data for secondary use, whilst ensuring appropriate protections for participants.
   - To develop recommendations for EAGDA funders on assessing, maintaining and supporting data access processes and governance for different study types.

METHODOLOGY

3. The research phase of the project initially involved two main components:

- An **analysis** of a selection of currently funded cohort, epidemiological and social science studies in 2013-14 (detailed in **Annex 4**):
  - A sweep of grants funded by EAGDA's funders, to explore how data access was managed across different projects and to identify any areas in which EAGDA could potentially provide guidance. The search strategy involved only the resources that would be available to a researcher unfamiliar with the study, and hence was restricted to information on studies' data access processes that is publicly available.
  - Study websites were analysed to assess the type of study, size, data access mechanisms, the degree of independence or oversight of access decisions, and the transparency of data access policies.
  - A total of 61 studies were included in the UK analysis. A further 20 were excluded from analysis as the studies either could not be found on the basis of their titles, did not have websites with details of the specific study, or they appeared to be out of date.
  - Five international initiatives with well-established data access mechanisms (MalariaGen, ICGC, dbGaP, The Cancer Genome Atlas (TCGA) and CARTaGene) were also analysed separately, to provide points of comparison with UK studies.
- A web-based **survey** of data users, structured around the evidence base from the data access analysis, to explore the issues secondary data users face in discovering, accessing and using shared data resources (detailed in **Annex 5 and 5a**):

[6] The project did not consider in detail the use of government administrative data or the use of NHS patient data.

- An article was composed by the EAGDA chair, Martin Bobrow, for a 'World View' column in *Nature*[7], briefly presenting some of the challenges of balancing access to research data with participant privacy, and inviting readers to complete the online survey which was linked to the article. The survey link was also circulated to the EAGDA members and funders to pass on to interested parties and funded researchers. The survey was targeted at as wide a research audience as possible, across different disciplines that involve data sharing.

- There were 111 respondents from a range of disciplines, approximately half of whom were involved in biomedical or health-related research.

- Although the article specifically discussed the challenges of sharing data from human participants, survey respondents included those who did not use human data at all in their research. These were included in the analysis to provide some cross-disciplinary context to the survey findings.

4. Following examination of the data access analysis and user survey, EAGDA discussed the key issues emerging for cohort and case-control studies in the UK and identified several areas in which the group considered strategic funder action would be appropriate. In order to ascertain whether these issues accurately reflected what was happening in practice in the research community, it was decided that cohort leaders should be approached to be interviewed on the issue of data access for their studies.

5. Therefore, in August to September 2014, a total of 16 study leaders and data managers from UK cohorts, selected by the funders, were contacted for a 30 minute phone **interview** (detailed in **Annex 6**). The interviews aimed to establish how the data access mechanisms for their respective studies worked and whether there were particular issues regarding data access that they felt funders could better support.

6. The interviewees were selected to represent a range of cohort studies, from small, specialised studies to large-scale projects set up specifically with data sharing in mind. The sample was restricted to cohort leaders and did not include case-control studies, as it was felt that the cohort criterion helped defined a reasonably sized group of interviewees to contact. All study leaders who were contacted agreed to participate.

7. This report summarises the key findings and conclusions of this project, detailing the evidence upon which the recommendations above are based.

8. A summary of EAGDA funders' data sharing policies is provided in **Annex 7**, together with a comparison between the governance principles clearly supported by these policies. It is intended to highlight the similarities and differences between the policies for the purposes of providing an overview to inform EAGDA's discussions.

[7] Bobrow, M. (2013) "Balancing privacy with public benefit", *Nature,* 500:746, 123. http://www.nature.com/news/balancing-privacy-with-public-benefit-1.13506, doi:10.1038/500123a

ANNEX 4: DACs AND MANAGED ACCESS ANALYSIS

High level summary

1. There are a range of data access mechanisms across the studies analysed, with varying types of control and oversight. A large number of studies (n=29; 47.5%) did not provide any information at all on data sharing and access on their public sites.

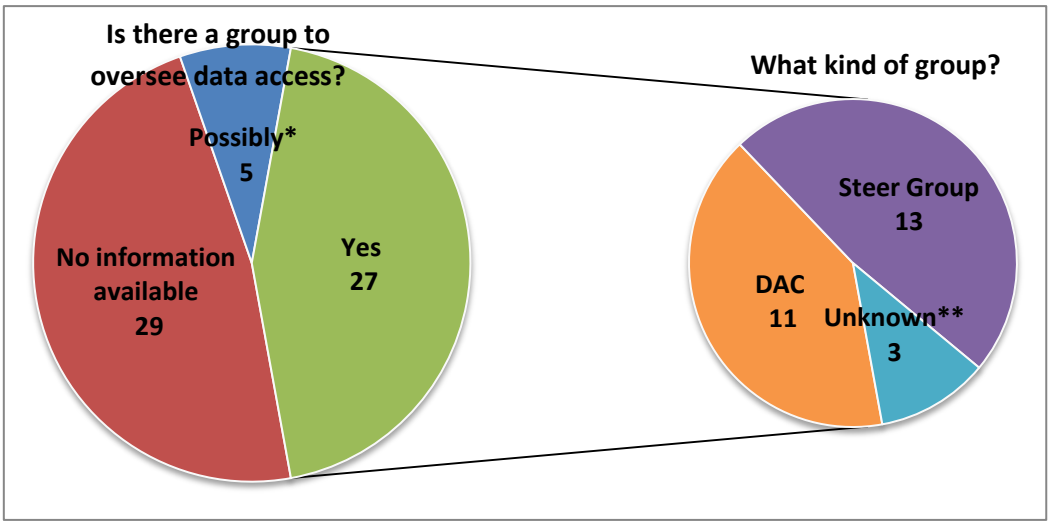


Figure 1: Analysis of UK studies *indicates there is some information suggesting data is available for access, but no specification of what the access mechanisms are. **studies are in EGA and so have some data access process in place, but no further information is easily available.

2. From the limited information available on study websites, some trends can be seen in different types of study. The most clear cut comparison can be drawn between studies of different sizes:

   - Large scale (>10,000 participants, n= 27 studies) cohort studies tend to have clear access policies, a group or committee to oversee data access, and their websites are designed for use by research participants as well as researchers. Information about data access is easy to locate, and in many cases the process of access is clearly defined, with named members of the study team and DAC available.
     - The Centre for Longitudinal Studies (CLS) cohorts use a consolidated DAC between them (1958, 1970, Millennium Cohort Study) for complex access decisions, which also provides a place of arbitration for contested access requests from Twins UK.[8]
   - Medium scale studies (1,000-10,000 participants, n= 18) vary in the availability of information about their data access procedures (see Table 1). The two studies with DACs formed part of larger groups (WTSI Cancer Genome Project and 1946 Birth Cohort), which may account for their relatively well-established data access procedures.
   - Small scale studies (<1,000 participants, n = 10)[9], across all study types, did not provide any information on data sharing. This may be because the study websites were poorly resourced and contained only very basic information about the study, with contact details of the study Principle Investigator (PI) I for further information. These

[8] At the time of writing this set-up is under review, with a view to further consolidating the DAC with Understanding Society.
[9] A further 6 studies had an unknown number of participants.

studies may also have a limited community for data sharing owing to the small scale of the data they produce.

	Large (n= >10000)	Medium (n=1000-10000)	Small (n= <1000)
DAC	9	2	0
Steering Group	6	7	0
PI	3	1	0
Other	2	0	0
Unknown	7	8	10

Table 4-1: Breakdown of access governance by size of study

3. Of the studies with a specific mechanism in place for managing data access:
   - 11 have specially constituted DACs. These are committees that are convened primarily for the purpose of making decisions about data access for at least some variables.
   - 13 have Steering Committees/Groups. These groups are convened to manage the study as a whole, and will consider data requests as part of their other business in running and overseeing the study.
   - There does not appear to be any clear distinction between studies opting to create a DAC and those using a steering or management committee, in terms of type of data or study.

4. There are more DACs than steering groups overall, particularly among large studies. It may be that smaller studies are less likely or able to set up DACs of their own. This is a point of interest for EAGDA: where governance structures are not in place specifically to manage access to data, the importance and value of widening access may be overlooked and projects may not recognise possibilities for enabling data access efficiently (e.g., through consolidation of data access mechanisms). Additionally, general steering groups may lack the expertise or the right composition to undertake realistic appraisals of the risks associated with granting access to data.

5. Level of demand may influence the type of access mechanism chosen for a study: for a well-known and established resource, it would be justified to devote specific administrative and technical resources to data access (e.g., the Cancer Genome study at the Wellcome Trust Sanger Institute receives 1-2 data access requests per week, and the Institute has a Data Officer to process these requests). However, it is difficult to ascertain cause and effect here: if a resource has straightforward access mechanisms and is well supported, this may itself create demand among researchers

Comparisons between UK and international DACs

6. Internationally, there is a high degree of variability in the way data access mechanisms are set up, that is similar to the variability observed in the UK. There may be difficulty in developing international frameworks given the different ethical, legal and regulatory requirements across borders, although as the Public Population Project in Genomics and Society (P3G) Generic Access Agreement for genomic studies demonstrates, it is possible to produce high-level standards and broad principles that a range of studies could subscribe to.[10]

[10] http://www.nature.com/nbt/journal/v31/n5/extref/nbt.2567-S1.pdf

7. For those studies that do have DACs, the access processes appear similar to those of the comparator international studies, and follow broadly the same series of steps.

8. One point of difference is the level of oversight the NIH has in the US, as a result of being an overarching funder for a large number of studies and institutions. The dbGaP database, for example, requires data users to apply to a centralised NIH-controlled DAC for access to controlled (individual-level) data resources. This requires users to register for an NIH extramural account if they are not NIH staff. The NIH DACs are established around programmatic areas of interest and the degree of technical and ethical expertise required to assess access requests for the particular type of data they control.[11]

Governance

9. DACs have varying roles depending on the purpose they are intended to serve. Five particular roles were identified from scoping the study sites:
   - establishing that researchers are bona fide (verifying credentials);
   - ensuring the utility and sustainability of a collaboration;
   - ensuring the terms of participant consent are adhered to;
   - ensuring compliance with legal and regulatory requirements;
   - ensuring high quality research is fostered[12].

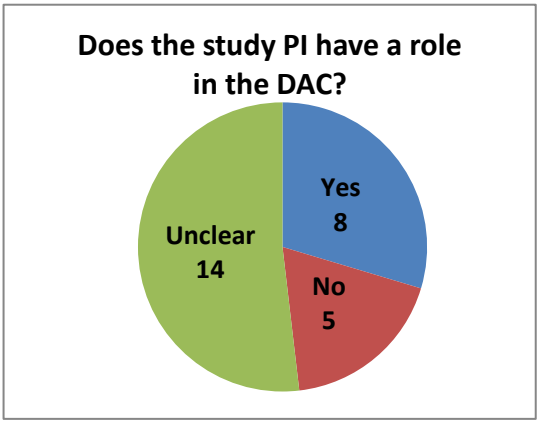


Figure 4-2: Role of PIs in DACs

DAC membership (independent oversight/role of PI)

10. In line with previous qualitative research indicating uncertainty over the appropriate composition of DACs[13] there is a high degree of variability in the role study PIs play in data access decisions. In some cases, it appears that the PI provides technical expertise to the committee and a deep understanding of the datasets (e.g., ALSPAC). In others, the PI is the first point of contact for those wishing to make an access request: the PI then collaborates with data requesters to develop their access proposal and may make a recommendation to the DAC. Where the PI is involved in the DAC, there is too little information on the DAC processes available to ascertain clearly whether the PI is responsible for the final access decision.

11. Some PIs elect to have an independent faculty leader sign off on data access requests; others prefer to retain full control (especially if their research involves vulnerable populations). In discussion with staff during the course of this research, it was noted that at the Sanger Institute a number of studies request a faculty member not involved in the study to oversee access decisions, in order to ensure a degree of independence to the process.

[11] An overview of dbGaP's data access policy is here: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html
[12] The CTSU DAC does assess the "scientific probity" of data access requests. This is unusual for data requests, although this may be explained by the fact that the unit controls access to samples as well as data. See p. 3: http://www.ctsu.ox.ac.uk/research/Data%20Access%20and%20Sharing%20policy%20V1%205.pdf
[13] Kaye & Phillips, Background briefing for EAGDA, 2012, p.5

12. Anecdotally, there were suggestions that PIs often lack trust in data access processes or wish to retain control even where there are clear mechanisms and protocols established (e.g., having a dedicated data access officer to process requests). As a result, it is difficult to implement policies to ensure adequate independence and oversight – this factor was unable to be picked up in the analysis, which focused purely on the policies in place.

Transparency

13. Nearly half of the studies analysed do not have any information publicly available on how researchers who may wish to use the research data should seek access. Although it was not always possible to establish when the studies were originally funded, from the information available on the study websites it is plausible that some began before the funders implemented their data sharing policies, and so provisions for data access were not standard practice at the time. There were also limitations for the search strategy in seeking this information: data access information may have been held on an institutional or group website that was not signposted from the study site.

14. However, given the current data sharing policies of the EAGDA funders, it is striking that such a high proportion of studies do not mention data access or provide any contact details for researchers who may wish to find out more about the study variables.

Accountability

15. There is a lack of clarity in data access mechanisms even for some well-established cohorts over who has the final say in access decisions and authority to sign off requests. This issue may not arise where there is agreement within the DAC and no conflict with funder or institutional policies. Nonetheless, it will be important to understand clearly the structures of accountability where disagreements do, and inevitably will, arise.

16. There may be wide scope for using different models of data access depending on the study and data type, but it is arguable that these models should be unified by clear structures of accountability and responsibility. It is often not clear to whom DACs are ultimately answerable, or whether there is a reporting or auditing structure in place.

Right of appeal

17. Four studies provide details of an appeals procedure for when a request for data access is rejected. Previous research has suggested that many studies do not have a formal appeals procedure as it was deemed unnecessary: there are few or no refusals of access, and no experience of dissatisfaction with the DAC decisions.[14]

Study	Appeals process
Twins UK	Criteria for rejection listed, and there is an entitlement to appeal decision: - Data requested is not stored - Data is embargoed for use in another collaboration - Lack of justification for use of depletable samples - Collection of the requested data/samples conflicts with our duty of care to not overburden the twins. A decision on rejection of collaborations will be overseen by the ACCC (Access Committee for CLS Cohorts) who will do an independent assessment of your application

[14] Kaye & Hawkins (2010) GWAS and Cohort Studies in the UK and USA – Access requirements and governance models, p.20

	before a final decision is reached
UK Biobank	Within 3 months of the relevant decision, the applicant PI should submit a written request, giving their reasons why they consider that the decision should be revised; the Access Sub-Committee or the Board (as appropriate) will aim to consider it along with the original application within 4-6 weeks.
Whitehall II	Outlined in application guidance
1958 Birth Cohort	Any applicant who wishes to appeal the decision of the ACCC will be able to apply to the BCSC, but this will require a documented (self-contained) description of all of the relevant background and a formal justification for why the decision is being appealed.

Compliance/ handling breaches

18. From the data access policies of the studies analysed, it is rare for details to be given about sanctions against breaches of the access conditions. Of note, no information was found about recourse participants might have against researchers or institutions in case of a breach of data handling.

19. Of the Data Access Agreements (DAA) for UK studies examined, only the ESRC End User License and Special License (used by the Understanding Society and other cohorts for some types of data) contained reference to the possible sanctions against an individual who breaches the terms of the license (see Box 1). Other DAAs emphasise that any attempts to re-identify individuals from the data being supplied would constitute a breach of the terms, but do not specify sanctions.

20. For Office of National Statistics (ONS) data, which is held under the Statistical Services and Registration Act (2007), the disclosure of personal information is a criminal act punishable by a maximum of two years imprisonment or a fine.[15]

Box 1: Abridged terms of the UK Data Service's Special License (Completion Notes s.2,17)
Sanctions that may be applied:
1. For a first offence, the penalty should be a minimum twelve-month non-discretionary suspension from access to any micro-data
2. An individual's second breach would, as a minimum, result in a suspension of access of two to five years, or permanently, on the individual, and would generate a written warning to the individual's institution.
…
4. Any discretionary penalty may be decided, including permanent suspension for the individual or other staff in the relevant department, and/or pursuing in the Courts an action for breach of contract.
5. Where the breach is the result of an institution's wilful or negligent action, then a minimum penalty of a twelve-month non-discretionary suspension shall apply to the relevant department within the institution. Repeated breaches will result in a letter with discretionary penalties to the institution as a whole including suspension of all data access facilities for all the institution's staff and/or an action for breach of contract.

21. Statistics from the UK Information Commissioner's Office (ICO) were examined in order to provide some context to the issue of data breaches in public body, administration and research settings. In the UK, the majority of breaches investigated by the ICO (in the first quarter of 2013) occurred because of errors in disclosure.[16] Health organisations account for the highest number of breaches. It is worth noting that academic research data do not, at the current time, tend to be hacked, but that data breaches are more likely to occur through carelessness or maladministration than deliberate misuse.

[15] UK Data Service Breaches Penalties policy, p.3
http://ukdataservice.ac.uk/media/176861/UKDA142_SDS_SecurityBreaches_public.pdf
[16] http://www.ico.org.uk/enforcement/trends

Sustainability

22. It is widely recognised that enabling data access comes with costs, both in terms of financial costs required to set up and manage mechanisms for access, and in terms of time commitments for staff. From the DACs analysis, only 11 studies provided any details on the costs associated with requesting access to data. These vary according to several criteria, depending on whether the data:
    - are encrypted and an encryption key needs to be sent to the data user securely via courier;
    - are sent in hard copy, which may give rise to postal charges;
    - need to be 'extracted' from the dataset and processed, or clerical or statistical support from the study team is required;
    - need to be extracted, analysed or 'cleaned' in order to be sent to the data user: this incurs financial and time costs for the data managers.

23. Only the UK Data Service makes reference to specific charges (£500) for commercial use of data. Twins UK standardly charges £500 for raw data access, with extras for additional work required; UK Biobank charges £250 for the preliminary application, with full applications charged on a cost-recovery basis. As of April 2014, ALSPAC charges on a bespoke cost-recovery basis depending on the nature and complexity of the access request.

24. Some studies (e.g. Born in Bradford) allocate a staff member to support each access request, providing guidance and feedback as the proposal is developed. We do not know how large a time commitment this entails, or whether staff are remunerated for additional time spent supporting data access requests. The Sanger Institute has a dedicated staff member for processing DARs, who is also helping to develop a more streamlined online access mechanism.

25. For datasets controlled by the study PI, there is a practical issue for the sustainability of data resources as it is not clear what happens to the dataset if the PI moves to another institution or retires and the study becomes an "orphan". One identified example of this problem was a dataset controlled by a PI who has subsequently moved abroad: it is difficult to maintain contact and the former PI is isolated from the rest of the collaborators in the study, meaning that if access decisions do get made, they are made independently of any scrutiny or dialogue with others. Continuity of access and longevity of the data resource could be assured if control and oversight of data access was managed by a body or individual independent of the study PI.

Different access levels

26. One of the main reasons for creating a DAC to oversee the management of shared data resources is to ensure that potentially identifiable data[17] are accessed only by legitimate researchers, who are bound by the terms of a data access agreement to use the data only in appropriate ways. The categorisation of data as either being potentially identifiable or not in the context in which it will be used is therefore crucial the access decisions DACs make. It is important that judgements about risk and potential identifiability are proportionate when categorising data.

[17] Both increases in the sophistication of data analysis and the scale of information that is available have implications for the possibility of re-identifying data that has been through a process of anonymisation.

Genetics and genomics

27. Generally, genetic and genomic data is managed through a two-tiered system, with controlled access of potentially identifiable and individual-level data and open access for aggregate level data that contains no personally identifying or identifiable information. The wide range of data types makes it difficult to ascertain whether risk of identifiability is the primary criterion for categorising the data as open or controlled access:

Name	Access levels	Details
dbGaP	2	**Open access**: Studies; Study documents; Phenotypic variables; Genotype and phenotype analysis. **Controlled access**: De-identified phenotypes and genotypes for individual study subjects; Pedigrees;
Sanger Institute	4[18]	**1**: no specific security requirements **3**: only specific people can access, but data don't require encryption (most WTSI data are level 3, e.g., DDD contains images, but no names, DoBs etc.) **4**: Data Protection Act applies, personal identifiable information. One level 4 social science study in WTSI, relating to attitudes towards genetics, incidental findings
ICGC	2	**Open**: raw genetic data **Controlled**: pathology data; histology data; personal data, genetic data

Social science

28. The UK Data Service is the primary UK repository for social science and economic research datasets. It currently contains around 7000 datasets, and access to these is provided around 54000 times per year. Few of the datasets are controlled by DACs, and the majority of data are available for download subject to users signing an End User License. There are four levels of access, categorised according to whether the data contains personal information, defined by the UK Data Service as "*information that relates to and identifies an individual (including a body corporate) taking into account other information derived from published sources*"[19].

Epidemiology

29. None of the studies that involved the collection of epidemiological data provided details on data categorisation. The Clinical Trials Study Unit (Oxford University) provides the most detailed information on data access processes for epidemiological data, and implies that all data will be controlled by a DAC or Custodian – usually the study PI.[20]

Government data

30. Government Business Impact Levels (IL) are standards for categorising government data according to their level of sensitivity and the protocols that should be followed for handling them. They are specified for different concerns (e.g., business impact, financial impact etc.,) but are broadly comparable and are intended to provide cross-departmental standards. This leaves open the question of whether and how the use of government data

[18] Pers comm. Carol Smee, WTSI, June 2013
[19] UK Data Service Breaches Penalty Policy, p.2 http://ukdataservice.ac.uk/media/176861/UKDA142_SDS_SecurityBreaches_public.pdf
[20] http://www.ctsu.ox.ac.uk/research/Data%20Access%20and%20Sharing%20policy%20V1%205.pdf

could be integrated into research, particularly as the data categorisations do not necessarily match up with one another across disciplines.

Different access protocols in single study

31. Some cohort studies specifically categorise data according to its risk of being identifiable, as it has different access restrictions for different categories. ALSPAC's categories include:
    - potentially identifying data (requires submission of a proposal that may be scrutinised by the Ethics and Law Committee, may require a Data Transfer Agreement (DTA) to be signed);
    - interpretable data (such as images and scans: access to raw data may require a DTA);
    - genotype data (requires a DTA and an additional agreement for non-Bristol staff);
    - GWAS data (available only to Bristol staff, but with plans to create a secure remote access facility in the future).

Data Access Agreements

32. Data Access Agreements (DAAs) form an integral part of all data sharing models except for some freely available datasets. Terminology differs between studies, but all DAAs share the following characteristics:
    - setting out the terms of ownership of the data, reporting and dissemination of results;
    - committing the data user to ensuring the usage of data is consistent with the terms of the participants' consent;
    - ensuring the data user handles data in accordance with legal and study protocols (e.g., complies with relevant data protection legislation; returns results to study leaders).
    - prohibiting the data user from attempting to re-identify participants.

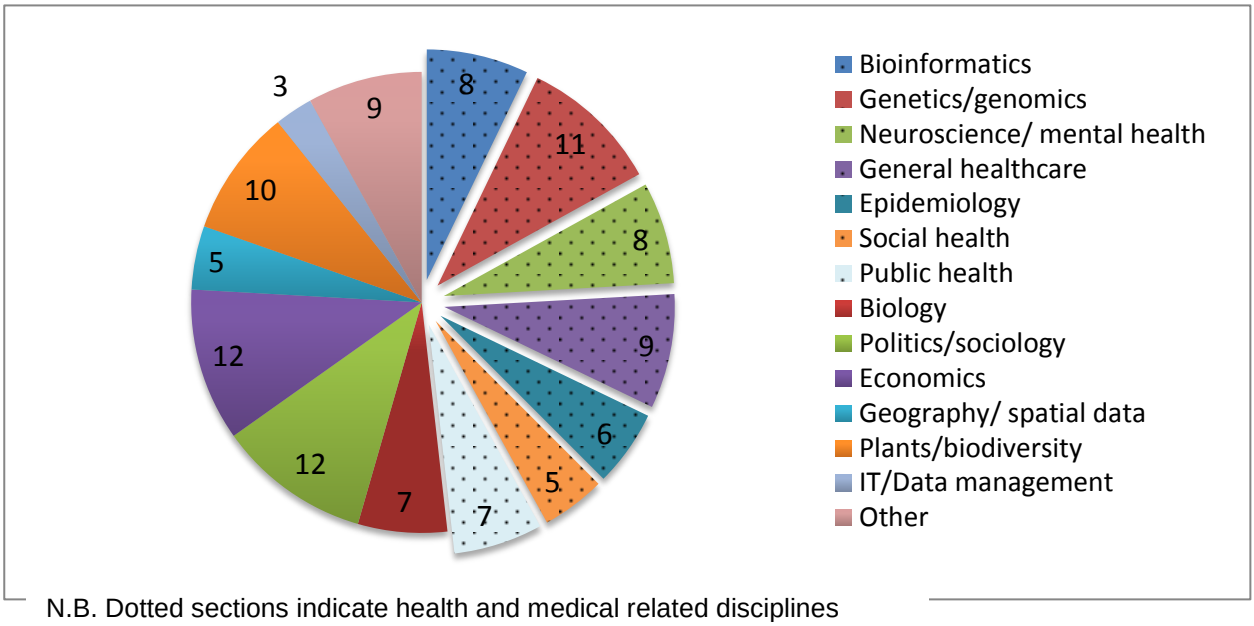
Language used: is there a shared terminology?

33. The language used to describe the processes for data access, the levels of control/restriction and the data use agreements that users must adhere to varies widely across study types. This may mean that data users find it difficult to navigate the range of terminology. For example, access may be restricted to "bona fide", "qualified" researchers, or just "researchers" with an institutional email address. The standard agreements signed may be an "End User Licence", "Data Use Certificate", "Data Access Agreement", "Data Transfer Agreement". The bodies responsible for overseeing access decisions may be termed: "Data Access Committee", "Steering Group" "Management Committee" "Executive Group" or "Co-operative Management Committee" and it is not clear to what extent the functions and terms of reference of these different groups overlap.

ANNEX 5: DATA USER SURVEY

Landscape of data users

1. There were 111 responses to the survey, with most (67%) based in the UK, followed by Europe and the USA. A detailed breakdown of the survey results can be found in **Annex 5a**. A wide range of research fields was represented, including several that were non-health/medical (e.g., economics; politics and sociology) and those that do not involve research using data from human participants (plants and biodiversity).



N.B. Dotted sections indicate health and medical related disciplines

Figure 5-1: Survey respondents grouped by field of research

2. 66% of respondents use research datasets generated by others to compare with or supplement their own data, the remaining third work exclusively on datasets generated by others. There did not appear to be any significant differences in the survey responses by these two categories.

3. More than 40 respondents provided email addresses for further contact, if EAGDA wishes to follow up with a focus group or pose more detailed questions.

Discoverability

4. The majority (85%) use research articles in journals to discover the datasets they use. Websites are a popular source, used by 76% of respondents. Online directories and conferences are also commonly used. In free text responses, informal networking was the most prominent source for discovering datasets, identified by 17% of respondents.

Constraints in discovering data

5. The three most common constraints in discovering datasets were:
   - Lack of information about variables contained in datasets (56%);
     - in the health and medical fields, this was most strongly identified by respondents in bioinformatics and epidemiology.
   - Uncertainties over data quality (53%);
     - majority response in neuroscience/mental health, epidemiology and bioinformatics

- Lack of a centralised directory (46%);
    - majority response in public health, general healthcare, and epidemiology.

Suggestions for improving discoverability

6. Free text responses for suggestions as to how to improve the discoverability of datasets:
    - Centralised repositories or databases;
        - majority response in neuroscience/mental health, epidemiology and public health.
    - Standardised annotation, metadata and documentation of data;
        - majority response in bioinformatics.
    - Better descriptions of datasets and the variables they contain.

7. Other suggestions included: enhancing open access publishing; improving search capabilities; enforcing data sharing policies; creating an international (not just UK/EU) registry of datasets; ensuring data components are machine readable.

Indicative quotes from survey responses

"Not all data is made available during publication. Journals and funding organisations need to enforce the obligation to make data public and in a useable form." (Postdoctoral researcher in bioinformatics)

"A central repository/search interface of databases from projects funded by the EU and other large funding bodies (e.g. Wellcome, RCUK) would be the natural way to go. Don't leave it to Google. The interface should be comparable to PubMed, but for databases." (Postdoctoral researcher in surgery)

"A directory, catalogue or register would indeed be nice. That might also bypass the problem caused by the widely differing websites of different projects which make it variably tedious to find out what is available and how." (Research student in genetics/genomics)

Accessibility

8. The issues identified to the greatest extent as barriers to accessing data were:
    - Availability of information on datasets;
    - Complexity of access procedures
    - Lack of information about how to access data;
    - Constraints on data use.

"There is a total divide between the curation of large scale social science data which is exemplary and easy to get at and the attitudes of some people who hold epidemiological data. The latter take the attitude that the data are their property." (PI in social epidemiology)

"The political and legislative complexity associated with accessing and using biodiversity-related data, coupled with the lack of any standardised approach between countries, is crippling some research activities…" (PI in biodiversity)

"Complexity of access procedures" is a nice way to describe the terrifying amounts of paperwork required to get the NHS and ONS to allow us to obtain and hold medical data... It takes months to sort out the paperwork, and then all the paperwork changes the following year when we want an update of the dataset. It is a massive burden on us and causes huge delays...." (Data manager in environmental epidemiology)

Responses varied significantly across disciplines, and it is notable that 'constraints on data use' affect those not using human data (e.g., 80% of respondents in plants/biodiversity) as well as those whose access may be constrained by consents or privacy concerns for human participants. The UK Data Service was most popularly cited as a good model of data access (n=13).

Timescales for access

9. Of 82 respondents who had to wait after submitting a data access request, the majority (54%, n=44) receive the data within 4 weeks. There was no significant pattern of differences between disciplines, and length of time to access varied widely within disciplines. 6 respondents reported waiting more than 24 weeks to access data.

Suggestions for accelerating access

10. Of 59 respondents who provided a free text answer for what actions funders and study leaders could take to accelerate access to data they use, the most common responses were:
    - mandates/ enforcing data sharing policies;
    - being pro-active in articulating benefits of data sharing;
    - standardizing annotation, metadata and documentation of data.

Adequacy of privacy protection

11. 54% (n=60) of respondents agreed that the access processes in place are appropriate for protecting participant confidentiality. 15% claimed they were not appropriate, while the other 31% said they did not know.
12. 48 respondents provided further reasons for their answer. 3 said that processes are appropriate because "confidentiality needs to be protected"; 3 said that processes are not appropriate for the same reason
13. Overcautiousness by data controllers and burdensome paperwork/bureaucracy were cited as two reasons for believing access processes are inappropriate.

> *"Some, especially NHS, use data protection as a barrier to sharing for the common good."* (Researcher, local community)
>
> *"I believe most of us in academia find our own research question more interesting than the true identity of the individual providing the data such that most will actually not bother to re-identify the individual. On the other hand, these regulations are severely hindering the progress of my work and that of my colleagues."* (Research student in genetics/genomics)
>
> *"Separate mechanisms for individual studies don't enhance participant confidentiality, but do introduce enormous potential for unfair (biased) restrictions on data access etc."* (PI in epidemiology)

Data linking

14. 72% (n=80) of respondent said that their research involved linking different datasets, and the examples given ranged across a wide number of disciplines. Of these 69% (n=55) said they had experienced obstacles in linking datasets. This response was particularly high in geography/spatial data, epidemiology, public health, genetics/genomics and bioinformatics (>60% in each field).

Obstacles

15. The main obstacles experienced by those who provided details (n=49) were:
    - incompatible or inconsistent formats, coding or software between datasets;
    - reluctance by data controllers to allow linking;
    - lack of consent for reuse of data.

The numbers of respondents were too small to break down these responses by discipline.

> *"Data not being made available - lack of willingness from government departments… to help external researchers link survey and administrative data (e.g. I was told I would need an act of parliament to link up such data)."* (Postdoctoral researcher in politics/sociology)

Recommendations

16. 56 respondents provided ideas for changes that would make it easier to conduct research involving data linkage, whilst maintaining necessary safeguards on the data. Most common (n=17) was the suggestion to standardise identifiers and vocabularies for describing datasets. Creating centralised databases or repositories (n=6), standardising/simplifying access procedures (n=6) and generating a cultural shift towards data sharing (n=6) were also suggested.

ANNEX 5a: DATA USER SURVEY DETAILED RESULTS

There were 111 respondents to the survey. The majority (67%, n=74) were based in the UK, with 16% (n=18) in Europe, 11% (n=12) in the USA, and 6% (n=7) from other countries. Most respondents were affiliated to a Higher Education Institution or University (73%, n=83), with others spilt between not-for-profit research institutes or charities (8%, n=9), government, industry and independent status.

47% (n=52) of respondents described themselves as PIs or study leaders; 18% (n=20) as postdoctoral researchers, 13.5% (n=15) as data managers and 9% (n=10) as research students. The 'other' category comprised mainly individuals working in the not-for-profit sector, at research institutes/charities.

Respondents were permitted to enter a free text response to define their field of research. These responses were subsequently grouped into 13 categories, plus one 'other' category for those field containing fewer than three respondents. Just under half (49%, n=54) of respondents were categorised into health and medical-related fields, although it is possible that researchers from other fields may conduct research that involves health data. The majority of categories included research that involved human participant data in some form, although it was not possible from the survey responses to ascertain this precisely for the category of 'biology'.

Although the numbers involved in each discipline were small and therefore lack statistical power, the survey provides a qualitative indication of the kinds of issues faced across different disciplines in discovering, accessing and using shared datasets.

Discoverability

Qs: How do you find out about the datasets used in your research? (tick all that apply)

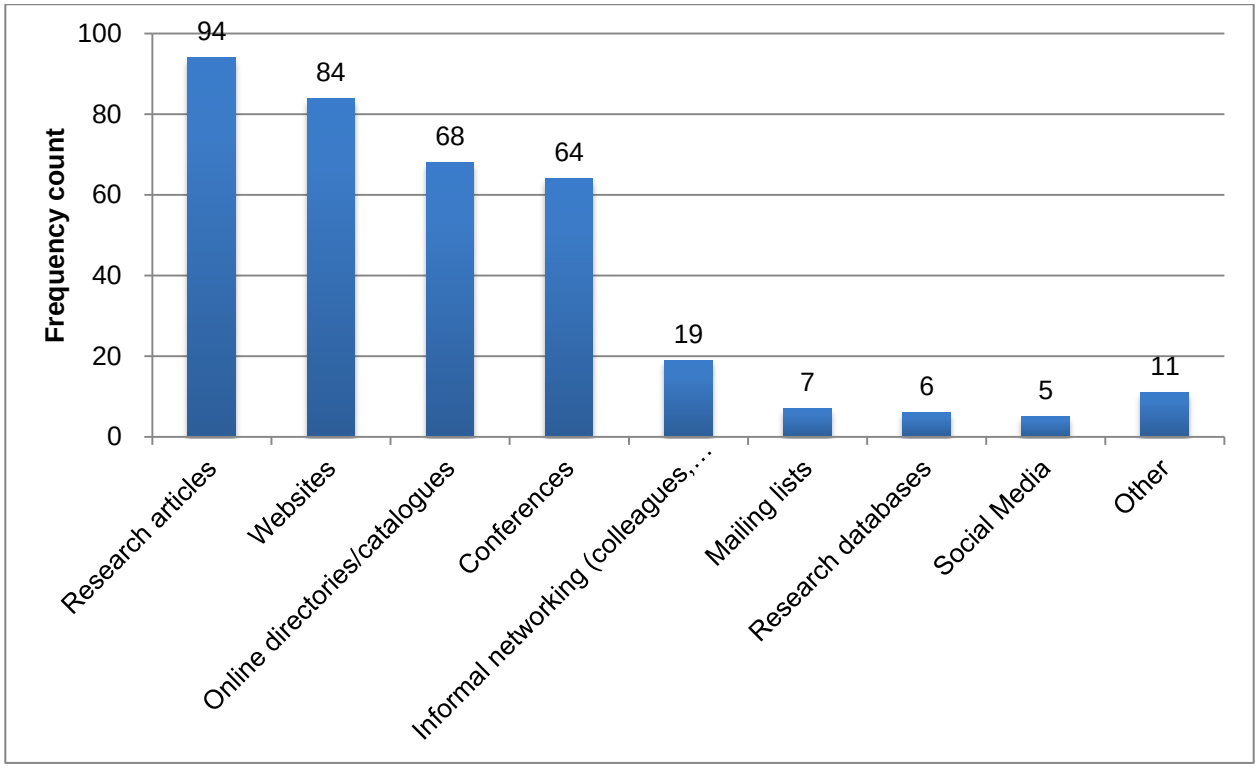


Figure 5a-1: Sources for the discovery of datasets

Qs: What most often constrains you in discovering the datasets you wish to use?

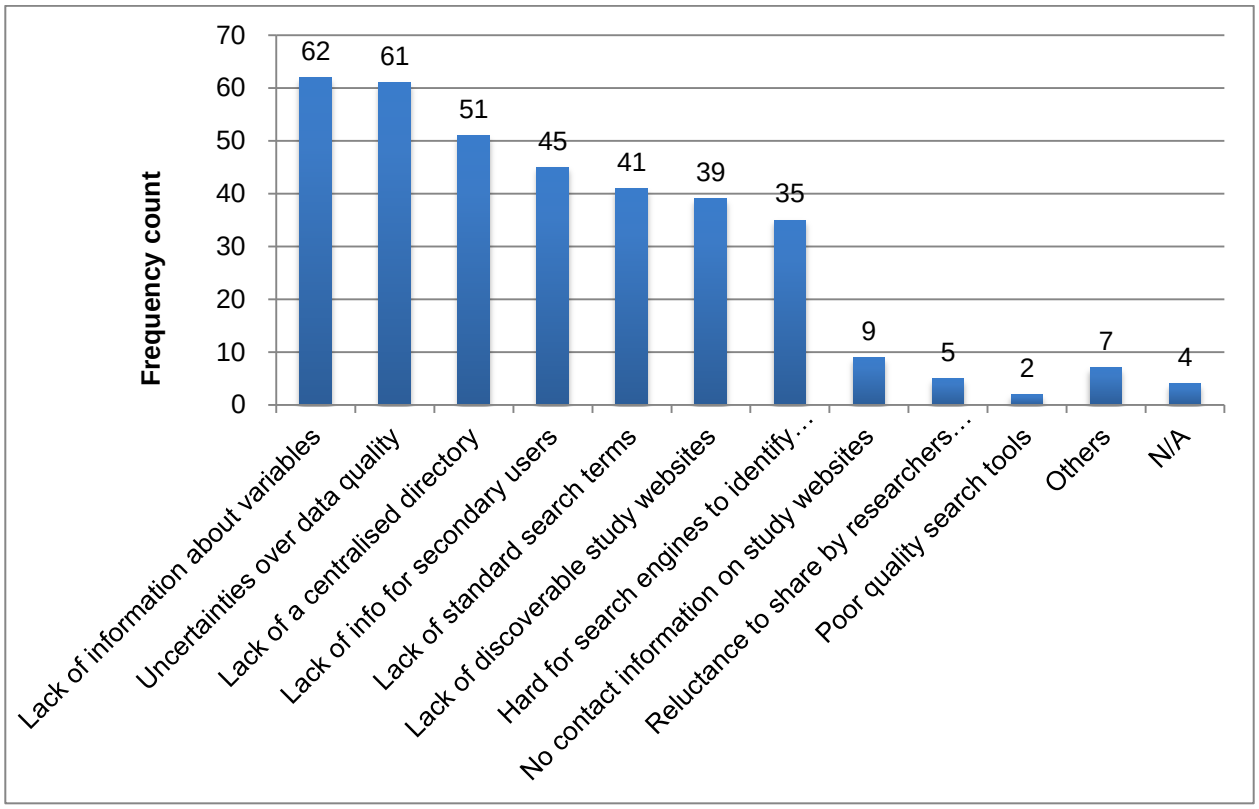


Figure 5a-2: Constraints on discovering datasets

Major constraints on discovering data:

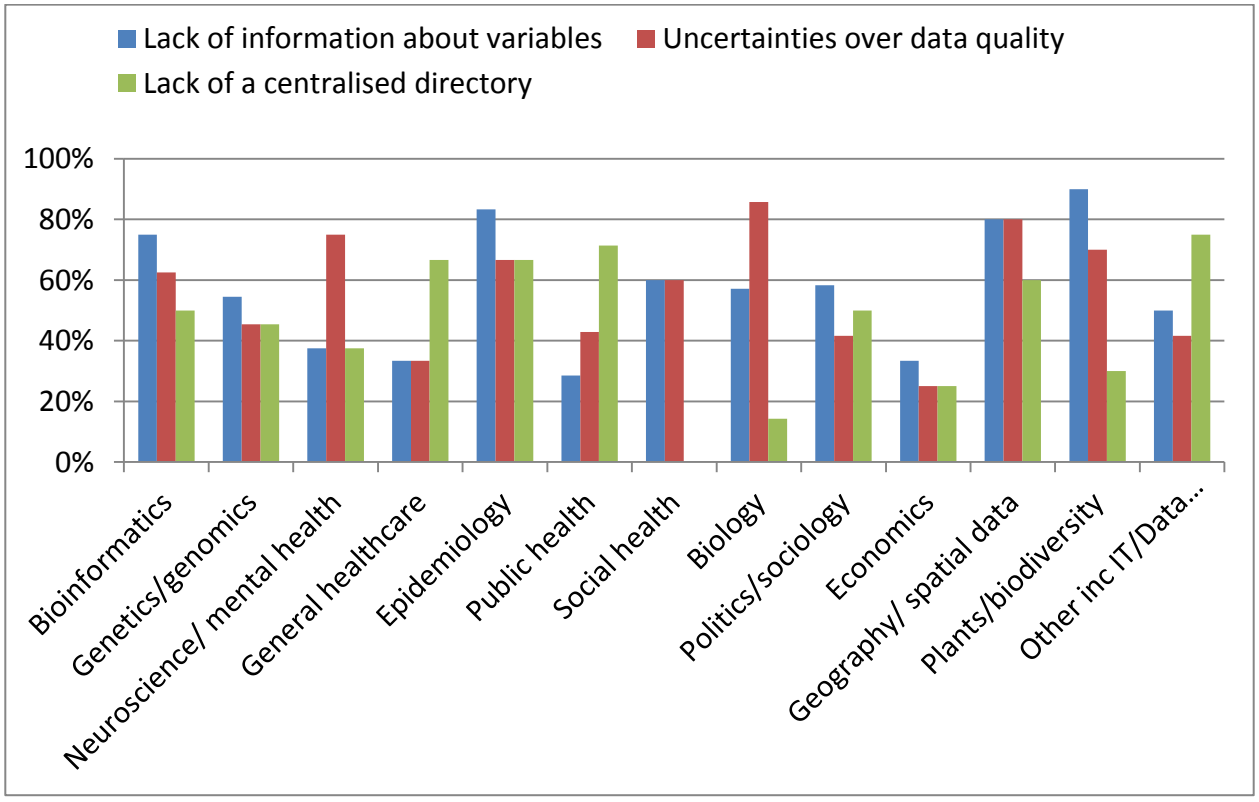


Figure 5a-3: Leading constraints on data discovery, by research field

Qs: What, if anything, would make it easier for you to discover the datasets you wish to use? (free text answer)

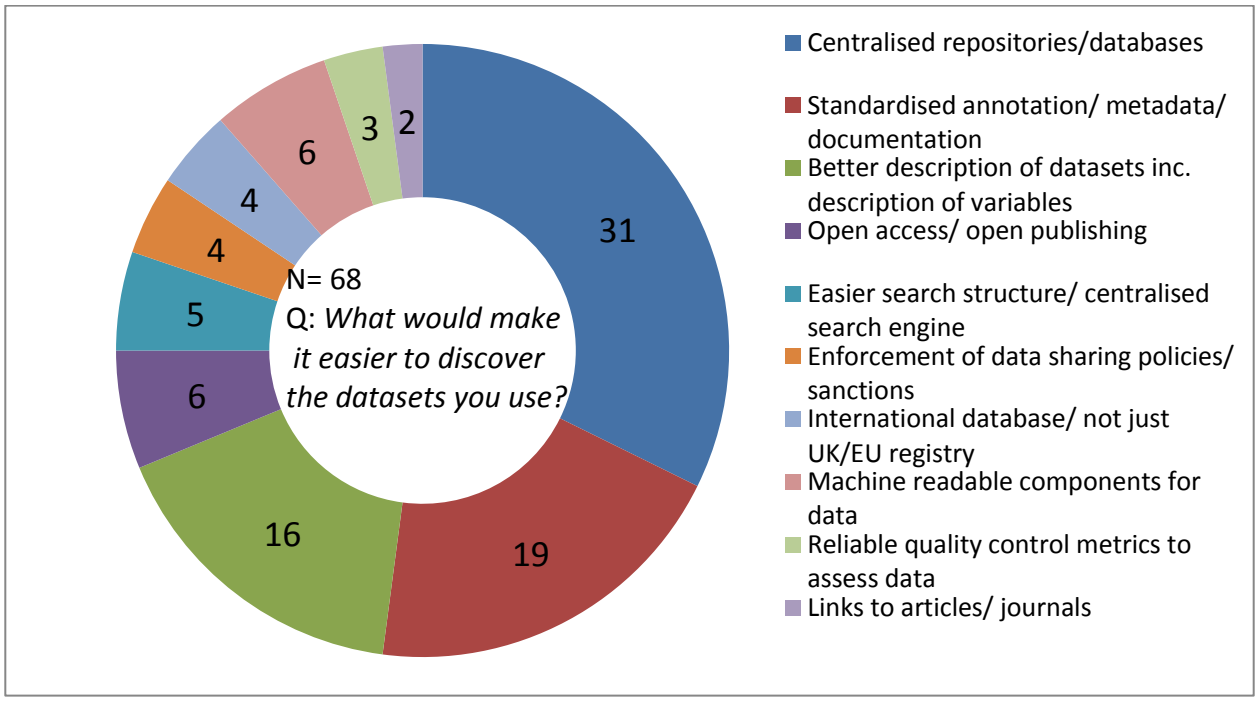


Figure 5a-4: Suggestions for improving discoverability of datasets

Of 68 who provided answers to this question, the suggestion for centralised repositories was made by 4 out of 5 respondents in neuroscience/mental health; 4/6 in epidemiology and 4/6 in public health. The suggestion of standardising annotation/metadata was made by 3/5 respondents in bioinformatics.

Accessibility

Qs: To what extent do you think the following are barriers to accessing the data you need?

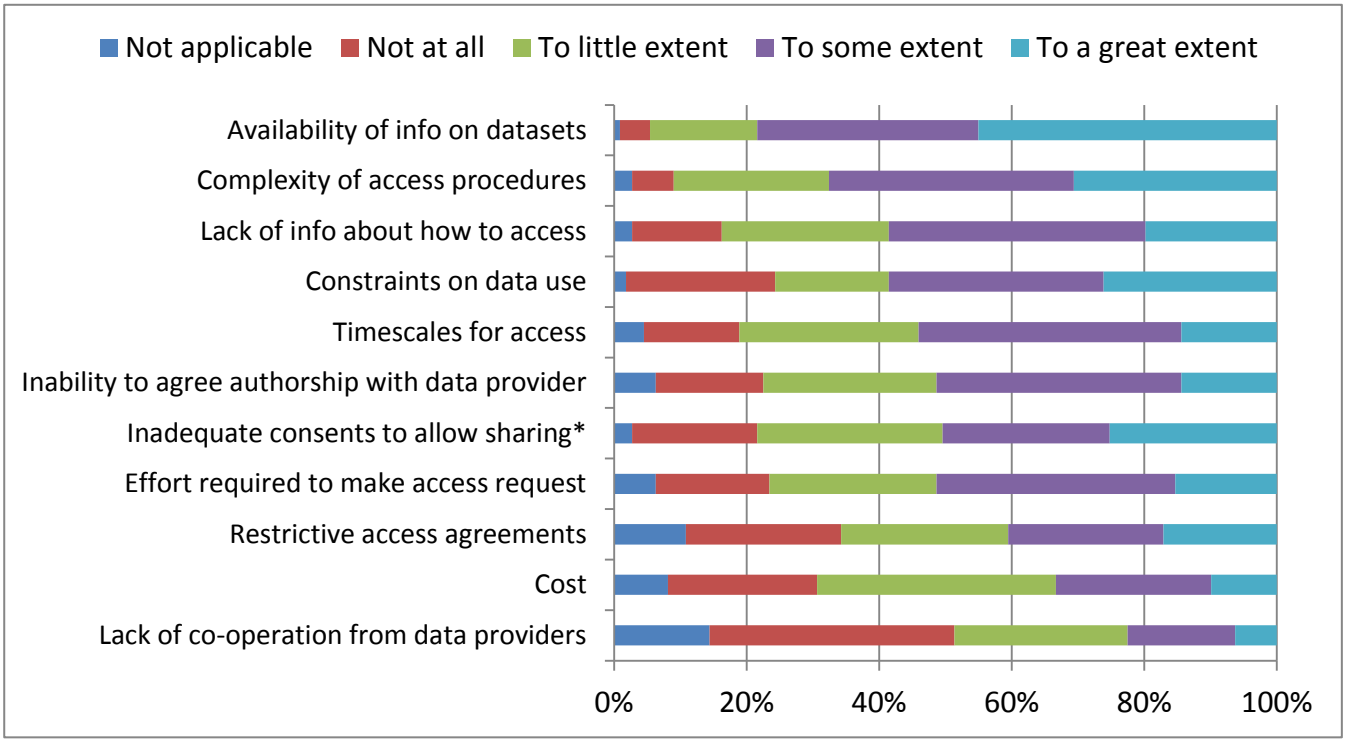


Figure 5a-5: Constraints on access to data

*This question may have been confounded by lack of clarity or confusion over what was meant by 'participant' consent, as 50% of respondents from plant/biodiversity disciplines not involving human participants answered that inadequate participant consents were to some or great extent a barrier to access.

Main constraints on access, divided by research field. Respondents who answered 'to some extent' or 'to a great extent':

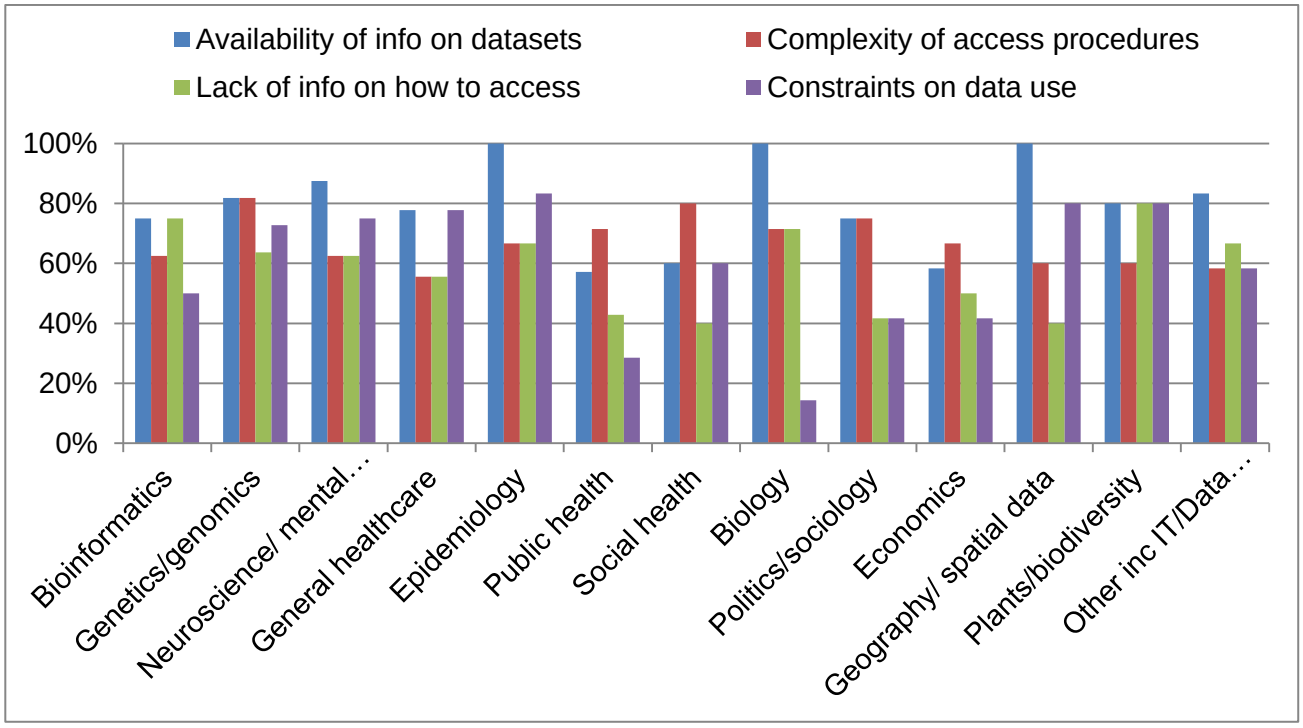


Figure 5a-6: Main constraints on access, by research field

Qs: Are the processes for accessing data whilst protecting participant confidentiality appropriate?

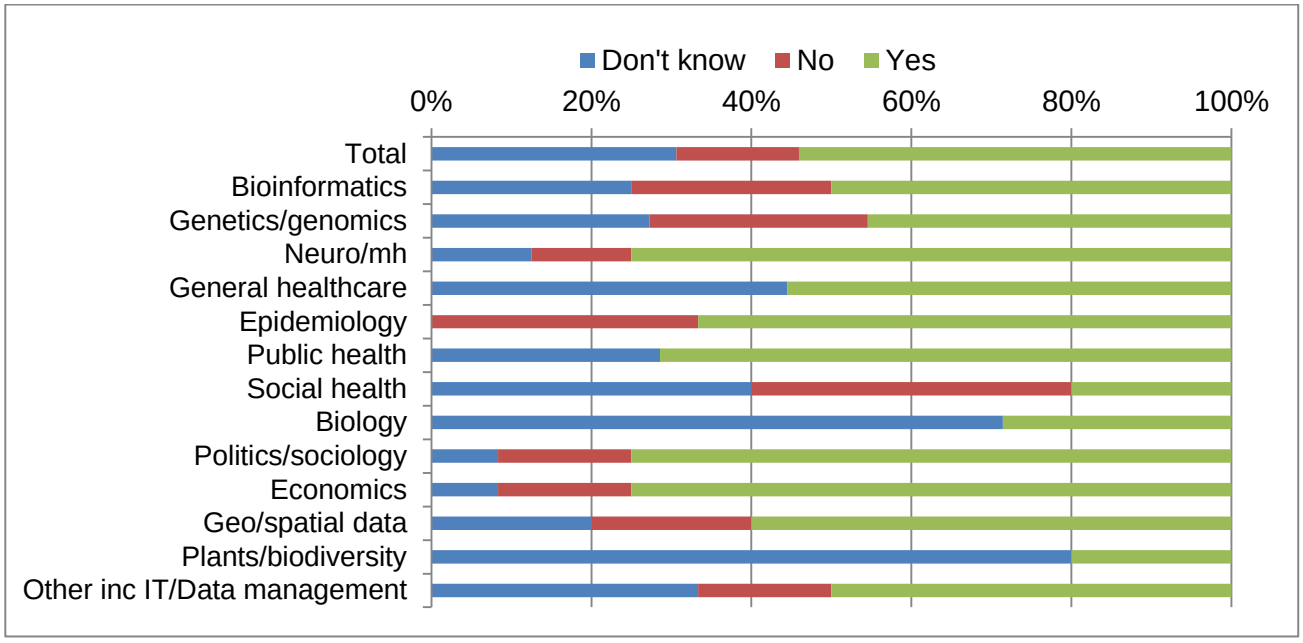


Figure 5a-7: Appropriateness of data access procedures

Qs: In general, how long do you usually wait from the time you submit a data access request until the data are released?

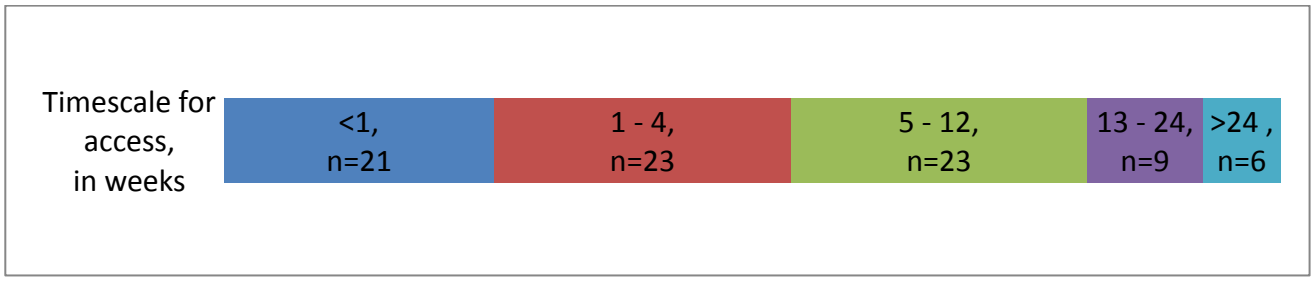


Figure 5a-8: Timescales for accessing data

Data Linking

Of the 80 respondents who reported that their research involves linking different datasets, 69% (n=55) reported experiencing obstacles with linking. This was particularly marked in geography/spatial data (100%, n= 5) and epidemiology (83%, n= 5), with more than 60% of respondents in bioinformatics, genetics/genomics and public health also reporting obstacles.

Qs: what obstacles to lining datasets have you experienced? (Base: n=49)

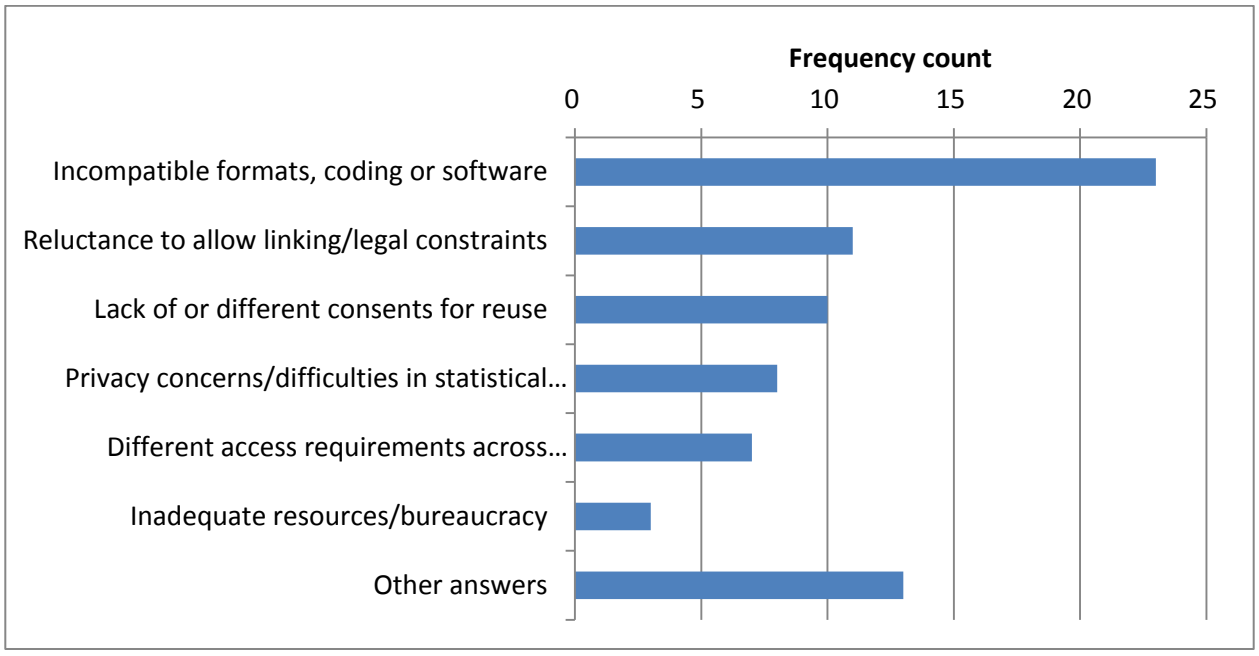


Figure 5a-9: Barriers to linking different datasets

Qs: What key changes, if any, do you think would make it easier to conduct research that involves linking datasets across studies, whilst maintaining necessary safeguards on the use of those data? (Base: n=56)

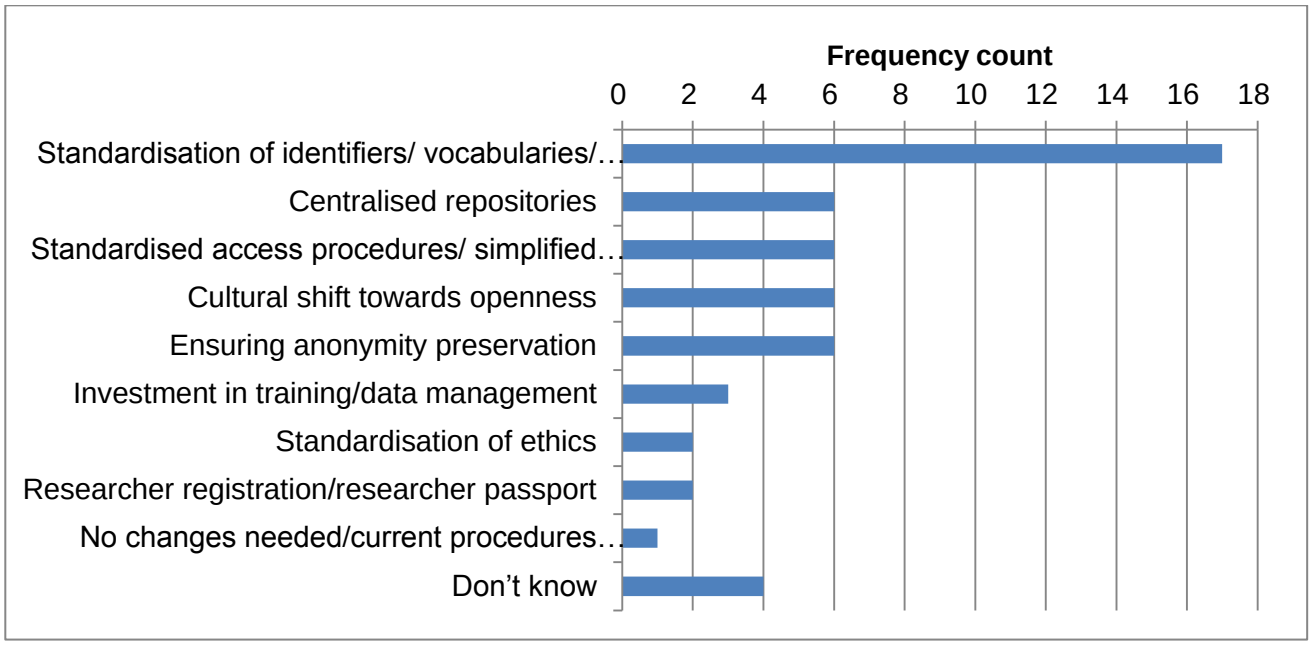


Figure 5a-10: Suggestions for making it easier to conduct research that links datasets

The majority of respondents who had used shared data had taken steps to improve its quality (n=91), with 72% (n=80) 'cleaning' the data; 68% (n=75) labelling data for clarity or consistency, and 43% (n=48) improving the metadata. However, only 41% of those who had improved the data had returned these improvements to the study leader or data producer.

This trend was strong (i.e., fewer than 25% returned results) in biology, politics/sociology, public health, and genetics/genomics.

ANNEX 6: DATA ACCESS INTERVIEWS WITH COHORT LEADERS

1. In August 2014, the EAGDA funders identified 16 cohort study leaders to approach for interview, representing a range of study types and sizes, at varying stages of their grants, and using a number of different approaches to data access. Every investigator contacted responded positively to the request.
2. Cohort study leaders were contacted and asked to participate in a 30 minute interview. The purpose of the interviews was to clarify data access mechanisms for their respective studies, ascertain the costs of data access if feasible, and to identify any particular areas in which they felt funder support or guidance would be beneficial.
3. Interviews were conducted in August-September 2014 for the most part over the phone, with one face-to-face interview, using a semi-structured approach. A set of key questions was circulated to interviewees in advance as a starting point for the discussions. During the interviews, notes were taken, written up and fed back to the interviewees who could then correct any errors or misrepresentations of their views. A thematic analysis was undertaken on the resulting interview notes, drawing on EAGDA's previous discussions regarding some of the key themes and issues in data access already identified.
4. The summary of views presented here focuses on responses to questions concerning data access and funder roles in supporting this. Owing to EAGDA's additional interest in potentially exploring sustainable sample access strategies interviewees were also asked about their processes for sample sharing.
5. The following questions of relevance to this particular report were sent to interviewees in advance as a basis for discussion:
    1. Do you have a data access system which allows researchers from outside your own study to access detailed study data? Could you outline how and why the system has been set up the way it has?
    2. What are the costs associated with setting up and maintaining data access (in terms of staff resources, administrative burden, actual costs of providing data)?
    3. What more could funders do to support data and/or sample access mechanisms for cohort studies?
6. The short time available for the interviews entailed that some of the questions could not be fully explored in depth, and the interviewer therefore sought to gain an overarching understanding of the data access setup for each study.
7. In addition, interviewees were asked to provide information, if available, in answer to the following questions:

    - On average how many data access requests do you receive per month?
    - What is the usual timescale from receipt of a data access request to release of/access to the data?
    - What proportion (if any) of access requests received are declined?
    - What are the usual reasons for declining an access request?
    - Have you had any issues with data users breaching the terms of an access agreement (if this is monitored)?

8. Not all studies were able to provide these data so the summary presented below is necessarily an incomplete picture of the landscape of secondary data access requests and use.

SUMMARY OF KEY THEMES

9. Several key themes emerged during the interviews, which cross-cut the questions and were fairly consistently raised by most of the interviewees. These are described in more depth in paragraphs 42-end:

- **Value of data sharing** – allowing access to data is widely recognised as an important function for cohort studies, and interviewees were largely positive about the value of sharing data. This supports the findings of the EAGDA *Incentives*[21] report.
- **Varied nature of cohorts and access mechanisms** – cohorts have different characteristics and there is a range of different setups for data access. The studies explicitly set up as resources for the scientific community have the most well-established systems of governance; others have systems that have evolved over time as the studies have become established.
- **Collaboration as a mechanism for data sharing** – several studies primarily share data through collaborations with other research teams, rather than allow data to be "handed out" to any qualified researcher. This ensures that participants and the study reputation can be protected, and allows the study staff, who have expertise and detailed understanding of the datasets, to valuably contribute to any further uses of the data. There are a range of models for collaboration, but collaborations of this sort often mean that the study PI retains control over who has access; and/or that the PI can expect to be awarded authorship in papers even if they contribute little; and/or that the data are used only to further the interests of the study itself (which is understandable when further funding decisions are made on the basis of the scientific quality of the study team's outputs). This can mean that datasets are not able to be repurposed to explore wider hypotheses beyond the interests of the study team.
- **Tensions between different demands** – different policies and requirements investigators need to adhere to lead to some tensions with their efforts to act in the best interests of their studies.
- **Costs, resources and capacity** – costs are not generally calculated specifically for supporting data access and in many cases access activities are undertaken within study staff's day to day roles on an *ad hoc* basis. Nonetheless, it is commonly felt that there is inadequate resourcing in terms of staff time and capacity for supporting data access.
- **Standards and formats** – lack of a common approach to data and metadata standards limits the utility of some datasets for further use, particularly between disciplines.

Question 1 part 1: Summary of data access mechanisms

10. Fifteen of the sixteen studies have a mechanism in place for allowing secondary users to apply for access to the study's datasets, with varying degrees of formalisation of the processes. These range from a study steering committee, responsible for the general management and oversight of the study, reviewing access requests on an *ad hoc* basis, to a specifically and formally constituted data access committee (DAC) with members independent of the study. In the remaining case, data are widely shared within an

[21] EAGDA's report on Establishing Incentives, available at: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/WTP056496.htm

established collaboration but there is not a formal mechanism for users beyond that collaboration to seek to access the highly specialised datasets.

11. In ten of studies that require an access decision to be made, prospective users informally contact the PI or a member of the study team in the first instance, to discuss their request and ascertain if the study can feasibly supply the datasets they require.

12. Studies are split between those who routinely use formal committees to review and assess all access requests, and those who revert to these committees only in the case of complex requests, depletable sample requests or those with potential ethical complications (such as access to linked genotype and phenotype data). For straightforward data requests in six out of the sixteen studies, access decisions are handled by the PI alone or with co-investigators, drawing on additional expertise if needed.

13. Four studies have access to a further independent committee to oversee appeals or disputes, although anecdotally appeals are rare.

Issues assessed by data access committee/governance mechanism

14. There is a range of different mechanisms for assessing requests to access data across the cohorts, and different issues are considered by the PI or committee. Making decisions on access requests involves assessing any or all of the following issues:

    - Feasibility of request (availability of data; staff capacity to fulfil);
    - Risks to participants (disclosure risks; protect ion of confidentiality; adherence to consents);
    - Whether access request would lead to duplication or conflict with other current studies;
    - Whether the proposal furthers the aims of the original study;
    - Whether regulatory and legal requirements will be adhered to;
    - Credentials of researchers (ensuring only *bona fide* researchers have access);
    - Whether access would lead to high quality science and thus protect the reputation of the study.

For long running cohorts that maintain contact with their participants and return to them for further data collection over time, the ethical issues appear to predominate in the decision-making.

Requirements on data users

15. Several interviewees outlined the requirements they place on users of data. Where a member of the study team has been heavily involved in managing the datasets, studies generally indicated that co-authorship of any resultant publications would be a requirement of allowing access.

16. In the case of TwinsUK, all papers must be submitted for approval by the Steering Committee prior to publication to check the acknowledgements. For other studies, the requirement for acknowledgment is more informal. In some cases, it is requested that papers are submitted to the steering or management committee prior to journal submission in order to ensure there is no risk of participants being re-identified from the information provided in the paper and to ensure the data are not being misrepresented

Question 1 part 2: Rationale for data access set up

17. Each cohort had its own rationale for setting up data access mechanisms and governance in the way that they had, but there was a great deal of informal learning and discussion

between studies in developing their access mechanisms based on others. For example, Life Study is modelling its data access and governance to some extent on UK Biobank.

18. There were several common reasons underpinning these respective approaches:

- **Ethical** – several studies indicated that their steering or data access committees considered ethical issues as part of the assessment of data access requests. The types of ethical issues raised included:
  - Re-identification risk and ensuring the confidentiality of participants would be protected. There was some uncertainty expressed in one interview over how to ascertain identity disclosure risks, with another indicating that in studies involving genomic data this was an increasingly salient issue.
  - Ensuring participants were not unduly burdened by additional requests for data collection.
  - Checking whether the proposed uses conformed to the terms of participants' consents (especially with respect to commercial usage).
  - An ethical imperative to ensure the interests of the study and the participants would be best promoted.
- **Scientific** – many access committees based the rationale for their data access set up on the need to ensure high quality science resulting from the use of their data. This includes the need to:
  - Ensure data are not misused or misrepresented;
  - Protect the reputation of the study;
  - Reduce the risk of duplications of research or conflicts in using the same datasets;
  - Ensure data users respect the privileged nature of access to the resource. Some data dictionaries or metadata lists are not openly available and searchable because PIs prefer to allow access only once they have established contact with a prospective data user.
- **Legal** – in some cases the legal requirements on data owners and institutional uncertainties can create complexities in the sign off process for MTAs or DAAs between the sharing institutions, and this can limit what data are accessible given the resources available to process requests.
- **Resources** – some data requests require significant time and expertise for study staff to extract, format or create specific datasets as needed. Some access decisions are therefore based on the feasibility of requests given the constraints of time and resources that would be required to fulfil them.

19. Decisions over what kind of data access to permit, whether open, controlled, restricted to certain academic partners (such as consortium members) or primarily through collaborations, and what governance is appropriate in each case, are largely based on balancing these considerations.

Question 2: Costs

20. Specific breakdowns of costs were difficult to obtain for setting up and maintaining data access mechanisms in most of the studies. This may have been due to the nature of the interview methodology and short timescale for discussion, which precluded in-depth questioning about resource allocation and costings.

21. The majority of interviewees indicated that not having a clear sense of costs was in part because funding requirements do not stipulate costings for data access and these activities are expected to be included in core costs. One implication of this is that it is extremely difficult to quantify what additional resources may be needed to better equip studies for wider data access, in the short-term (setting up access mechanisms), medium-term (maintaining data access without overburdening study team resources) and long- term (sustainable access to data beyond the life of the study grant).
22. The interviews revealed three distinct areas in which costs involved in enabling data access for other users are usually incurred:
    - **Data formatting and management**: cleaning and preparing data; formatting and documenting metadata for submission to a repository; creating data dictionaries; understanding user requirements and extracting data to specification.
    - **Governance mechanisms:** decisions on some datasets are made by a committee, which incur costs in terms of staff time and administrative resources required to run the committee. In cases in which the study PI makes decisions with some technical or ethical support as needed, investigator time is the primary cost.
    - **Data administration:** handling incoming requests, processing any access agreements as required; sending out data or liaising with repository to authorise access.
23. The majority of interviewees stressed that the major costs of data access concerned the staff time and expertise required to discuss user requirements and extract the relevant variables as required. Some requests are complex and require substantial time, an intricate grasp of what the datasets can and cannot do, and in some cases sophisticated statistical expertise to provide the required data. Resources such as those being developed by CLOSER[22] to improve metadata documentation were cited as extremely valuable in helping cost efficiency.
24. For studies that use a repository such as the UK Data Service or EGA, the administrative process is largely outsourced to those services. It is only the governance (for controlled access datasets) and initial formatting that needs to be covered by core costs in these cases. However, whilst using a repository mitigates the cost issue for researchers, it remains a significant issue for funders as the repositories need to be well-resourced, sustainably, if they are to provide storage and curation facilities for studies.
25. Two studies currently use a cost recovery model for data access:
    - ALSPAC (cost determined on a case-by-case basis, reflecting the true cost to ALSPAC of providing the resources required[23]);
    - Born in Bradford (charging a fixed fee of £900 per access request, calculated based on the cost of the average number of hours required to extract data).
    - A third study, ELSA, is currently considering charging £800 for up to 40 phenotypic variables.
26. Cost recovery can ensure that there is sufficient resourcing for managing data access requests that is not contingent on core grants and is therefore more sustainable in the short-

[22] www.closer.ac.uk
[23] http://www.bristol.ac.uk/media-library/sites/alspac/documents/research/Access%20Policy_v6.0.pdf Section 1.3

medium term. However, it is a controversial model and several interviewees indicated that they would prefer not to charge users if at all possible.

27. ALSPAC began operating on this model in April 2014. It is too early to formally evaluate whether the switch to charging users has resulted in a decrease in the number of data requests, although it is suspected that there may have been a slightly negative impact.

28. Cost recovery based on fixed fees does not necessarily cover need for the expertise of data managers with the interdisciplinary skills required to handle requests from different data users. Such managers become more valuable over time as they gain expertise and knowledge of the datasets.

Question 3: Views on funder actions/support

29. Several themes emerged from the interviews in relation to issues the funders should consider and actions they could take to support data access in the studies they fund. These are divided below into issues concerning grant applications and funder policies, infrastructure, co-ordination and culture.

Grant applications and funder policies

30. Grant applications focus on fieldwork, data collection and publication, not data preparation for further use. If data sharing is to be embedded in the culture of scientific research, funders should consider ensuring there are realistic assessments of data sharing and management plans as part of their funding decisions, and due reward and recognition for those who make data accessible. Plans should take into account the methodological process of cleaning and preparing data to make them useable for others and build these considerations into both the costs and timeframes for grants.

31. Clarity over what funder policies require would be welcomed: study leaders need to know what is expected of them with regard to data access when studies are setting up, with guidance on what it takes to create data resources, what governance requirements should be in place, how much time should be devoted to these activities and what expertise are needed to maintain data access.

32. If funders want studies to improve and enhance data access, PIs and study teams need to be empowered, with recognition and credit for their leadership and knowledge. Any push for independent oversight of access governance needs to be balanced with practicality and common sense. One participant indicated that the focus of governance should be on transparency of decision-making, not on taking access decisions out of the hands of PIs entirely, as this can be disempowering.

33. Funders could also support innovative collaborative ideas to add value for bringing data resources together, e.g., through a special funding stream or call. Studies need both financial support and access to expertise to support data sharing.

Infrastructure

34. Infrastructural support is needed to develop tools such as data dictionaries, online access request systems and metadata documentation tools as these are time consuming and complex to create.

35. Data management and sharing is only going to get more technologically complex. Funders ought to be looking to the future and supporting forward-looking initiatives (such as DataSHIELD) to enable the research community to stay abreast of these developments.

Co-ordination

36. It would be extremely valuable to the research community if consensus among major funders could be reached on issues such as the implications of cost recovery models and long-term sustainability for data and sample access.
37. Given the number of studies that interact with NHS resources and ethics committees, co-ordination or harmonisation of governance mechanisms and processes between funders, institutions and the NHS would significantly speed up access to healthcare data used in research.
38. Administration of data access can take up a lot of time for study teams. Centralised repositories such as the EGA and UK Data Service can markedly improve efficiency by handling administration and ensuring studies only deal with complex requests of those requiring particular controls. Could funders create or support a more centralised administrative system for data access?
39. Universally recognised mechanisms, such as Digital Object Identifiers (DOIs)[24], could be gainfully used to track dataset usage and citation, which would reduce administrative burdens on study teams to keep track themselves of how their datasets are used.

Culture

40. Several interviewees felt that funders were well placed to increase the prominence and value of data and data sharing in the research community and among institutions.
41. It was noted that institutional buy-in will be needed for recognition of the valuable work data managers do even though these are often considered to be non-academic roles. The development of cohort management and leadership skills will need to be supported by institutions.

DETAILED ANALYSIS OF KEY THEMES

42. The key themes outlined in paragraph 9 are described in greater depth here, to provide a sense of the range and diversity of opinions expressed by the interviewees.

The value of sharing data

43. All interviewees recognised the value and importance of sharing data, with several of the larger studies being set up or developed specifically as resources for the wider scientific community. In general, enabling datasets to be further used beyond the study team is perceived as an important part of the scientific mission of the studies. However, there was much variation in attitudes towards how access to data could and should be managed.

[24] DOIs are unique alphanumeric strings that identify content and provide a persistent link to an electronic object, such as a document, paper or dataset. They will not expire or become outdated like a URL. DOIs are assigned by a central registration agency, the International DOI Foundation www.doi.org

44. The majority of studies are established contributors to consortia or academic collaborations that pool datasets from across different studies with similar or complementary interests. Within these communities, sharing of data is either a formal part of the relationship between groups (for example, the IARC consortium of which EPIC Norfolk and EPIC Oxford are part), or undertaken more informally on a *quid pro quo* basis – this sharing can be either based on a common disease area or data type. Thus access to data beyond the members of study teams is well-entrenched in many cohort studies.

45. There was a strong message from several interviewees that maximising the value of datasets (a key tenet of many funder policies) does not necessarily mean sharing data as widely as possible in practice. The reasons for these views are detailed below.

Varied nature of cohorts and data access

46. For most cohorts that have been established for some years, their approaches to data access and sharing have evolved over time, in response to changes in scientific culture, funder policies and increasing moves towards collaboration between groups. The majority of these studies have been funded by a succession of short-term project or programme grants, which have been subject to competitive grant applications. One implication of this mode of development over time has been that there has been little or no specific funding for supporting data access in most cases, as investigators have needed to focus on the aspects of their studies that will achieve the greatest credit and recognition from both funders and their institutions.

47. It was stressed by several interviewees that cohort studies vary enormously in their local characteristics, and they indicated that a “one size fits all” approach to guidance or policies on data sharing and access would not be appropriate across the types of cohorts supported by the EAGDA funders. It was argued that there are legitimate reasons for studies having different governance arrangements from one another, and furthermore that burdening studies with prescriptive regulation or policies could be detrimental to research.

Benefits of collaboration versus open/wide sharing

48. For several studies, the preferred mode of data sharing was through collaborations between the study team and other data users, in most cases exclusively other *bona fide* researchers. There were several benefits to these arrangements cited by interviewees:

    a. Some types of data, even when well-documented, require a thorough understanding of their history and context if they are to be interpreted accurately and not misrepresented in publications. Allowing secondary users to work with a member of the study team is often considered the best way to enable this.

    b. Collaboration allows the use of data to be easily tracked, through co-authored publications and the PI’s awareness of who is using the datasets for what purpose. This enables studies clearly to demonstrate both impact and publication outputs generated, for use in future grant applications. The development of citation metrics for datasets could greatly improve studies’ capacity to track how their data is used without requiring the formalities of collaborative work.

    c. Collaborations also ensure the PI can remain accountable to the study participants for how the data are being used, and can help maintain the study’s reputation for

undertaking high quality science. It was noted by several interviewees that the UK's cohorts are frequently considered world-leading.

d. Collaboration enables the study team to gain credit for their time, effort and original research using the study datasets. This provides strong positive motivation for the researchers, which can be undermined if data are made available to any qualified user without adequate recognition for the data producers and curators.

Tensions between different demands

49. In seeking to understand the rationale underlying the particular set up of data access mechanisms for cohort studies, several tensions emerged in the demands placed on interviewees. Firstly, a tension was felt between funder policies regarding data access and the value they, and institutions, appear to place on these activities in practice:

    a. Some interviewees mentioned that although funders require data management and sharing plans, the focus of assessing grant applications competitively is on the core functions of data collection, conducting original research and producing publications. Data sharing and management plans do not appear to be given much weight or credit, which means that investigators may not specifically plan or put in place costs for enabling data access.

    b. Developing the infrastructure and maintenance of datasets to support data access are time and resource intensive exercises, requiring significant data management expertise and long-term forward planning. In most cases (with the exception of grants specifically to support data access) data sharing activities are supposed to be covered by core costs of short-term grants. This does not allow for developing strategies for long-term curation and access mechanisms.

    c. Institutions do not give academic credit for data management or for the research leadership skills required to manage a cohort as a resource for secondary data use. The issues raised by interviewees with regard to recognition and credit closely mirror those raised in the previous EAGDA report on 'Establishing incentives for data sharing.'

50. There was also a tension between the cultural shift towards wider, more open sharing of data and the duty of responsibility interviewees felt both towards their study and towards participants.

    a. With regard to the study, several interviewees emphasised their obligation to protect the reputation of the study and to ensure that the datasets were used for high-quality science. This responsibility influences the decision-making process for access requests, and this is seen as both an ethical responsibility and a scientific one: the complexity of some datasets means that it would be irresponsible to send out data to a user unfamiliar with the methodology, context and history of the data collection, as they may misrepresent or misinterpret the data as a result.

    b. With regard to participants, many cohort leaders have worked with participants over a period of time and feel a strong duty towards them to ensure that their commitment and generosity in contributing to the study results in quality science and good advances in scientific understanding. Again, this often leads to a higher threshold for allowing access to the data and a degree of peer review for access requests.

    c. There were also historical consent issues in some cohorts, as the terms of the original consent, taken at a time when data sharing was not the norm in scientific research, did

not allow for use of the data beyond the study team. There is therefore an extra onus on study leaders to ensure data are used responsibly, to know who is using the data and how, to have records of its use, and to ensure the data are valued and stored securely. Consents in some cases limit further use to collaborative work in which the study team has a major role.

d. Pressure from funders to release datasets at an early stage also generated concerns regarding the potentially negative impact on motivation of investigators: investigators devote substantial time and intellectual effort to generating data and seeking to explore hypotheses using their datasets, and it was felt by some interviewees that early release of data in the interest of maximising their use could jeopardise investigators' own interests. Several interviewees indicated that his may result in putting investigators off running such studies in the longer term if they are not given the best opportunities to exploit the resource they dedicate themselves to creating.

51. A further tension was evident in the way that the requirement to "maximise the value of data" generated through public or charity funding should be interpreted. Several of the longer established cohorts in particular expressed the view that maximising value does not necessarily entail sharing data as widely as possible, as this does not always result in better science. For certain types of data an in-depth understanding of the methodology and rationale of collection is required before the data can be interpreted thoroughly and accurately. If data users do not spend the time carefully considering the questions they want to address, this can lead to poor science.

52. Another aspect of this tension was particularly relevant to observational data. Observational data are not definitive: the expertise of study team may be vital for interpreting the data and providing essential context. Additionally, early data releases could generate a tension between potential research benefits versus scientific quality, as conflicting results can sometimes be obtained.

53. There was also some evidence of difficulties in enabling data (and sample) sharing owing to the legal requirements of institutions rendering various aspects of access laborious and complex. Difficulties with establishing Material/Data Transfer Agreements, information governance requirements, questions of data ownership, questions over which policy (funder, institution, local ethics committee) trumps others, and responsibility for storage and curation of samples and/or data have all either created challenges between studies and their host institutions, or are anticipated to in future.

Resources and capacity

54. A recurring theme throughout the interviews concerned the need to adequately resource data access and sharing activities. There was an overall feeling conveyed that funders did not sufficiently recognise the staff time, resources and expertise that are required to create and maintain datasets for secondary use.

55. For many datasets, data managers are required to perform extractions of specific variables in response to the requirements of each access request and it would not be feasible to simply deposit entire datasets in a repository and allow users to browse for the relevant variables. Understanding user requirements and extracting, formatting, annotating and cleaning data to maximise its utility and value to other users can take up a significant proportion of staff time. These efforts are often considered to be activities that should be undertaken as "part of the day job" yet are rarely recognised or credited academically.

56. As data access activities are subsumed into general staff time and not specifically credited, it was felt by some interviewees that data access is supported primarily through general goodwill of staff, which is not a sustainable basis for maintaining access.

57. This concern over staff capacity and credit was related to the preference by some interviewees for sharing data via collaborations rather than simply giving access to datasets to secondary users: it was perceived as worthwhile activity to share data through a collaboration, as academic credit through paper authorships and a deepening of the study team's understanding of the data can be obtained through this mode of data sharing.

Standards and formats

58. It was felt by one interviewee that a lack of guidance on standards for metadata and file formats leads to uncertainty for the study teams over what the most useful and interpretable formats for data would be. A further two interviewees indicated that the sheer quantities of data being produced meant that it could be difficult to establish how the data should best be formatted and interpreted for secondary use.

59. Strategic co-ordination and cross-funder guidance on how data should be quantified and processed was perceived as a potential route through which to make data more available and useful to secondary users.

Interviewees:

Name of interviewee	Study
Alissa Goodman	NCDS/1958 Birth Cohort
Andrew Steptoe	English Longitudinal Study of Ageing
Anthony Swerdlow	Breakthrough Generations
Cyrus Cooper	Southampton Women's Survey; Hertfordshire Cohort Study
Dan Mason (Data manager)	Born in Bradford
Carol Desateux, Rachel Knowles	Life Study
Diana Kuh	National Survey of Health and Development/1946 Birth cohort
Doug Easton	EMBRACE
Deborah Hart, Chris Hammond	Twins UK
Ian Deary	Lothian Birth Cohort
Jane Green	Million Women Study
Kay Tee Khaw	EPIC Norfolk
Lynn Molloy	ALSPAC
Nick Buck	Understanding Society
Peter Whincup	British Regional Heart Study
Tim Key	EPIC Oxford

ANNEX 7: SUMMARY OF FUNDER DATA SHARING POLICIES (2014)

The EAGDA funders have data access policies that promote data sharing and require funded researchers to plan for how the value of their data can be maximised. The funders differ significantly in the level of guidance they provide on how researchers should ensure these policies are adhered to. The policies are summarised below.

Funder	Details of data sharing policy
CRUK	- Data arising from CRUK research should be made as widely and freely available as possible to maximise public benefit. - All grant applications must include a data sharing and management plan. - Period of exclusivity can be negotiated, to protect patents and IP rights. - Appropriate method of data sharing contingent on type, size, complexity and sensitivity of data. - No specific guidance on how researchers should preserve and share data.
ESRC	- Overarching commitment to long-term preservation of data, high quality management and strengthening provision for secondary data use. - Policy builds on OECD key principles[25] that publicly funded research, produced in the public interest, should be openly available to the maximum extent possible. - All grant applications must include a data sharing and management plan. - Data should be available for preparation for reuse or archiving within an ESRC data service provider within 3 months of the end of a grant award. - Sensitive and confidential data can be shared ethically, providing researchers anticipate data sharing in their research plans from an early stage, e.g., including sharing in consent processes, ensuring data can be anonymised, building in access restrictions.
MRC	- Actively promotes research collaboration and data-sharing, emphasising how the value of data can be increased throughout 'data lifecycle', and maximised for the public good. - Policy based on OECD key principles applicable to publicly funded research, and RCUK common principles on Data Policy[26]. - Very detailed data sharing and access policy, detailing requirements and expectations on studies for: discoverability of studies and datasets; proportionality of access mechanisms; data governance; prohibitions on re-identification; participant consent and transparency in criteria for access. - Highlights need for independence in oversight for access decisions
WT	- Researchers should seek to maximise public benefit from their research, and so data should be made available in a timely and responsible manner. - Grant applications that are likely to lead to the development of a shareable data resource must include a data management and sharing plan.

[25] http://www.oecd.org/science/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm
[26] http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

Both the ESRC and MRC data sharing policies refer to an OECD report on access to research data from public funding. The report was developed in response to a meeting of OECD countries' science and technology ministers in 2004, at which it was agreed that the return on public investments in scientific research should be increased by maximising the use of digital research data from public funding.[27]

Thirteen key principles governing the use of and sharing of data in publicly funded research are advocated in the report. It does not provide any guidance on how these principles should be implemented or suggest particular models of good governance. None of the EAGDA funders' data policies use these principles specifically, but the values underlying the policies can be roughly mapped on to the OECD principles (Table 7-1). Although the MRC and ESRC policies are most closely aligned with them, the principles of openness, I.P protection and professionalism are evident as drivers for all of the policies. The table highlights the range of similarities and differences between EAGDA funders' policies.

Table 7-1: Comparison of EAGDA funder data sharing policies on OECD principles

	CRUK	ESRC	MRC	WT
Openness*	✓	✓	✓	✓
Flexibility	✓	×	×	✓
Transparency*	×	**	✓	✓
Legal Conformity*	×	✓	✓	×
I.P. Protection*	✓	✓	✓	✓
Formal Responsibility (over access mechanisms)	×	✓	✓	×
Professionalism* (standards and codes of practice)	✓	✓	✓	✓
Interoperability*	×	✓	✓	✓
Quality*	✓	✓	✓	×
Security	×	✓	✓	×
Efficiency* (inc. recognition of good data management)	+	×	+	+
Accountability (inc. evaluation of data access arrangements)	×	×	×	×
Sustainability*	✓	✓	×	✓

* These principles are reflected in the RCUK Common Principles on Data Policy,[28] although framed slightly differently.

** The ESRC guidance is very clear that plans must be in place for sharing data, but given that ESRC datasets will be submitted to the UK Data Service, it does not outline separate requirements for researchers to make information about the data collected and the data access process available.

[27] http://www.oecd.org/science/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm

[28] http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

Table 7-2: Abridged OECD Principles for access to research data from public funding

Principle	Description
Openness	Access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based.
Flexibility	Flexibility requires taking into account the rapid and often unpredictable changes in information technologies, the characteristics of each research field and the diversity of research systems, legal systems and cultures of each member country. Specific national, social, economic and regulatory implications should be considered when organisations develop research data access arrangements
Transparency	Information on research data and data-producing organisations, documentation on the data and specifications of conditions attached to the use of these data should be internationally available in a transparent way, ideally through the Internet. Lack of visibility of existing research data resources and future data collection poses serious obstacles to access.
Legal Conformity	Data access arrangements should respect the legal rights and legitimate interests of all stakeholders in the public research enterprise.
I.P. Protection	Data access arrangements should consider the applicability of copyright or of other intellectual property laws that may be relevant to publicly funded research databases.
Formal Responsibility	Access arrangements should promote explicit, formal institutional practices, such as the development of rules and regulations, regarding the responsibilities of the various parties involved in data-related activities. These practices should pertain to authorship, producer credits, ownership, dissemination, usage restrictions, financial arrangements, ethical rules, licensing terms, liability, and sustainable archiving.
Professionalism	Institutional arrangements for the management of research data should be based on the relevant professional standards and values embodied in the codes of conduct of the scientific communities involved.
Interoperability	Technological and semantic interoperability is a key consideration in enabling and promoting international and interdisciplinary access to and use of research data. Access arrangements, should pay due attention to the relevant international data documentation standards. Member countries and research institutions should co-operate with international organisations charged with developing new standards.
Quality …	The value and utility of research data depends, to a large extent, on the quality of the data itself. Data managers, and data collection organisations, should pay particular attention to ensuring compliance with explicit quality standards

Quality (cont'd)	Data access arrangements should describe good practices for methods, techniques and instruments employed in the collection, dissemination and accessible archiving of data to enable quality control by peer review and other means of safeguarding quality and authenticity
Security	Specific attention should be devoted to supporting the use of techniques and instruments to guarantee the integrity and security of research data. With regard to guaranteeing the integrity of a data set, every effort should be made to ensure the completeness of data and absence of errors. With regard to security, the data, along with relevant meta-data and descriptions, should be protected against intentional or unintentional loss, destruction, modification and unauthorised access in conformity with explicit security protocols.
Efficiency	One of the central goals of promoting data access and sharing is to improve the overall efficiency of publicly funded scientific research to avoid the expensive and unnecessary duplication of data collection efforts. Data access arrangements should promote further cost effectiveness within the global science system by describing good practices in data management and specialised support services.
Accountability	The performance of data access arrangements should be subject to periodic evaluation by user groups, responsible institutions and research funding agencies. Although each party is likely to use somewhat different evaluation criteria, the sum total of the results should provide a comprehensive picture of the value of data and of data access regimes. Such evaluations should help to increase the support for open access among the scientific community and society at large
Sustainability	Due consideration should be given to the sustainability of access to publicly funded research data as a key element of the research infrastructure. This means taking administrative responsibility for the measures to guarantee permanent access to data that have been determined to require long-term retention.

Cancer Research UK
The Angel Building
407 St John Street
London EC1V 4AD
T +44 (0)20 3469 8360
E publicaffairs@cancer.org.uk
www.cancerresearchuk.org

Economic and Social Research Council
Polaris House
North Star Avenue
Swindon SN2 1UJ
T 01793 413000
E comms@esrc.ac.uk
www.esrc.ac.uk

Medical Research Council
Polaris House
North Star Avenue
Swindon SN2 1FL
T 01793 416200
E corporate@headoffice.mrc.ac.uk
www.mrc.ac.uk

Wellcome Trust
Gibbs Building
215 Euston Road
London NW1 2BE, UK
T +44 (0)20 7611 8888
F +44 (0)20 7611 8545
E contact@wellcome.ac.uk
wellcome.ac.uk