Jisc

FAIR in practice

Jisc report on the Findable Accessible Interoperable and Reuseable Data Principles

May 2018

“FAIR in practice
Jisc report on the Findable Accessible Interoperable and Reuseable Data Principles”

Authors
Robert Allen, David Hartland

DOI: 10.5281/zenodo.1245568

Contents

Executive summary

This report investigates the meaning and (potential) impact of the FAIR data principles in practice. These principles were established by a group of diverse stakeholders engaging via a working group in FORCE11[1]. They are referenced in many policy documents and in developments of open science[2], for example, the European Open Science Cloud[3]. This report explored FAIR with stakeholders in the UK academic research community. Its aims were to understand how using or being inspired by these principles improves the findability, accessibility, interoperability and reuse of research data, including consideration of disciplinary differences.

Jisc established a group of research data experts and Jisc staff who provided expertise and helped to validate findings. Interviews and focus groups were conducted, primarily with researchers, but also involving input from research support professionals, publishers and funders. An assessment of the use of FAIR principles in institutional research data management (RDM) guidance was also carried out. Data collected from these activities was synthesised and structured using a PEST framework - grouped into factors relating to Political, Economic, Social and Technical aspects.

Explicit use of FAIR was seen to be limited, in many cases, to discussion at a fairly conceptual level amongst those most heavily involved in best practice of data management. Even where FAIR was fairly well established as a term and a concept, it tended to largely reflect existing practice. Arguably this was without significantly influencing many practical changes for those at the "leading edge" of data management, although it did provide a new and effective common way of communicating best practice.

However, in exploring the underlying practices of research, demonstrating findability, accessibility, interoperability and reusability, there was considerable evidence that good practice existed in all of these elements. In many cases this was both well established (over many years) and continually improving.

Significant findings included:

- There are low levels of understanding around data ownership in the research community
- Views on what adherence to FAIR means in practice and how to evidence it are inconsistent
- Strong support for growing the body of tools and resources available that reduced the burden of data management
- Lack of awareness and practice in machine readability of data
- There is a significant volume of data that is not effectively managed, particularly raw or unprocessed data, and the supporting infrastructure is absent
- Lack of good tooling to support metadata capture at data generation

Key differences across disciplines included:

- Diversity of data types
- Variation in corresponding tools and systems to support data management
- Divergent views on the "burden" of data management
- Attitudes to sharing and perceived individual benefits
- The influence of publication pressure and career issues

Common across disciplines, the FAIR data principles were:

- Considered a helpful concept in bringing together various aspects of data management best practice
- Seen as "going beyond" open access
- Recognised as allowing for flexibility and clarity in access

» Useful for supporting usage of different types of data for example through licences

A number of recommendations were identified (listed below) with potential for further exploration, or possible intervention from Jisc, working with its partners in the sector and interested parties from national services and stakeholders. They reflect the current early state of FAIR within the community, including coordinating with key groups and stakeholders on policy harmonisation. The recommendations also reflect a need to further consult and explore on a range of more specific issues before moving to any form of development or implementation phase.

Recommendation 1: The findings of this report should be considered in the context of the institutional RDM landscape in order to identify ways to improve RDM support. Jisc should consider the findings of this report in the context of the institutional RDM landscape and how its support of RDM provision within institutions can be improved by applying the FAIR principles. In particular, there is a gap in the provision of training and skills resources, tutorials and materials which could be filled by Jisc and/or others. Jisc and others should consider applying the FAIR principles to improve support in particular for training and skills resources.

Recommendation 2: Assess existing research data resources, guidance and services through the lens of FAIR to identify gaps and opportunities to include the FAIR principles and influence practice, with the aim to ensure these relevant resources are "FAIRised". For example, the provision of FAIR focused resources that help bridge the gap in knowledge and skills relating to data ownership.

Recommendation 3: Improve the understanding of how existing tools and resources can be better aligned with FAIR. Where tools don't currently exist or there are gaps, investigate the tools that are required with a view to developing them and taking advantage of a "market" for FAIR-branded tools and systems to support RDM.

Recommendation 4: Identify and adapt existing case studies to add exemplars demonstrating FAIR in practice. Where case studies do not exist, create new ones, highlighting examples of research data management processes in different disciplines and demonstrating the benefits of FAIR.

Recommendation 5: Jisc, research funders, key publishers, data services and other stakeholders, such as the learned societies should explore commonalities in approach and policies for advocating FAIR. Coordination could take a number of forms including: events, online forums, resources and guidance materials, and Jisc is well placed to lead some of this.

Recommendation 6: Stakeholders supporting and promoting RDM should consult with appropriate FAIR metrics groups[4] and FAIRsharing[5] to ensure that they are up to date with metrics and standards to measure the fairness of digital content, and that these are incorporated into new or existing exemplars, case studies and guidance; this could be a useful additional aspect to Jisc's support for RDM.

Recommendation 7: Create a roadmap that documents the steps needed to incorporate FAIR metrics into an institution's research practice. Whilst recognising the practical concerns and potential limitations of a metrics approach the road map will provide practical implementation benefits. It will lead researchers through the key elements needed to implement FAIR metrics successfully, including identifying stakeholders and risks, the use of exemplars and pilots, and clarifying what information is needed to move forward.

[1]

1 force11.org/group/fairgroup/fairprinciples
2 Open science represents an approach to research and is relevant across disciplines including in the arts, humanities, social sciences and physical sciences
3 https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
4 http://fairmetrics.org
5 https://fairsharing.org

Introduction, background, and method

"One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows."

Findable, Accessible, Interoperable, Reusable

The above quote is part of the preamble to the FAIR Data Principles, established by a group of diverse stakeholders engaging via a working group in FORCE11[6]. The FORCE11 definition of the FAIR principles is provided in Appendix A. Following the publication of the "The FAIR Guiding Principles for scientific data management and stewardship" paper[7] in Scientific Data, introducing a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable the FAIR "brand" has become a rallying cry for many in the research data community. It has come a long way in a very short time and been very successful in terms of coordination and influencing at a policy level. The principles are now referred to in many policy documents and referenced in open science[8] developments, for example, the European Open Science Cloud.[9].

As part of Jisc's work on open scholarship and science, that supports the sector with a range of work including that with regards to research data creation, management, access, re-use, and impact. I this is especially of interest given the fact that the FAIR principles itself do not define how to implement and achieve FAIR. This allows for, and presumably will lead to various approaches, and these need to be better understood. To respond to this Jisc[10] commissioned Hapsis Innovation[11] to undertake a consultation and analysis to better understand FAIR in Practice.

The aim of the project was to:

- Advise the research community on the current state of development, uptake and perceived value proposition of compliance with FAIR principles, identifying commonalities in research disciplines, with regard to technical as well as non-technical aspects
- Provide a general picture of the current practice with regards to FAIR principles
- Provide Jisc, and stakeholders in the national research community, with suggestions and recommendations on the basis of the evidence to inform further activity and decision-making considerations on the FAIR principles
- Develop knowledge and understanding of existing views and current practice, as well as perceptions, of future usage and development of FAIR principles across disciplines, including:
  - What the FAIR principles actually mean and accomplish in the context of current research practice of finding, accessing, exchanging and (re)processing research data in all phases of the research lifecycle
  - Opportunities and impact of adoption of and adherence to FAIR principles
  - Challenges to establishing common adoption of and adherence to FAIR principles
  - Discipline specific differences

- Stakeholders and their roles and responsibilities, for example, research institutions, researchers, funders and infrastructure and service providers

The report investigates the meaning and (potential) impact of the FAIR principles in practice, to understand how using or being inspired by these principles improves, or can improve, the findability, accessibility, interoperability and reuse of research data, and what these principles mean in disciplinary groups.

The methodology employed in gathering the information, evidence and stakeholder input needed to create this report on FAIR in Practice is described in detail in the appendix. This methodology included expert and practitioner interviews, focus groups and analysis of institutional websites and RDM guidance.

The range of disciplines to investigate were agreed as:

- Biological sciences
- Digital humanities - including history and archaeology
- Chemistry - including computational chemistry and crystallography
- Social sciences - including sociology and longitudinal studies

Methods used

Jisc established a group of research data experts and Jisc staff to provide expertise – at the start of the project they were interviewed and during the project they were consulted to validate findings. Key publications and sources related to the FAIR principles in the UK research context were reviewed (including those listed in bibliography at the end of this report)

The methodology employed in gathering the information, evidence and stakeholder input to inform this report on FAIR in Practice, consisted of a number of phases:

- Background research
- RDM and university website analysis
- Jisc Research Data Network (RDN) event discussion
- Experts group interviews and meetings
- Participation in EU online consultation
- Researcher and support staff focus groups
- Interviews with practitioners from a range of disciplines
- Interviews with research funders and publishers
- Synthesis and analysis to create the final report

An extensive account of the methodology and project activities, as well as the names of the research data experts, can be found in Appendix B. Outcomes of the RDM guidance and university website analysis are presented in the next paragraph.

In summary the methodology consists of iterations of information gathering and analysis, discussing the analysis with experts – including those working in research data practice - and formulating conclusions and recommendations. Types and settings of the discussions as well as the composition of participants varied. The

[1]

6 force11.org/group/fairgroup/fairprinciples
7 nature.com/articles/sdata201618
8 Open science represents an approach to research and is relevant across disciplines including in the arts, humanities, social sciences and physical sciences
9 https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
10 jisc.ac.uk
11 hapsis.co.uk

disciplines that were selected to be investigated to provide an informative overview of differences were:

- Biological sciences
- Digital humanities - including history and archaeology
- Chemistry - including computational chemistry and crystallography
- Social sciences - including sociology and longitudinal studies

At the final stage of the project a SWOT analysis was conducted using findings from the interviews. Three focus groups with over 50 research practitioners and research support staff discussed and refined outcomes and recommendations around the following themes:

- Awareness and understanding of FAIR
- Current use of FAIR in policies and guidance
- Motivations and barriers for implementation of FAIR
- Examples of implementation across the FAIR elements
- Disciplinary differences and non-data artefacts

Challenges and limitations of this study

Three main limitations and challenges were identified in conducting this study:

1. **Scope:** The study is a relatively small sample of the overall research community and inevitably contains major gaps in our understanding of the sector.
2. **Range of interviewees:** Involved were the prominent experts in their field that are more informed in terms of data management and the FAIR principles; the report can be seen to represent more of the "leading edge" rather than typical practice.
3. **Engaging researchers:** This was a particular challenge, evident in difficulty attracting researchers to focus groups, in contrast to research support professionals who were enthusiastic and highly engaged.

FAIR and RDM institutional guidance

An initial step in the work to inform background to the further investigation was an analysis of institutional websites (detailed in Appendix B), this showed that very few (17% of Russell Group universities and 6% of the remaining universities) made any explicit reference to FAIR, although focus groups identified recent activities such as awareness presentations and draft policies as beginning to include FAIR.

Using an adaptation of the Horizon2020 DMP guidelines[12] as a benchmark, a number of institutions were seen to provide guidance at some level against all of the questions relating to Accessible and Reusable. In contrast, few institutions provided details that were clearly identified as answering the questions on Interoperable, which was by far the weakest element.

Perhaps surprisingly Findable showed less coverage than Accessible and Reusable. This may be due to the very specific questions used in the Horizon2020 DMP guidelines eg listing naming conventions, versioning, registration and indexing. In contrast the questions in other elements may be thought to be more general.

[1]

12 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Synthesis of interviews and focus groups

The following section consists of a distillation of the 20 interviews conducted with research practitioners representing the disciplines identified for this report plus the contributions from the three focus groups and input from the experts group. More detailed findings structured by discipline group are presented in Appendix G.

As a useful framework for structuring the findings, we adopted the commonly used PEST[13] model, incorporating the political, economic, social and technological/technical aspects of a phenomenon. Whilst other, more sophisticated, models[14, 15] have been developed in the context of data and open science, the PEST model is widely known, matches well with the findings and includes the important "economic" perspective which is relevant in some of the issues raised, albeit implicitly in some cases. The PEST model has been use with excellent results for the KE report 'Sowing the seed: Incentives and motivations for sharing research data, a researcher's perspective'.[16]

From the topics covered in the interviews and focus groups, we present them under PEST in broadly the following structure, with some interspersing of the findings as appropriate:

- Political - guidance on data management, awareness and understanding of FAIR
- Economic - whilst this is largely a gap explicitly, a number of motivations and challenges have economic implications and underlying drivers
- Social - ownership of data, motivations and challenges of FAIR
- Technical - types of data, data storage, formats and standards, implementation of F, A, I, R

In this chapter we will address the findings on FAIR principles in the research community using the PEST model. In the following chapter 'Funders and Publishers' complementary findings can be found that came out of interviews with two funding organisations and two publishers. In the chapter 'Summary of disciplinary differences' we will once more use the PEST model to identify how disciplines differ in their approach to, and use of the FAIR principles.

Political

Political factors include institutional or national policies, and mandates or guidelines from funders and publishers. These were generally mentioned in responses to questions about motivations and challenges for FAIR and were fairly common across disciplines. In the context of this study, political is also interpreted to be issues that can be considered at a policy level and in terms of the wider agenda of FAIR in promoting sector wide good data management and open science - including the level of awareness and understanding of FAIR, and the sources of guidance. The latter covers aspects of sharing practice at the "coalface" of research, but also significantly the influence of national services, the subject of national level funding and infrastructure investment.

[1]

13 https://en.wikipedia.org/wiki/PEST_analysis
14 scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf
15 https://zenodo.org/record/20637/files/Deliverable_2._4.pdf
Open science represents an approach to research and is relevant across disciplines including in the arts, humanities, social sciences and physical sciences
16 http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf

Policy and funding

Requirements from funders and publishers were seen as a key driver across disciplines but different strategies and funding mechanisms across the funders led to challenges. Mandates alone were seen as insufficient, with the need to be supported by systems or expert practical guidance. In addition, whilst some funders ask for adherence to FAIR-like principles, typically they don't follow-up or police these principles. Submission requirements from national repositories were seen as a more concrete and enforceable set of conditions. However, the lack of a consistent set of FAIR metrics was seen as a barrier, and there needs to be a process to define the metrics and indicators that determine what is FAIR.

A number of practical issues, such as the level of heterogeneity in biological data, were identified, and communities were thought to avoid difficult issues unless it was clear there is a consistent long term funded direction of travel, with a link to policy.

A key challenge identified for major longitudinal studies in sociology was sharing practices across countries. Other agencies such as national statistical institutes reportedly don't share the same principles, with some not making data available. Some of this is attributed to national policy-level attitudes and regulations.

Awareness and understanding of FAIR

As many of those selected to contribute to the study were known to be working at the leading edge of data management, it is not surprising that they had all been aware of FAIR for some time, some since its inception in 2014. However, many also reflected on a lack of awareness more widely within their institutions and research communities, also represented by a few interviewees whose first contact with the concept was our interview. The source of awareness was often workshops, conferences or meetings, either focused around open science, data and archiving themes, or finding its way into national and international groups and communities that were more discipline focused or relating to research practice.

All contributors to the study agreed that FAIR was useful, as a set of principles to guide a general direction towards better data management, and the first whole and coherent attempt to encapsulate multiple aspects. In particular, the value of addressing data access and reuse where more complex ethical or legal considerations were in play was recognised.

FAIR was also noted as a useful "banner" or "rallying call", as a "brand" to follow, a vehicle to focus energy and commitment, and a tool for promoting good data management principles and practice. Is was thought to be particularly useful at a political level for arguing the case for clearer or more consistent policies and standards to drive improved tools and services.

Guidance on data management

Researchers across the subject groups generally sought guidance in similar places. Differences were seen to be more likely to relate to the data expertise of their research group, the level of institutional support available, the maturity of data management in their field, and the perceived relevance of national resources and services to their discipline.

Large research groups in some cases were supported by "facilities" managers or even a data team, providing highly specific subject relevant guidance on collecting, managing and sharing data. The expertise in such groups often has been developed over many years and is shared with new researchers. Where this was not available, the institutional support through research data services was generally thought to be good, often providing sufficiently subject-specific advice. Where commercial interests were involved, institutional services such as the knowledge transfer office were a key source of support. In some cases, notably social sciences, researchers felt they were more likely to seek support from researchers conducting similar studies, and there was a view that institutional services were more of an "overhead".

National data services typically offer help desks, documentation, guides and training courses, and these were seen as valuable. Examples include the UK Data Service[17], Archaeology Data Service (ADS)[18] and the Digital Preservation Coalition[19] - seen as having useful resources such as a handbook[20]. Other national and international initiatives, groups and annual meetings were an important source of standards and good practice within the discipline such as Bioschemas[21].

National tools and services provide some "implicit" guidance within the deposition process, including syntax checking, checking for process data and to seek justification where it is absent. One example is the Cambridge Structural Database[22] from the Cambridge Crystallographic Data Centre (CCDC).

Economic

A number of issues, largely identified in responses to questions on motivations and challenges, can be aligned with economic factors, either explicitly or implicitly. This includes the time/effort/cost of preparing and managing data and the associated productivity or efficiency of the research process; the investment in tools and infrastructure; and incentives and rewards for researchers.

Costs of data management

The cost of preparing and managing data suitable for sharing and reuse was repeatedly mentioned by contributors across all the disciplines, sometimes expressed as an overhead of time. All disciplines commented on the increasingly large size of datasets and the corresponding growing overhead for managing data, particularly manually finding and adding metadata.

In biological sciences the overhead was acknowledged as necessary, and helped avoid later problems, but still led to the perception of "wasting their time on data management". Groups such as experimental researchers were reported to be struggling with these issues, lacking resources in their grants and personnel with the right skills and training.

In digital humanities, it was highlighted that other researchers were not always asking the same questions and data may not have been collected with that research question in mind, leading to difficulties determining a suitably appropriate set of metadata.

In social sciences, the time to properly curate the data was also thought to be a potential barrier, but for well-established longitudinal studies and annual surveys, repeated methodology had led to a streamlined process that incorporated good data management with minimal additional effort.

Infrastructure

Other costs implications overlap with political considerations such as the amount of funding available for infrastructure to support increasing amounts of data, including the focus on "new features" rather than maintenance reported in biological sciences. This included the institutional costs of storing raw data, and the difficulties in making it accessible, leading to small proportions (if any) of raw data being made available. In different disciplines, ad hoc methods were sometimes used to store raw data, and it is unclear if this is strictly an infrastructure or cost issue, or more of a culture and skills matter. There was also recognition that appropriately designed systems can provide data management as a side effect of day to day research practice, but it was suggested that such RDM systems must give immediate benefit to that researcher. Technical aspects of infrastructure, tools and services are covered in a subsequent section.

[1]

17 ukdataservice.ac.uk
18 http://archaeologydataservice.ac.uk
19 dpconline.org
20 http://dpconline.org/handbook
21 http://bioschemas.org
22 ccdc.cam.ac.uk/solutions/csd-system/components/csd

Incentives

Hand in hand with the time overhead commonly perceived as associated with managing data, was the lack of alignment with incentives for individual researchers. Partly a cultural or organisational issue, it was nevertheless raised in the context of rewards, recognition and career advancement, which can be thought to be an economic factor at the individual level. Individual self-interest was unsurprisingly evident across disciplines. This included the perceived pressure to publish for career progression outweighing the demands to curate data in biological sciences. More directly, it was suggested that there was "almost no immediate return to any researcher for following FAIR principles". This was echoed in digital humanities where there was some reluctance to spend time on making data FAIR as it was not easy and doesn't bring any apparent benefit to their individual career. However, it was noted that some institutions were working to provide institutional mechanisms for recognition. The concept of 'producer' and 'consumer' of data was introduced, with benefits being more obvious by those who identified as both. For example, in digital humanities, researchers wanted other people's data accessible "in the best way that suits me" and it was felt that it would "make my life a lot easier if it were FAIR". Some benefits were seen to accrue over a long period of time, but sometimes to the principal investigator (PI) and the benefits needed to be more immediate to individual researchers.

Social

Motivations and challenges for FAIR generally were common across the disciplines studied in this report, with some differences in emphasis described below. Particular differences stood out in research culture, primarily attitudes towards sharing and openness. Whilst many of the themes are not specific to FAIR, the underlying attitudes and behaviour will impact on the future adoption of FAIR so are noteworthy here.

Within the social aspect of the PEST framework, the two key motivations/challenges were centred around principles of impact and altruism on one hand and skills and culture on the other. Ownership of data is considered as a dedicated section within the social aspects due to its significance. Other factors from motivations and challenges are covered elsewhere in economic and technical aspects.

Impact and altruism

Publishing data that led to reuse was acknowledged in many cases to be important for research impact, resulting in increased citations and dissemination of research. However, in some disciplines, such as history, more "traditional" forms of output were said to be more obvious forms of impact, such as publishing a book or running an exhibition.

More broadly, altruistic motivations were cited, that sharing (good quality) data was for the advancement of research, for the public good, or simply the "right thing to do". Some, particularly from digital humanities and social sciences took this further to argue it was the main purpose of research. However, as discussed in economic issues, a counter argument, particularly prevalent in biological sciences was that any altruistic principles were outweighed by the imperative to publish linked to the pressures of day to day research, limited resources and the time overhead to prepare data.

Skills and culture

There were references to data management not being part of the culture in some areas (eg experimental biology), researchers needing to be trained, and universities needing "to get better at this". Various reasons were given for the cultural attitudes. In many cases, openness and sharing were approved of in principle, but other issues reduced the levels of implementation.

In chemistry this was linked to the competitive nature of research, in other areas these related to the time/effort required, lack of appropriate tools (eg metadata capture at the point of data collection for biosciences), a sense of not being "finished" with the data (eg historians), or not wanting others to benefit from it at the original

researcher's expense. Some of these can be seen as issues of personal attachment to the research and data, and not wanting to lose control. In some cases this was suggested to be compounded by a risk-minimisation approach at institutional level.

Skills in managing data were noted as being important across disciplines, but with varying degrees of understanding about exactly what skills were needed and assessing the levels of competence. Managing good data practice was reportedly difficult in large academic schools where the scale/granularity of data and research varied considerably. For example, differences in costs and appropriateness of reproducibility were cited across different areas of biological sciences.

Some arguments were put forward across disciplines that detailed technical skills shouldn't be necessary and that experts (in the case of humanities), and/or services or tools (particularly in biological sciences and chemistry) should take away the need for individual researchers to get involved at that level. However, examples were cited of researchers in biological sciences not complying with, or subverting, the mechanisms of tools designed to guide good practice, for expediency.

Contributors from within the biological sciences acknowledged the prevalence of "sloppy science" in regards to managing data without sufficient attention on metadata and other practices. In contrast, interviewees from sociology felt that good practice in data was an essential skill and was ingrained in practice from an early stage as "good research". In the digital humanities it was felt that students, as the future of research, were not sufficiently equipped in data skills, and researchers more broadly were not quite ready to fully embrace digital culture, although things were improving particularly in findability and accessibility.

Where research culture was less open to sharing, differences could also be identified across the careers of researchers. It was suggested in multiple areas that early career researchers identified more closely with the short term pressures of publication, whereas those more "secure" in their career often were seen to have a broader, longer term perspective, and possibly a greater awareness of impact issues. However, some views in focus groups identified early career researchers as having better digital skills and more positive attitudes towards sharing in principle, so perhaps are driven by conflicting loyalties.

Ownership of data

It could be argued that ownership of data is a legal issue, and as such should be part of the PESTLE[23] extension of the PEST model, it features here as a social/cultural issue due to the substantial influence of the perception of ownership and influence of research culture, rather than ownership of data in a 'technical' sense.

Generally findings on data ownership were significant and revealing at a disciplinary level and the full, discipline specific, results are available in appendix G. However, a number of data ownership issues raised within each subject area are not specific to that field, for example different contractual arrangements for staff and research students affecting intellectual property, the interests of commercial stakeholders, funders or international partners, or the large proportion of data currently not released. However, the emphasis of the issues does vary, not only by the subject groupings we explored, but between sub-disciplines and research groups depending on the context of their research.

What is common is the wide lack of understanding of data ownership, in all except those closely working in areas with commercial, legal, ethical or contractual concerns, or with responsibility for groups with these concerns. In many cases there are assumptions that

[1] 23 https://en.wikipedia.org/wiki/PEST_analysis

"someone else" is dealing with the matter of ownership, that it "doesn't matter" because it's openly accessible anyway, or should be "given" to the public as it is funded primarily through public funding. Some of these attitudes demonstrate general confusion between controlling the access of data and the ownership of it.

In research groups where there are significant external stakeholder interests in the data, and where the consequences of breaching legal contracts or ethical codes are serious, ownership and licensing are taken very seriously. Elsewhere practices appear to support a more community-spirited collaborative approach that largely ignores ownership issues in deference to what is effective or productive in the pursuit of research.

There is a key opportunity in the work of FAIR to separate the issues of control and ownership more clearly with more effective use of, for example, licences and data access statements.

Technical

Although the FAIR Principles have consciously avoided technical implementation guidance it is clear that there are many technical issues associated with data management so these were discussed as part of this project.

The broad areas discussed across the disciplines related to the types of data, the storage of data, and the formats and standards that are used.

Many discipline specific examples, illustrating these broad areas, were given by interviewees and these have been reproduced in appendices G and H but have largely been omitted for this summary for the sake of brevity.

Types of data

Some of the disciplines investigated reported that there were a wide range of data types used within just their field. There were, however, a surprising number of commonalities across the disciplines in terms of the data types used. All disciplines reported that they worked with primary or raw data as well as secondary or processed data. Most researchers were creating their own primary data and then processing it in some way.

Many practitioners reported that they worked at times with very large types. Generally this referred to data quantities of many terabytes or petabytes. The size of imaging datasets was specifically mentioned in the biological sciences, the digital humanities and crystallography.

The types of data collected, even within a discipline, can be a complex picture. It may be that a vast majority of data is produced by a few large research facilities, with a small number of well recognised data types. However, there is often a very long tail of small, but very useful data sets created and processed by small groups, or individual researchers, leading to a wide range of data types.

Datasets created in disciplines such as the biological sciences, chemistry and crystallography are often derived from instruments, and are highly structured, while the digital humanities and social sciences use large quantities of textual data. The text based data often needs to be transcribed and encoded in order to be further processed. Indeed a major part of the work for one research project involved the encoding and harmonisation of data from a large number of countries.

All discipline areas were to some extent using historical or longitudinal data varying from annual survey data collection, to 25 year instrument results, to text from several hundred years ago.

Another common feature was the use of data from others' research. Many interviewees reported that they combined data from a number of resources in the research. For example, an archaeological researcher reported that their own primary data is rarely used on its own, it would be compared to other's data, or cross-referenced to primary historical sources.

It was notable that only in the chemistry and crystallography field was the issue of complex workflow information needing to be considered part of the data set in order to provide context. This group also mentioned that data should be machine readable, which is at the core of the FAIR Principles.

Data Storage

The storage of research data, across the disciplines cover in this report, tends to fall into three categories:

1. **Local storage** held on an individual researcher's, or research group's hard drive, on a personal network drive. Data held locally is often considered "at risk" – it may not be backed up adequately, could be insecure or on a vulnerable part of a network. This data may not findable or accessible to external researchers since such arrangements do not support the access function of a repository. It could be restricted because of bandwidth limitations, particularly for very large data sets. However, it allows researchers to have complete control over data, maintain confidentiality and keep costs very low. Much of the "long tail" research data is held in these local storage facilities as is initial instrument results. It should be noted that there is also local storage that is managed by the institution (albeit arranged via different providers), this does not have the same vulnerabilities that individual researcher or researcher group arrangements might have, for example, there will be systematic back-up arrangements, however, these arrangements do not support access and findability, the storage needs to link to repositories to address these aspects.

2. **Institutional repositories** – most research-oriented universities provide their own research data repository, often managed by the library and provided either as a local system or via some shared or cloud based model. Some researchers felt institutional repositories were limited in terms of the data types available with only the most generic formats available.

   They also had issues with:

   - The admin overhead in some cases
   - A lack of control for the researcher
   - Poor specialist discipline knowledge from the repository manager
   - Findability, with no generic search system across all institutional repositories. Note there is the development of the UK research data discovery service by Jisc and this is coordinated with the disciplinary data centres in the UK as well as with other international infrastructures and metadata approaches that support findability via search engines.[24]

   Some research funders require long term data storage (>ten years) which institutions provide through their repositories

3. **National or international specialist research databanks** – set up and managed by discipline specialist these databanks offer the gold standard in terms of data storage. They are comprehensive, well-structured and organised. Also they tend to be well known and used within the discipline with high levels of accessibility. In some cases their use is mandated, for example, the Archaeological Data Service. Journal publishers may require data to be deposited with these databanks. These databanks are also more concerned with the use of identifiers than many research practitioners and by linking the identifiers to indexes findability is improved. However, many discipline areas do not have national/international databanks, also, resourcing can be an issue, so some services require subscriptions to allow access. A number of these services are mentioned in appendix G.

[1] 24 http://researchdiscoveryservice.jisc.ac.uk/dataset

A few interviewees mentioned that they generally only made available the processed data that was directly related to published papers. The primary or raw data might only be held locally or may not be kept at all.

Standards and formats

As might be expected, contributors discussed a wide range of specialist and technical data formats and standards used across the disciplines.

Many formats were associated with specialist research instruments and software. They tend to be proprietary and often only available to those research groups. These formats are discussed and identified in more detail in appendix G. The specialist knowledge and the licences required (often at a high cost) make it difficult for this type of data to be FAIR.

At the other end of the scale, considerable quantities of data is held in very common, basic formats such as Excel spreadsheets, plain text, XML and .csv files, along with Tiff and Jpeg for images. This is particularly the case for data held in institutional repositories. While this data is superficially accessible it may suffer from lacking context, or data loss, in being formatted.

Data that is made available through national/international databanks is often both raw and processed with appropriate formats, along with useful metadata. These services, along with institutions such as the Library of Congress, also help in standardising formats across a discipline. For examples services like FAIRsharing.org[25], provide a **"curated, informative and educational resource on data and metadata standards."** across multiple disciplines, including the life, environmental and biomedical data resources. The dedicated curation teams associated with these services assists researchers in their use of standards and in providing the necessary metadata elements during submission.

In some disciplines the validity of research may depend upon agreed de facto standards, for example, mass spectrometer results need to adhere to very strict international standards. Without standards the research will be considered of little value. In turn, some journal publishers insist upon standard data formats from researchers submitting papers, with systems in place automatically check this.

When discussing the machine readable aspect of the FAIR principles it was clear that there is a desire from increased automation in managing data which is enhanced by wide use of recognised standards. This automation enables submission tools to pipeline richly described data into data repositories. It is particularly the case with large quantities of data, where manual checking for compliance is impractical.

In some fields the uptake of standards has been problematic, for example where researchers do not recognise that they are even using data, or they feel that standards may constrain their working methods. Similarly, some formats are historical or legacy making current usability difficult. Finally, the "long tail" of research under taken by small groups and individual researchers often produces data that does not adhere to widely used standards.

[1] 25 fairsharing.org

Implementation of FAIR

This section summarises the main elements of the current implementation of the four principles of FAIR. A full review of the contributions and the discipline based results from the interviews and focus groups is available in appendix H.

Findability

Most of the practitioners taking part in this project reported that they used one or more of the following methods to find data:

- Well known, discipline based repositories and databanks
- Search engines such as Google
- Research papers published in journals containing data or its location
- Word-of-mouth with colleagues at conferences etc, or through social media

The need for rich metadata was discussed in terms of enhancing findability but the overhead entailed in its creation means that metadata is often patchy at best. There are also issues relating to uncertainty regarding which metadata standards to use in some fields along with poor documentation and the need for goods tools to increase automation.

However, metadata schemes supported by repositories and databanks are often more developed meaning the data they hold tends to be more findable.

In some fields the concept of findability was not a high priority. One researcher commented, when asked how can researchers can find their data, that "mostly they don't". The issue of confidentiality was also raised meaning that some data is intentionally not findable.

The findability of raw data, and whether this is desirable, was discussed at length with many preferring only processed data to be made findable.

Where findability is desirable a number of possible improvements were suggested including the need to build catalogues, the requirement for closer links between researchers and those building tools, and increasing the findability of the "long tail" of research data.

Accessibility

The use of well-run repositories and databanks, for example in the biosciences, using web based interfaces, APIs and, sometimes, simple spreadsheets means that data can be very accessible. However, there are many challenges, for example "long tail" data, poor quality metadata, along with the need for additional tools, expensive, specialist software, standard formats and high quality documentation.

In some disciplines data produced may be instrument specific reducing its accessibility to those not using vendor licensed software.

Embargos are seen as an issue in other fields, caused by IP, ethical, commercial and legal restrictions. Data may also be held back until further clearance from the community being studied is obtained or where it might be misused, for example, ivory trade data.

The size of data was also noted by some researchers, it could be too big to download and work on with major implications for university infrastructure.

Interoperability

A majority of contributors found interoperability to be the most challenging of the four FAIR principles. This, in part, is due to interoperability not being well understood, for example, the use of XML was considered by some to mean data was fully interoperable.

Some repositories and data services aim to improve interoperability by providing tools to map between datasets. They endeavour to address issues including:

- **Content** - is it there, is there a standard?
- **Syntax** - what format, can it be read, is it valid, can it be transformed?
- **Semantic** - is the same language used to describe the datasets?

The issue of differing terminology and the need for common vocabularies was noted across disciplines with some services attempting to match terminology and map standards landscapes. FAIR is seen by some as helping to improve standard vocabularies and mapping tools.

The digital humanities field reported some good levels of interoperability with the researchers mainly using processed rather than raw data, plus good documentation and standards available. The principle of interoperability is also considered very important to chemistry researchers. They are keen to access other researchers' data, however, much of the data produced from instruments is in a proprietary format or requires specific software.

Reusability

Reuse of data was acknowledged to be beneficial to research across all disciplines and is often considered a result of the other three principles - findability - accessibility - interoperability.

However, levels of reusability were very variable, depending upon the specific research field and data type. Data is most reusable where data types are simple and easy to describe, and when the community is organized and collaborative.

Other important factors include:

- Ease of data creation
- Quality and availability of tools including software with versioning information
- Need for rich metadata that describes the whole research process
- Use of standard, up-to-date, formats

Many of the repositories and data services discussed consider reuse to be an important function so provide high quality metrics, check for errors and clean up semantics.

The trend towards standard licences, the use of creative commons and open source software were discussed by some researchers but in many cases licensing was not explicit, well understood or highly visible. When using creative commons the most restrictive version is often chosen by researchers, out of fear that "someone might use it for something I don't want them to" for example, for commercial gain.

Data reusability was seen by some as "an extra thing they have to do" and was therefore a low priority. However, some researchers are producing "data papers" - short articles that describe the data, how it was collected, what it is for, the methodology, the context and a DOI to the raw data. One practitioner mentioned that to make data more reusable peer reviewers should be required to include a view on the level of data reusability in the paper.

Funders and Publishers

The main focus of this report has been practice within the research community, however, representatives from Economic and Social Research Council (ESRC)[26] and the Wellcome Trust[27] were interviewed to explore the funder role, and Elsevier[28] and Springer Nature[29] on behalf of publishers. These are detailed in Appendix I.

ESRC have a data infrastructures team and fund major longitudinal studies in addition to an administrative research network, and the UK Data Service[30]. More recently they have added flexibility to allow repositories other than the UK Data Service, as long as the principles are met. They publish a research data policy that has been recently updated and includes explicit reference to the FAIR data principles as does their research guide. The RCUK[31] position statement is a key influence and the council benchmark against that. The FAIR principles have been on Wellcome's radar for over a year and consider the FAIR "brand" to be very useful. While it is not yet specified in policies, Wellcome aims to follow the principles and is particularly interested in the cultural barriers. Funding longitudinal studies and data infrastructures was seen as key for ESRC, and a lot of money is about preparing data for reuse, providing staff and technical infrastructures.

In terms of challenges for FAIR, implementation of the principles was not straightforward, in particular with interoperability. It was recognised that the more you try to bring together different data infrastructure and different data, the more difficult it became to align standards, access, metadata and different cultures and infrastructure.

Other challenges noted were:

- Public perceptions and attitudes on how they want their data to be used
- Skills for researchers in knowing how to work with data
- The need for incentives for researchers
- The lack of available repositories making data harder to find
- Sharing personal and sensitive data and balancing openness with confidentiality

Publishers Elsevier and Springer Nature work with a range of international data management groups, including Research Data Alliance[32] CODATA[33], FORCE11[34] and ELIXIR[35]. The types of data are highly variable, reflecting the variety of disciplines covered by their journals, from humanities to high energy physics. Some broad discipline-specific differences are recognised, such as life sciences data sharing being generally more established than social sciences. However, this is not necessary true in sub disciplines.

[1]

26 esrc.ac.uk
27 https://wellcome.ac.uk
28 elsevier.com
29 springernature.com/gb
30 ukdataservice.ac.uk
31 rcuk.ac.uk
32 rd-alliance.org
33 codata.org
34 force11.org
35 elixir-europe.org

A pragmatic approach is taken by providing a range of data management policies, recognising that some journals and editors are only just starting to think about policies. The policies do not currently refer explicitly to FAIR, and this is seen to reflect current practice of researchers, who are generally not asking about it.

The main challenges to the implementation of FAIR that have been identified include:

- The lack of incentives or credit available to researchers to adhere to FAIR
- Variable interpretations of FAIR eg specifying what machine readable means
- The need to improve use of metadata and DOIs and standard file formats
- Lack of clarity and low awareness of embargos, restrictions and licensing arrangements
- Poor levels of interoperability
- Low awareness and usage of appropriate infrastructure
- Low awareness of, andinvestment, in data curation

It was noted by other contributors to this study that small publishers are involved in innovative practice and that some of the bigger publishers are supporting F and A from FAIR but not I and R. Details and comparisons of publishers' data policies are covered elsewhere[36].

[1]

36 dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

Summary of disciplinary differences

The current uptake of the FAIR principles varies widely between disciplines. Some of the fields studied in this report are leading the way, and some have established good practice in data management that has evolved over decades, but the penetration in many fields is almost non-existent. It is early days, in fact, the recognition in the research community of FAIR as a concept has been very fast. The following sections summarise some of the key points identified in disciplines, using the PEST model once more.

Political

In biological sciences numerous different repositories were noted, sometimes specialising in a particular data type, but sometimes competing for the same data. In contrast, The UK Data Service[37] is the primary deposition place for social sciences, funded and supported by the ESRC[38]. Its position allows it to develop tools and provide guidance to support standards and policy. Elsewhere, small repositories lack the guiding tools that help make data deposition supportive of good data management and sometimes limit the size of datasets. In some cases, researchers in other disciplines rely on their institutional repository, some of which are "data neutral", but may lack the sophistication of mechanisms for good metadata specific to the data type.

Some internet journals and data repositories are beginning to provide live links between data and articles, for example, the Internet Archaeology Journal[39] and the Archaeology Data Service (ADS)[40]. Similar developments can also be observed in chemistry where researchers produce live linked FAIR data tables able to present data in different formats, and accessed via DOI links. These interesting examples provide a glimpse into possible future integration of publications and data that may also help with the issue of giving credit to researchers. Another example is the production of "data papers" in history, with academic credits assigned.

Guidance from national tools and services helps researchers to identify how to document non-data aspects of the research process to aid reusability and interoperability, for example, the Digital Curation Centre[41] provides guidance for data management planning that includes suggestions on what to consider:

Well established disciplinary networks and good practice initiatives such as Bioschemas[42] were influential in developing standards in biological sciences. In archaeology examples were given of de facto standards developed through publication of a good practice article detailing data handling techniques. Through recognition as a published article and from wide citation, the good practice became shared through the community. This is in contrast to the more formal bodies and working groups in other areas that develop and publish standards.

Economic

In every case senior researchers and research managers across disciplines supported the FAIR principles, however, some reports in biological sciences stressed the overriding driver of getting to publication quickly at the expense of good practice in managing data. This was usually attributed to the incentive and career progression mechanisms in place that emphasised publication.

Researchers who use very expensive equipment, such as the hadron collider, of which there is only one, the researchers have to collaborate and share by necessity. Where costs are low and researchers can work in small independent groups the need to share is not as pressing, but also where the perceived benefits were to the individual rather than the community, "protectionism" was more prominent.

[1]

37 ukdataservice.ac.uk
38 esrc.ac.uk
39 http://intarch.ac.uk
40 http://archaeologydataservice.ac.uk
41 dcc.ac.uk
42 http://bioschemas.org

Social

Almost all of the experts we talked to felt that FAIR would be appropriate in all research disciplines. However, many felt the differing cultures across disciplines may provide barriers. There was discussion regarding the sharing culture within disciplines, some disciplines have a deep and long standing sharing culture, but others do not.

Some experts thought that the differences were sometimes overplayed, every discipline feels it is "special" and "unique" but generally the results of their research could be handled using generic principles. There were however, clear differences in the attitudes and norms, also reflected in training of new researchers.

The consequence highlighted in biological sciences was said to be that "fast science" became "bad science" or "sloppy science" in terms of the quality of data and metadata, especially for any meaningful reproducibility or reusability. In contrast, strong views representing major longitudinal studies in social sciences treated data as the main purpose of the research, and reusability as a primary aim, viewing other practices as "shameful".

Technical

Other differences were more technical. Experts suggested a bigger barrier was likely to be the type of data rather than its discipline focus. With the huge variety of data types in existence, even within a field such as biological sciences, there is wide variation in the level of standardisation of formats and processes. Differences were observed to follow some generic principles. Commonly used data types, and ones that were less complex, attracted widespread support and momentum for developing standards, tools and infrastructure. Such support allowed for easier adherence to FAIR principles.

The lack of standard vocabularies was a barrier to interoperability and occurred in different disciplines. In some cases this was due to national or institutional differences. However, there were some key disciplinary features where data for research crossed boundaries. An example of this is in biological sciences where data is originally in a clinical record, possibly from another country, and using a clinical vocabulary. When used in other research contexts, translation is needed and tools exist, but this is not a problem of lack of standards as one of research across contextual boundaries. Similar issues are likely in other cross-disciplinary research.

A wide variety of types of non-data material was described and discussed, although the relationship to FAIR was not always apparent. Materials mentioned included artefacts that support the research life cycle and can be used to help validate and reproduce research. Indeed they may be needed to allow the research to be built upon. Examples mentioned include: laboratory notebooks, software, reports, workflows, publications, documents, manuscripts, photographs, physical objects, samples, mathematical models, performance, equipment types and settings, video and images of many kinds. It was acknowledged by many interviewees that this is an area of weakness in terms of true reproducibility, due in part to the lack of complete data, insufficient process description or context, or poor metadata capture - either human or equipment. In some areas, such as sociological longitudinal studies, it was felt that complete process information was included including questionnaires and user guides, to provide a "large and comprehensive description of the survey". In biological sciences, the analysis process may be documented within the paper or provided in a supplemental file in publications, or scientific workflows deposited or code shared using eg Github.

Summary of key drivers, challenges and lessons learned

Research data management practice around the FAIR data principles was seen to be influenced by a range of different drivers and barriers, described in earlier sections and summarised below.

Political

1. Levels of awareness and understanding of FAIR were thought to be quite low amongst researchers generally

2. FAIR was very rarely mentioned in institutional policy and guidance, although there was evidence that less formal presentations, workshops and training sessions were starting to incorporate the principles

3. FAIR was only sometimes mentioned with national level tools, policies and services, although recent updates and new versions were seen to be starting to incorporate FAIR

4. Lack of awareness, and in some cases agreement, of what was considered good practice in FAIR data management, particularly across disciplines, limited the sharing of good practice

5. Current lack of metrics and measures to assess FAIRness led to inconsistent views on adherence to FAIR

Economic

6. Funder mandates were a significant driver, although there was said to be a lack of harmonisation across funders, and the lack of compliance monitoring potentially led to a disparity between intent and actual practice

7. The "burden" or overhead of data management was considered a significant issue for many

8. The lack of appropriate recognition and incentives for individual researchers for good practice in managing and releasing data (as opposed to publications) within institutions was a significant issue for some, and in some cases a reported lack of resourcing or funding

Social

9. Cultural attitudes to sharing and openness varied between disciplines but there were many examples of others who wished to protect their data in pursuit of a perceived personal advantage in doing so

10. Challenges across national boundaries, in particular working with other agencies that have different laws, attitudes or technical standards make integrating data complex for some

Technical

11. Repository deposition requirements were found to be helpful in indicating good practice for data management

12. Tools to support data management, through the entire process from collection to deposition were considered critical to the success of effective data handling

13. The large scale of data in some cases was a barrier to some data handling processes, including the omission of complete "raw data" where it was considered too large to store or share

14. Complexity/variety of data types and formats leads to inconsistent standards, support, tools and infrastructure - with the most commonly used (and simpler) data types attracting good support

15. Standard vocabularies were a particularly challenging area in several disciplines, where tools and services were beginning to be developed to aid translation, but manual interpretation still took place

16. Lack of good data collection tooling in particular to support metadata capture at the earliest point of data generation

Conclusions and recommendations

This report aimed to explore "FAIR in Practice" in UK academia, and one of the most common questions that arose concerned the meaning of the term "in practice". Did it, for example, refer to people explicitly (and knowledgeably) referencing FAIR in their work, or to their everyday research practice that reflected the individual elements of FAIR?

In response to the former question, explicit use of FAIR was seen to be limited in most cases to discussion at a fairly conceptual level amongst those most heavily involved in best practice of data management. Even where FAIR was fairly well established as a term and a concept, it tended to largely reflect existing practice, forming a new way of describing data management processes. Arguably this was without significantly influencing many practical changes for those at the "leading edge" of data management, although it did provide a new and effective common way of communicating and advocating best practice.

However, in exploring the underlying practices of research demonstrating findability, accessibility, interoperability and reusability, there was considerable evidence that good practice existed in all of these elements, and in many cases was both well established (over many years) and continually improving. The FAIR data principles were seen as a helpful concept in encapsulating, and bringing together more coherently, various aspects of data management best practice. In particular it was seen as "going beyond" open access to allow for flexibility and clarity in access and usage of different types of data for example through licences.

There was considerable variation in the levels of "adherence" to FAIR, between researchers, disciplines and universities. Some of that related to wider cultural aspects of research practice and attitudes towards openness and sharing, and the perceived benefits to individuals, organisations and society. In other areas it reflected technical issues of the research process and the tools and infrastructure currently available. However, a crucial gap was observed in the lack of a clear mechanism to determine FAIR "compliance", such that any discussion around "FAIRness" was currently without an agreed set of measures to provide a consistent answer. Instead, as with this report, the approach tended to be to seek "evidence" of where aspects of FAIR were closely followed and where they were not. Developments in policy and community standards look set to address the issue of assessing FAIRness, for example with the work on FAIR metrics[43] and the DANS badges concept, so the timing is good for FAIR to become more widely "implemented" in a concrete sense.

One key area that was not widely evidenced (or lacked awareness) was the machine readability or computability aspect of FAIR, with the exception of some tools and infrastructure. This was linked primarily to the FAIR element of interoperability, seen by a majority of interviewees as the most challenging of the FAIR elements in practice.

Some saw machine readability as core to FAIR, many others felt it was the next step along a gradual path, for which the necessary skills, tools and infrastructure was yet to be commonplace. In that respect, there is a clear opportunity for Jisc to support the growing momentum behind the FAIR concept, and enable it to move more quickly into established practice.

Challenges and limitations of this study

Three main limitations and challenges were identified in conducting this study, these are outlined in the Methods chapter, but are set out here again for convenience:

1. **Scope:** The study is a relatively small sample of the overall research community. Whilst it gives some rich insights into practice relating to FAIR in some areas, it inevitably contains major gaps in our understanding of the sector. It should be considered as a sample of material to feed further discussion and consultation, highlighting some key areas to consider for the future.

2. **Range of interviewees:** Within each of the discipline areas there was still some variance in attitudes and practice, and particularly in the biological sciences,

there was heterogeneity in terms of data and research methods. The interviewees were typically suggested by the experts group, formed of those more prominent in their field, and more informed in terms of data management and the FAIR principles. Whilst they also gave useful reflections on their field in general, the report can be seen to represent more of the "leading edge" rather than typical practice.

3. **Engaging researchers:** This was a particular challenge, evident in difficulty attracting researchers to focus group, in contrast to research support professionals who were enthusiastic and highly engaged. Alternative means of engaging researchers, possibly through established groups, learned societies, discipline networks, existing events, or using research support professionals as a conduit, should be considered for future activity.

Recommendations and areas for further study

A number of areas were identified during the study with potential for further exploration, or potential intervention from Jisc, working with stakeholders across the sector and interested parties from national services.

The recommendations reflect the early stage of FAIR within the research community, leading to a need to further consult and explore on a range of more specific issues before moving to any form of development or implementation phase.

Whilst the influence of culture and attitudes in research was highlighted as significant, this is seen to be broader than FAIR and an issue likely to be dealt with via other mechanisms. However cultural issues should be taken into account in assessing the feasibility of other potential interventions. In areas of skills and training, particularly of early career researchers or research support professionals the FAIR data principles should be part of the conversation on good practice in data management. However, with difficulties engaging researchers in the topic of FAIR, careful consideration would need to be given to the communication and message to create a compelling narrative that attracts and encourages researchers. FAIR can also be seen as creating a 'market' for tools and services. The FAIR 'brand' could be used as a marketing device for national support, for example Jisc could consider its use both in skills areas (eg digital capabilities and researcher skills) and in products and services (such as the Research Data Shared Service).

This report provides a sample of four discipline areas with supplementary evidence from other experts and stakeholders. Most of the evidence is from researchers but it is clear that research support staff play a pivotal role. Such staff have stronger engagement with FAIR, and also a strong relationship with Jisc through existing initiatives. There are opportunities in strengthening consistency in the application of good practice in RDM provision across the sector, this could be via Jisc and its partners such as ARMA[44] and possibly working with partners such as the UK Data Service or Digital Curation Centre.

Recommendation 1: The findings of this report should be considered in the context of the institutional RDM landscape in order to identify ways to improve RDM support. Jisc should consider the findings of this report in the context how its support of RDM provision within institutions can be improved by applying the FAIR principles. In particular, there is a gap in the provision of training and skills resources, tutorials and materials which could be filled by Jisc and/or others.

[1]

43 http://fairmetrics.org
44 https://arma.ac.uk

Resources, tools and infrastructure

Having appropriate tools to manage data throughout the data lifecycle from data collection to deposition is seen as critical to enable FAIRness. For some, this was the biggest factor in FAIR adoption and the only way to get researchers to improve data handling methods. More detailed discipline-specific consultation would be needed to determine priority areas, but infrastructure is seen as having some "gravitational force", as one interviewee put it "there are some really good examples where it's been possible to organise scientific communities around providing infrastructure."

Many good tools and services already exist but the following suggestions have been made by interviewees that could benefit FAIR adoption:

1. Review current tools to assess FAIRness including the Jisc Research data shared service and any related suppliers
2. Actively promote tools and services that are supporting the FAIR principles
3. Work with tool developers and data service providers to better understand how to improve FAIR, including use of metrics
4. Explore the development of further FAIR tools and services in key areas of deficiency for example metadata capture, translation tools across vocabularies

Recommendation 2: Assess existing research data resources, guidance and services through the lens of FAIR to identify gaps and opportunities to include the FAIR principles and influence practice, with the aim to ensure these relevant resources are "FAIRised". For example, the provision of FAIR focused resources that help bridge the gap in knowledge and skills relating to data ownership.

Recommendation 3: Improve the understanding of how existing tools and resources can be better aligned with FAIR. Where tools don't currently exist or there are gaps, investigate the tools that are required with a view to developing them and taking advantage of a "market" for FAIR-branded tools and systems to support RDM.

In addition, there were many excellent examples of good practice highlighted by interviewees and other stakeholders, demonstrating how FAIR can be, and is, implemented in practice. Such practical illustrations are seen as highly beneficial to those who have less understanding of the concepts and implementation detail. Providing detailed case studies and exemplars of FAIR in practice across different disciplines would be extremely useful for promoting FAIR and giving practical support. The benefits of providing case studies would be to highlight real examples of research processes in practice in different disciplines, to demonstrate to others how and why FAIR is beneficial. However, many case studies already exist, so rather than create new ones, it is proposed that following from the assessment in the previous recommendation, suitable existing case studies are identified.

Recommendation 4: Identify and adapt existing case studies to add exemplars demonstrating FAIR in practice. Where case studies do not exist, create new ones, highlighting examples of research data management processes in different disciplines and demonstrating the benefits of FAIR.

Advocacy

Jisc is seen in the sector to have influence as an advocate for good practice in research data management, and it is fitting therefore for Jisc itself to make FAIR more prominent in materials and events relating to data management. An extension to this would be promoting the inclusion of the FAIR data principles in guidance and policy in Jisc-funded tools and services and in partner organisations that provide data tools and services widely used and respected by the sector. This includes repository and data services at institutional, disciplinary and national levels. Recognising that many of these already have FAIR on the agenda, a practical step might be to hold a series of conversations to see how best this can be introduced.

A fuller picture of the status of policies and intentions towards FAIR from individual funders, RCUK/UKRI[45], publishers and other key stakeholders would also be of interest in providing knowledge of the wider landscape.

Jisc is in a good position to support the sector to effectively work towards 'FAIRification', but it is essential to work with a range of stakeholders across policy, funding and infrastructure development to see how to address this at different points within the research process.

Recommendation 5: Jisc, research funders, key publishers, data services and learned societies should engage to explore commonalities in approach and policies for advocating FAIR. Coordination could take a number of forms including: events, online forums, resources and guidance materials, and Jisc is well placed to lead some of this.

Standards and metrics

Having some way of determining FAIRness was a key question from respondents, and could supplement case studies and work with service providers to provide practical illustrations on how to achieve FAIR in practice. The work of the FAIR metrics group[46] is seen as pivotal in this and could provide timely insights as they develop the metrics. As stated on their website:

"Communities must not only understand what is meant by FAIR, but must also be able to monitor the FAIRness of their digital resources, in a realistic, but quantitative manner."

However, there are also concerns that a "compliance" approach could counteract momentum in supporting FAIR as a set of principles, so a focus on specific exemplars is recommended to assist in practical application of FAIR rather than measurement. The work of the FAIRsharing[47] resource can also help mapping the landscape of standards in different disciplines, monitoring their use and adoption by data repositories and data policies (by funders and journals). An example of working across the sector to promote RDM in FAIRsharing is the work Jisc has done with FAIRsharing by providing journals' data policy information collected as part of a Jisc pilot project[48].

Recommendation 6: Stakeholders supporting and promoting RDM should consult with appropriate FAIR metrics groups[4] and FAIRsharing[5] to ensure that they are up to date with metrics and standards to measure the fairness of digital content, and that these are incorporated into new or existing exemplars, case studies and guidance; this could be a useful additional aspect to Jisc's support for RDM. National providers such as Jisc and the UK Data Service may wish to consider how they could be incorporated into tools, services and infrastructure.

While metrics are frequently felt to have practical benefits in assisting implementation, potential drawbacks can occur if focusing on "compliance" at an early stage of engagement with the FAIR principles

Recommendation 7: Create a roadmap that documents the steps needed to incorporate FAIR metrics into an institution's research practice. Whilst recognising the practical concerns and potential limitations of a metrics approach the road map will provide practical implementation benefits. It will lead researchers through the key elements needed to implement FAIR metrics successfully, including identifying stakeholders and risks, the use of exemplars and pilots, and clarifying what information is needed to move forward.

[1]

45 rcuk.ac.uk and ukri.org
46 http://fairmetrics.org
47 https://fairsharing.org
48 jisc.ac.uk/rd/projects/journal-research-data-policy-registry-pilot
4 http://fairmetrics.org
5 https://fairsharing.org

About this report

This work on the FAIR principles was initiated as part of Jisc's work on open scholarship and science that supports the sector with regards to research data creation, management, access, re-use and impact, alongside other aspects of open scholarship and science, for example open access to publications.

The creation of this report overseen by Jisc staff Rachel Bruce - Director open science and research lifecycle, Dom Fripp - Senior curation metadata developer, and Bas Cordewener, International facilitator.

This report was commissioned by Jisc and the authors were Dr Robert Allen and David Hartland from Hapsis Innovation Ltd.

The project was advised by an experts group comprising the following who are thanked for their contributions to the project:

- Professor Simon Coles - Director of the UK National Crystallography Service, University of Southampton
- Dr Veerle Van den Eynden - Research Data Services Manager, UK Data Archive, University of Essex
- Professor Carole Goble - School of Computer Science, University of Manchester
- Professor Cameron Neylon - Centre for Culture and Technology, Curtin University
- Professor Susanna Sansone-Assunta - Associate Director of the Oxford e-Research Centre, Department of Engineering Science, University of Oxford
- Professor Melissa Terras - Director of Digital Scholarship, College of Arts, Humanities and Social Sciences, University of Edinburgh, formerly Director of the Centre for Digital Humanities, University College London at the initiation of the project

Appendix A: FAIR guiding principles as defined by FORCE11

force11.org/node/6062

1. To be **Findable** any Data Object should be uniquely and persistently identifiable
    a. The same Data Object should be re-findable at any point in time, thus Data Objects should be persistent, with emphasis on their metadata
    b. A Data Object should minimally contain basic machine actionable metadata that allows it to be distinguished from other Data Objects
    c. Identifiers for any concept used in Data Objects should therefore be Unique and Persistent
2. Data is **Accessible** in that it can be always obtained by machines and humans
    a. Upon appropriate authorisation
    b. Through a well-defined protocol
    c. Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object
3. Data Objects can be **Interoperable** only if:
    a. (Meta) data is machine-actionable
    b. (Meta) data formats utilise shared vocabularies and/or ontologies
    c. (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible
4. For Data Objects to be **Re-usable** additional criteria are:
    a. Data Objects should be compliant with principles 1-3
    b. (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources
    c. Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation

Appendix B: Methodology and project activities

In order to advise the project team, Jisc established a group of research data experts and Jisc staff who provided expertise and helped to validate findings. The group also suggested a range of disciplines to address and research practitioners to interview.

The project team reviewed the key publications and sources related to the FAIR principles in the UK research context. These include those listed in the report bibliography. Additionally discussion and interviews with the project expert group provided the team with valuable background knowledge and insight.

Data for this study was collected between June and November 2017. The methodology employed in gathering the information, evidence and stakeholder input needed to create this report on FAIR in Practice consisted of a number of phases:

- Background research
- RDM and university website analysis
- Jisc Research Data Network (RDN) event discussion
- Experts group interviews and meetings
- Participation in EU online consultation
- Researcher and support staff focus groups
- Interviews with practitioners from a range of disciplines
- Interviews with research funders and publishers
- Synthesis and analysis to create the final report

Additionally, the European Union (EU) Experts Group - Turning FAIR Data Into Reality[49] initiated an online consultation[50] during this study (3rd July – 26th July), and has received contributions[51] from some of the experts and groups included in this report. While the EU Experts Group has a different focus, and their report has yet to be published, many of the themes arising are in line with the findings of this study, and their report is expected to contain evidence that will complement this study.

Desk analysis

Explicit reference to FAIR on university websites

The project team reviewed the extent to which FAIR is explicitly mentioned on the websites of all 135 UK universities.

It seems likely that the Russell Group[52] universities, which receive the overwhelming majority of research funding within the UK, will be in the vanguard of implementing the FAIR principles. To get a snapshot, albeit superficial, of the current status of FAIR at these institutions the websites of all 24 Russell Group universities were reviewed using keyword searches and menu browsing. All provided considerable information on general research data management principles and guidance, additionally they also discussed open access. However, only four (17%) of them mention FAIR principles at all. The team then looked more widely at the other 111 UK universities. Only a further seven (6%) referenced the FAIR principles in any way.

What is revealing is the context of how and where FAIR is mentioned. This can be summarised as:

- Three are an external link to Horizon2020 data management planning (DMP) guidelines[53]
- Two are a link with some additional description but fairly brief about FAIR
- Four are blog posts
  - 1 of which is an individual's blog, the others are institutional blogs dedicated to relevant themes such as open access, open research and research data
- One is a research publication
- One is discipline specific guidance on research data management (in bioinformatics)

Research data management guidelines on websites

Many research-focused institutions make available guidance to their researchers in the form of web pages giving guidance on how to manage data. In some cases this is a discrete and self-contained toolkit, in others it is a disparate collection of policies and advice documents. Despite the findings that few universities explicitly refer to FAIR, many established practices within research can fulfil aspects of the underlying FAIR principles.

Institutional websites were explored to attempt to assess the extent to which FAIR principles were addressed in current guidance to researchers. University websites were visited, identifying the most likely central source of guidance, such as the library or research support service. Toolkits, advice pages and policy documents directly linked to these pages were accessed to attempt to identify evidence against each of the FAIR components.
A checklist of criteria/questions was adapted from the Horizon2020 DMP guidelines, to rephrase the DMP questions into a form that could answer the question:

"Does the institutional guidance recommend or encourage the following..."

Other than the phrasing, the questions were left largely intact. The checklist is provided in an appendix.

Initially it was envisaged that this could be a simple tick list to verify the existence of each criteria. However, as discussed in the findings, it was soon found that this wasn't straightforward, and instead, evidence was captured in a spreadsheet, in the form of a relevant statement or extract relating to each criteria.

Websites were assessed from a selection of universities from within the following mission groups

- Russell Group[54]
- University Alliance[55]
- Million+ Group[56]

Many of these were previously identified as mentioning FAIR somewhere on their institutional website, as reported above. This sample was chosen to arguably represent the set of universities most likely to have a fuller coverage of FAIR principles in their institutional guidance. It is unlikely that expanding this sample would return much in the way of meaningful additional data given the findings and limitations noted below.

[1]

49 http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464&NewSearch=1&NewSearch=1
50 https://github.com/FAIR-Data-EG/consultation
51 https://github.com/FAIR-Data-EG/consultation/issues
52 http://russellgroup.ac.uk
53 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
54 http://russellgroup.ac.uk
55 unialliance.ac.uk
56 millionplus.ac.uk

RDM website findings

The process of assessing RDM guidance on websites proved to be far more difficult than anticipated for several reasons. Some were practical issues regarding the access and location of the information. Typically there was no single source of information, as advice on different topics was fragmented across multiple pages and sections, often referencing separate policy documents or guidance in related areas. In particular, requirements for, and facilities to enable, open access and research data management were often documented in different places.

The guidance varied in level and detail, from high level policy documents stating general principles and responsibilities, to detailed practical guidance aimed at practitioners. In some cases there was a clear intent (eg to be as open by default), and in others it was left to the discretion of the researcher in how to interpret and implement (eg whether data should have restricted access).

Most guidance had external links to many sources including:

- The Digital Curation Centre[57] and specifically its DMPOnline tool[58]
- Funder, research council and publisher requirements
- Repository input requirements
- National services and resources such as the UK Data Service[59]
- Useful pages on other university websites
- International good practice and policy including eg the Concordat on Open Research Data[60]
- Discipline specific guidance, formats, repositories

It became clear that in most cases the guidance assessed was not intended as the single comprehensive source of information on how to manage research data. Instead it was pitched as a starting point for general themes. More detailed guidance was provided either by professional services staff in a bespoke manner, through the research discipline or department's practice, or by external sources.

Evidence was captured against each of the questions within the four elements of FAIR (findability, accessibility, interoperability, reusability). This evidence was typically a partial sentence or extract relating to the criteria. In many cases the evidence was insufficient to say the question had been fully answered (ie the guidance satisfied the criteria). However, what it did provide was a picture of the areas in which guidance was stronger and weaker, in terms of the extent to which the topic was mentioned.

Examples of weak coverage were often general statements referring to the topic without specific guidance, and giving leeway to the researcher to interpret. These include the following as illustrations:

- **Clear version numbers:** Good metadata should include descriptions of how and when the data was collected, a key to abbreviations or codes used within the dataset, and details regarding provenance and versioning
- **Identifying clearly any reasons for restrictions on use or sharing tools needed for access:** In some instances, it may be possible to restrict access to the data. In this case, a Data Access Statement may be required
- **Timescales for the data to be made available for reuse (including any embargos for publishing/patents):** You may be able to delay sharing your data for a short, defined time to allow further papers to be published or for patents to be filed

Examples of better coverage gave more specific details on how to meet the requirement, such as:

- **Persistent and unique identifiers such as Digital Object Identifiers:** Will the repository issue your data with a persistent identifier, such as a Digital Object

Identifier (DOI) or an accession number that you can include in your data access statement? Once you have deposited your data, you should either create or update the record for the dataset in Pure. In the section 'Data availability', provide the name of the archive you used as the publisher, and if your dataset has been assigned a DOI, enter it in the appropriate place

» **Software (eg open source code) and documentation needed to access the data:** If you developed and used software to generate, process or analyse your research data, you may need or wish to make the software available as well as your data. The best way to do this is by archiving your software code in a suitable repository and making it available under an Open Source licence

» **Timescales for how long the data is to remain reusable:** Unless legal or funder requirements specify otherwise, data must be retained for ten years from the end of the project or the publication date of any research findings based upon them, after which retention will be reviewed

Even where the evidence more strongly satisfied the criteria, it may not be sufficient for a researcher to put into practice, however references to related tools and processes, such as repository submission, hinted that further guidance would be provided.

The graph on the next page captures the extent to which evidence was gathered against each element Findable, Accessible, Interoperable, Reusable, for each institutional guidance studied. The results are expressed as a percentage for each element, representing the proportion of the questions in the checklist against which evidence was found. It does not attempt to "grade" the evidence in terms of how well it answered the question.

What the graph shows is that a number of institutions provided some level of guidance against all (ie 100%) of the questions listed in the Horizon2020 DMP guidelines in the sections Accessible and Reusable. In contrast, few institutions provided details that were clearly identified as answering the questions on Interoperable, which was by far the weakest element.

Perhaps surprisingly Findable showed less coverage than Accessible and Reusable. This may be due to the very specific questions used in the Horizon2020 DMP guidelines eg listing naming conventions, versioning, registration and indexing. In contrast the questions in other elements may be thought to be more general eg "Identifying the location of the data and associated metadata through open access repositories".

The allocation of evidence against criteria is subjective, and as stated, the level of coverage is not graded. So for example, it might be argued that none of the institutions meet the criteria "Describing conditions for access (ie a machine readable license)?" as they don't refer to a machine readable licence. However, it was felt to be more useful to treat the evidence generously to provide general indicators rather than return zero coverage with a stricter interpretation.

Other limitations

These findings should be treated as a general indicator. In addition to the challenges of interpretation and assessment above, and the variety of sources, there are additional potential limitations to the method used:

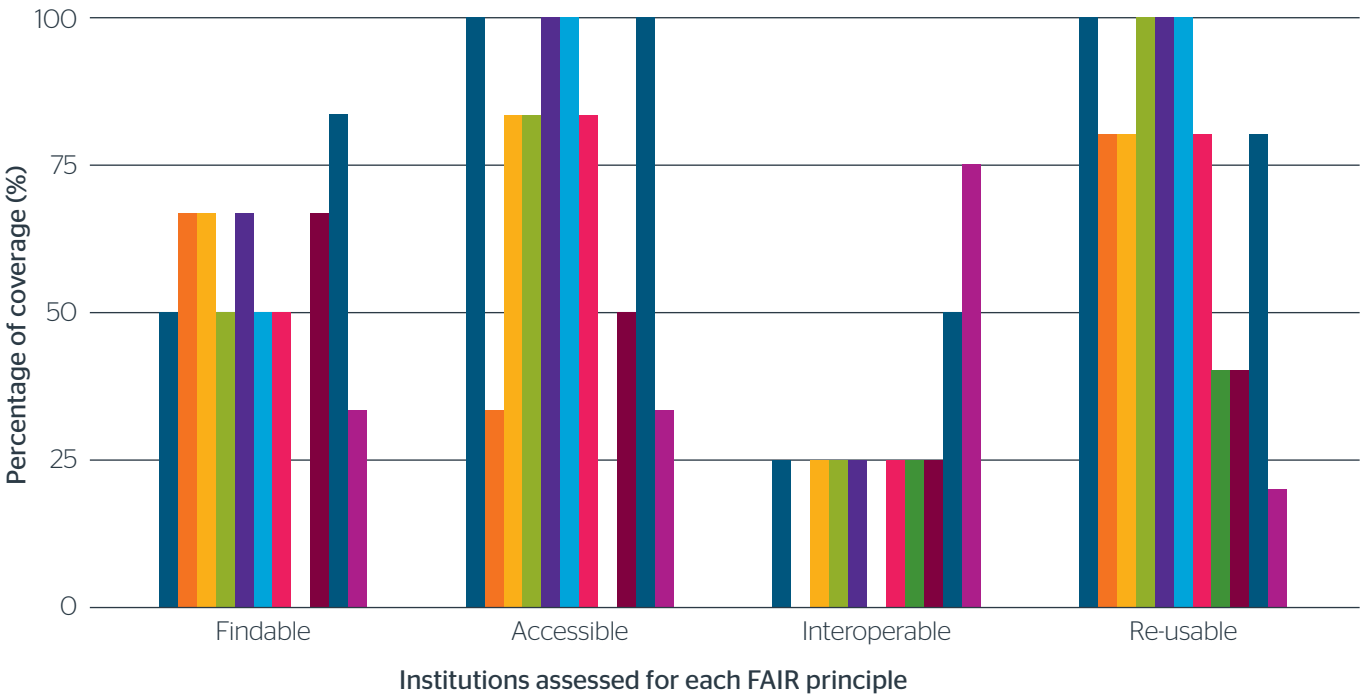
» Guidance may exist elsewhere on the website but wasn't easily located from the most obvious central point (such as research support service)

» Information may be hidden from public view in an internal document such as staff only intranets

[1]

57 dcc.ac.uk
58 https://dmponline.dcc.ac.uk
59 ukdataservice.ac.uk
60 rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf

Graph: evidence of F, A, I, R coverage in institutional RDM guidance



Other activities

Jisc research data network event

The project team attended one day of the Jisc Research Data Network workshop[61] at the University of York on 28th June 2017. The workshop included a very useful panel session entitled "FAIR in Practice, a reality check" which was chaired by Bas Cordewener from Jisc. The panel consisted of Ingrid Dillo from the Data Archiving and Networked Services (DANS)[62] in the Netherlands, Ingeborg Verheul from SURF[63], also in the Netherlands and Cameron Neylon who is a member of the FAIR in practice project expert group. The panel discussed some of the interim findings of this study with participants. The workshop also provided the team the opportunity to meet with the Jisc project staff as well as attending some other relevant sessions.

Experts group interviews and meetings

An initial group of data experts was constituted by Jisc staff in order to help guide and advise the project. The experts were interviewed individually and then brought together for an online meeting which also included appropriate Jisc staff.

[1]

61 jisc.ac.uk/events/research-data-network-27-jun-2017
62 https://dans.knaw.nl/en
63 surf.nl/en

The group members were:

- **Professor Simon Coles** - Director of the UK national crystallography service, University of Southampton
- **Dr Veerle Van den Eynden** - Research data services manager, UK data archive, University of Essex
- **Professor Carole Goble** - School of computer science, University of Manchester
- **Professor Cameron Neylon** - Centre for culture and technology, Curtin University
- **Professor Susanna Assunta-Sansone** - Associate director of the Oxford e-research centre, Department of engineering science, University of Oxford
- **Professor Melissa Terras** - Director of digital scholarship, College of arts, humanities and social sciences, University of Edinburgh, formerly Director of the centre for digital humanities, University College London at the initiation of the project

During the expert interviews, questions were asked about a range of themes. These included their perspective on:

1. The awareness, understanding and meaning of FAIR
2. The usefulness in principle and practice
3. Issues of responsibility within institutions over promotion, implementation and policing compliance
4. Levels of implementation and challenges
5. Motivations and differences between disciplines
6. Progress and timescales for adhering to FAIR
7. Known guidance, contacts and examples of non-data research artefacts

Similar questions were posed during interviews with individual Jisc staff with expertise across scholarly communications, research data management, metadata and curation, research and policy, e-infrastructure and shared research data services.

As a result of the online meeting the project was reviewed and a number of changes were implemented.

Additionally, the group provided assistance relating to:

- The range of disciplines that would be investigated:
  - Biological sciences
  - Digital humanities - including history and archaeology
  - Chemistry - including computational chemistry and crystallography
  - Social sciences - including sociology and longitudinal studies
- Contacting appropriate practitioners who might be willing to be interviewed
- Reviewing questions to be used in the interviews
- Promoting the focus groups

The results of the interviews and online meeting are summarised below.

A number of experts also attended the project focus groups and were asked to review the draft of the final report.

SWOT analysis of FAIR based on expert interviews and meeting

Initial individual interviews with members of the experts group and Jisc staff identified some key issues in the FAIR landscape that informed the early stages of the project. The main points from each interview were integrated and analysed around a SWOT[64] framework (Strengths, Weaknesses, Opportunities, Threats). The initial SWOT analysis was presented to the experts group in an online meeting, where the topics were discussed, and a voting mechanism was used to determine the priority issues and extent of consensus. The SWOT data is presented in appendix F.

The key strength identified of FAIR was as a well-defined, compact, clear, and concise brand. Other strengths were that it was easy to sign up to the principles to "do the right thing" and that it wasn't too technical, so valued as a communication tool. The discussion centred mainly around the "vagueness" of the principles, being at a high level, and whether this was a strength in leaving room for flexibility in implementation or creating uncertainty in how to implement. This last point connected clearly with the main weakness of FAIR, agreed to be the lack of implementation guidance, particularly across disciplines. Discussions included work on metrics and at what level the principles should/could be applied first, such as institutional or at community repositories for example.

By far the greatest opportunity identified was the sense of momentum in the FAIR concept, with international awareness and initiatives, and a common agenda. This was seen as also an opportunity to move beyond data and repositories, to non-data outputs and other services and tools, but also possibly to influence a culture change rather than just implementation of the principles.

The main threat was agreed to be the dangers of misuse of the brand, either intentionally or through ignorance, and the failure to engage researchers. Other issues identified included FAIR not being sufficiently part of research workflows and discussion more broadly about the purpose of FAIR. Distinctions were drawn between the FAIR principles and FAIR outcomes, and whether FAIR was most valuable as a direction of travel or practical guidance.

The discussion around these themes helped guide the subsequent stages of the study, in particular designing the questions for the practitioner interviews and focus groups.

Focus groups

Three face-to-face focus groups, with over 50 attendees, took place during September in London and Newcastle. One was aimed at research support staff, with the other two targeting research practitioners. The attendance at the support staff event was excellent. Despite the help of the experts group and considerable efforts from the project team it proved much more difficult to get practitioners to attend. However, the events were still very valuable with considerable useful data being gathered.

The focus groups took the form of interactive workshops using a range of techniques including small group work, presentations, guest speakers, flip charts, post-its and voting activities. The themes included the following, discussed in subsequent sections:

- Awareness and understanding of FAIR
- Current use of FAIR in policies and guidance
- Motivations and barriers for implementation of FAIR
- Examples of implementation across the FAIR elements
- Disciplinary differences and non-data artefacts

[1] 64 https://en.wikipedia.org/wiki/SWOT_analysis

Appendix C: RDM guidelines FAIR checklist

Making data findable by humans and machines, including provisions for metadata

1. Metadata – using what if any standards?
2. Persistent and unique identifiers such as Digital Object Identifiers?
3. Naming conventions?
4. Search keywords for optimising reuse?
5. Clear version numbers?
6. Registration and indexing?

Making data openly accessible

1. Making data and metadata open by default, using an open, free and universal protocol allowing authentication and authorisation?
2. Identifying clearly any reasons for restrictions on use or sharing tools needed for access?
3. Deposition in a repository with metadata available even if data is not?
4. Software (eg open source code) and documentation needed to access the data?
5. Identifying the location of the data and associated metadata through open access repositories?
6. Describing conditions for access (ie a machine readable license)?

Making data interoperable

1. That data and metadata should be interoperable - allowing data exchange and reuse between researchers, institutions, etc and combining with other data?
2. Data should adhere to standards for formats (compliant with open software applications, and facilitating re-combinations with different datasets from different origins)?
3. The use of data and metadata vocabularies, standards or methodologies (for all data types allowing inter-disciplinary interoperability)?
4. Providing mappings to more commonly used project specific ontologies and to other metadata?

Increase data reuse (through clarifying licences)

1. That data and metadata that they can be replicated or combined accurately with other research data?
2. How data is licensed to permit the widest reuse possible?
3. Timescales for the data to be made available for reuse (including any embargos for publishing/patents)?
4. Making the data produced or used useable by third parties, particularly after the end of the project, with explanations of any restrictions?
5. Timescales for how long the data is to remain reusable?

Appendix D: Interview template for researchers

- What sort of data do you work with?
  - Distinguish between primary, secondary and tertiary data (or just primary and processed)
  - Do you use existing data in your research?
  - Do you combine data from different sources?
- Where is it stored - in a repository, or where else?
- What format is it stored in. Are there community "standards" for data sharing?
- Do you know who actually owns your data? Do you have funders, research institutes, collaborators, etc that have explicit policies / demands on what you do with your outputs?
- Where do you turn to for guidance?
  - What guidelines do you follow in storing your data?
  - Who supports you?
  - What systems, services or tools help you?
- Are you aware of FAIR?
  - Where have you found out about FAIR from?
- What is your understanding of what FAIR means?
- What does/could motivate you to follow FAIR principles?
- What are the attitudes towards openness and sharing within your research area?
- What aspects of FAIR do you think you currently implement?
- What are the main challenges to implementing FAIR? Probe areas:
  - Understanding eg what does it mean
  - Practical eg knowing how to do it/what to do
  - Process eg integration with workflow
  - Technical eg use of systems/repositories
  - Cultural eg attitudes to sharing
- Questions on F
  - How can people find your data?
  - What documentation and metadata will accompany the data (assist its discoverability)?
  - Does it use a persistent and unique identifier such as Digital Object Identifiers?
  - What metadata standards do you use?
  - When using existing data - where do you find it, is it easy, what could improve?
- Questions on A
  - How is the data accessed? Is it openly available? Include what DB, repository etc
  - When using existing data - how do you access it, what could improve?
  - Is software or other tools needed? Include what documentation needed, what is made available with the data
  - What restrictions are there? Include embargos, IP, ethical/legal
- Questions on I
  - Is your data in a standard format that allows it to be recombined with other data or makes it compliant eg with open software applications?
  - Use of standard vocabularies and mappings?
  - When you use other existing data - What facilitates easy interoperability of data? What is lacking; what could improve?
- Questions on R
  - To what extent can others use your data during/after the end of your project?
  - When you re-use other data - are data easy/difficult to reuse; what is lacking for easy reuse; what makes data easy to reuse?
  - What timescales apply to release and lifetime for reuse?
  - What licences apply? Are any used, are they standard templates?
  - What other aspects of your research are necessary to allow it to be validated, reproduced or built upon (eg software, documentation, process information)

Appendix E: Interview template for funders and publishers

- What is your involvement/role in working with research data -
  - Do you work directly with data?
  - Do you oversee data submissions from others?
  - Or is it a policy role?
- What sort of data is it?
- Do you store, hold or archive data?
- Where is it stored - in a repository, or where else?
- What format is it stored in? Are there "standards" for data archiving and/or sharing?
- Who owns the data?
- Where do you turn to (or point others to) for support and guidance regarding data, when required?
- What are the attitudes towards openness and sharing within your organisation?
- Are you aware of FAIR?
  - If so where did you find out about FAIR from?
- What is your understanding of what FAIR means?
- What role do you see your organisation having with respect to FAIR?
- What impact do you see FAIR having on data sharing and reuse?
- Where do you feel your organisation is in adopting FAIR compared to other funders/publishers?
- Is FAIR included in any policies, mandates, requirements etc
- What are the main challenges and barriers to implementing FAIR? Probe areas: understanding, practical, process and workflow, technical, cultural

Followed by questions relating to each principle (F, A, I, R) as appropriate for their role/organisation.

Appendix F: SWOT Analysis from expert interviews

Strengths

		Vote
1.	Easy to sign up to as principles to do the right thing	3
2.	Well defined, compact, clear, concise brand	4
3.	Not too technical, communicates the concepts well	2
4.	Moves the debate beyond open access	0
5.	Links with the use of repositories and research collaboration agenda, builds on existing practice	0

Additional comments:

- The vagueness can be a strength, not tie people down too much. OA was tied to a definition, people became obsessed by it, damaged by focusing on the specifics
- Importance of machine readable/actionable as well as human readable

Weaknesses

		Vote
1.	FAIR not that widely known by name	1
2.	Open to interpretation, vague, not specific enough to implement	2
3.	Less clarity about dealing with non-data artefacts	1
4.	Lack of implementation guidance particularly across disciplines	4
5.	I and R are difficult	1

Additional comments:

- Different levels to consider this at - repository, institution... where will it be applied best first?
- Wanting to extend it, not the full story
- Lack of metrics to measure FAIRness

Opportunities

		Vote
1.	Lots of good practice that could considered FAIR	0
2.	Recognition by publishers and funders	1
3.	International awareness and initiatives, common agenda, momentum	6
4.	Developing metrics, tools and discipline-specific guidance	1
5.	Repositories as mechanism for enabling FAIR	1

Additional comments:

- Talking about why the principles are needed
- Extending FAIR beyond data (also a threat?)
- An opportunity to drive cultural change rather than "mere implementation"
- More broad than repositories - services, tools, etc

Threats

		Vote
1.	Difficulty translating enthusiasm to implementation	2
2.	Research culture of (not) sharing, competition	0
3.	Lack of incentives for researchers and links with career	1
4.	Not sufficiently part of research workflow	2
5.	It's open access again, failure to engage, brand may be misused	4

Additional comments:

- Which are threat to FAIR principles and which to FAIR outcomes? More about timing than balance of these
- What is FAIR for? What is the direction of travel - engage researchers or practical guidance?

Appendix G: Data from interviews and focus groups

This appendix consists of a distillation of the 20 interviews conducted with research practitioners representing the disciplines identified for this report plus the contributions from the three focus groups.

Data Types

Biological sciences data types

In the biological sciences, a variety of specialist data types are produced and used, including DNA sequences, genomic variation data, protein structures and sequences, and biological assays. In some cases this is the full range of data produced throughout the stages of the research lifecycle, from primary data from different instruments in the research facility through to processed data for the final published article. Very large quantities of image data is collected to create 3D reconstructions from 2D slices, for example the EMAP Anatomy Atlas of Mouse project[65].

The volume, variety and velocity of the datasets generated varies. For example a single biosciences facility routinely generates large amounts of a very similar data type. In contrast, long tail data represents heterogeneous small data sets produced by individual researchers from multiple different facilities. These unpredictable experiments, follow various lines of enquiry, with different designs and scope, leading to a wide range of resulting data.

Digital humanities data types

Data in the digital humanities can be extremely varied, a broad range of datasets, including from an access database, tabular data, rich visual data, raw data and processed data. The most common dataset type in digital humanities is text: encoded in XML, for which the datasets can be very large. The next most commonly used is digitised image data.

Specialist data types include Reflectance Transformation Imaging (RTI) for varying lighting conditions to reveal surface phenomena, and also Geographic Information System (GIS) or AutoCad data used in other fields. Many of the data sets are very large.

Digitised datasets are commonly used, can be large scale, and often cover a range of topic types, such as catalogue data, historical models, or economic data of 100 years ago or older.

In some areas of digital humanities data can be on a large scale, some of which is very structured (eg twitter data, or ALTO XML output from mass digitisation of newspapers) some not (web archives). It can include text, multimedia, metadata and derived data such as links. It can be existing data that it is captured from other sources and analysed as derived data, and can be data the researcher has created themselves, or gained access to from an institution, for example the datasets available through the OpenGlam (Galleries, Libraries, Archives and Museums) movement. An example is tables of the number of unique domains in a part year. Statistical information can then be an addition later, and this is reported to be a common way of working in the humanities. Data is frequently combined from different sources, including for example notes taken on analogue material combined with digital data.

In archaeological studies, it is often primary data, based on excavation, where the data becomes the primary research but primary sources can also be used too. Excavation data is the single source, so once the excavation happens, it becomes the record. In some cases this can be data extracted from samples of an artefact, such as numeric data from a mass spectrometer. Other researchers in a different research group may gather qualitative data on

[1] 65 emouseatlas.org

the same samples. One example given was both quantitative and qualitative data produced during the excavation of a skeleton. Categorical data describes the gender, age, etc., plus bone quality is assessed using scaling. Then in addition, a mass spectrometer can generic numeric data, comparing for example the isotype value of tall and short people. Samples run through a spectrometer provide raw data, then an initial processing is carried out immediately to get a primary data point per individual analysed. This is very precise to an internationally defined standard, and differs from mass spectrometry values in biosciences that are more flexible. Every sample is calibrated against standards and would be corrected. This is still considered primary data, as it is "tidied up" rather than interpreted.

Primary data is rarely used on its own, it may be compared to other people's data about the same kind of samples, or cross-referenced to primary historical sources, and other available datasets. Another comparison is one of interpretation comparing to other similar studies in the same geographic location or similar samples in other locations. In this sense these researchers are both producers and consumers of data.

Chemistry and crystallography data types

A very wide range of data is used by researchers from highly structured, raw, primary data produced by instruments to derived secondary and tertiary processed data. In addition, data produced previously by others may be combined with their own data. In some cases the data sets are huge, for examples x-ray crystallography imaging data. Other researchers produce software as a result of their work which is then made available to others.

One example is crystallography data and the 3D structure of a molecule. The coordinates of an atom in a molecule lacks a chemical description so this is added. Data is combined from different sources, typically software is used to mine data, analyse, extract knowledge and apply it in ways that might be relevant to what they are doing. Drug discovery is one of the most common applications - researchers extract information about certain properties of a molecule or set of molecules. This is used along with data from other technologies to design better drugs. The sophisticated workflow associated with this work is important in establishing the context of the data results.

In many cases published papers have the data associated with them which will need to be usable by the community.

The size of some data can lead to funding issues with respect to long term storage, but high levels of reuse make this very desirable.

Some researchers discussed the difference between datasets being human or machine readable, for example, images versus coordinates.

Social sciences data types

Data came from various sources and in different types for sociology, including surveys, interviews, questionnaires, diary templates, and also some activity monitors. Most data was collected digitally, such as a computerised questionnaires and computer assisted interviewing, although some was transcribed from traditional interviews. Longitudinal studies used large numbers of people, more than 10,000, and for diary data, it could amount to tens of thousands of days of data.

Data was collected at different points of time, for diaries this could be time intervals of five-ten minutes throughout the day. In other examples, there were intervals of several years between interviews. Data collection was sometimes subcontracted to fieldwork agencies and in some cases a data team was used to clean and process the data, producing some derived variables. Some surveys were repeated annually and had developed templates that eased collection and standard data. Some studies integrated data from other surveys, taken in other countries or at different times. One multinational study harmonised 75 national randomly sampled data sets.

Data storage

Biological sciences data storage

Biosciences data is often stored in special databanks and repositories rather than institutional repositories due to the large amounts of specialised data. Databanks, archives and repositories exist in a wide variety of specialist areas. Examples include PRIDE database for proteomics; the pep2pro dataset for plant systems biology; the Protein Data Bank for the 3D shapes of proteins, nucleic acids, and complex assemblies; GenBank for genetic sequences to name a few. The number of databases is said to be in the thousands, although they can be categorised into three types: global resources, comprehensive databases, and special-purpose or community databases. One example is discussed in the paper: **Databases for Microbiologists**[66]. The FAIRsharing resource provides an insight into resources across multiple disciplines, including the life, environmental and biomedical data resources, along with the wealth of data and metadata standards developed by the communities to better annotate and share data[67].

Global resources provide interconnected databases and tools with access to different data types. EBI is an example of this, with UniProt as one of the most well-known databases accessible through it. As well as providing access to other databases, EBI derive additional data to enrich the original raw data, providing different representations and processed data of interest to the sector.

Researchers are generally aware of, and encouraged to use, the public/community databases in their field. An example is a researcher who has sequenced a bacterium, and wants to compare this with other bacteria and look at how it fits with others, to do this they will need to retrieve thousands of similar sequences and thus would turn to a database specifically hosting bacteria sequences.

Raw data is often kept locally, and might be discarded. Usually only final data linked with publications is stored in community tools and what data goes into publications has been referred to as a "moving target". Long tail data was identified as a particular challenge and in that case, an institutional data repository that was "agnostic" to the type of data could be used. One example of this is the Edinburgh University DataShare repository.

However, in some cases, considerable quantities of data were held locally and sometimes not made available because of confidentiality issues. There was a commitment from the institution to keep these large quantities of raw data for the very long term, referred to as indefinitely, but often in the order of ten years.

Different analysis tools are available with different databases and sometimes the choice of database is influenced by the researcher's experience of using those tools and the perceived ease of use.

Digital humanities data storage

In digital humanities, working data is typically extracted from its source, sometimes extracting smaller subsets of the original dataset. In some cases the data is worked on in situ at the dataset host (such as at the British Library). The working data is sometimes stored on a local networked drive or the computers used by the research group, but some large data sets are not able to be hosted locally due to their size. It may mean that data is held locally, for example on high performance systems, but is too large to be open to other researchers because of bandwidth and infrastructure issues. Some datasets are published in an institutional repository in formats such as .txt, .csv and .xls, these being derived data, the final versions associated with publication. However, the data in general, across primary and derived versions can be stored in many different places.

[1]

66 http://jb.asm.org/content/197/15/2458.full
67 See: https://fairsharing.org

For archaeological data, the final dataset is often stored with the Archaeological Data Services (ADS), in keeping with mandates linked to regulations about building and discovery of archaeological remains. The journal itself provides the "story" about the data, the layer that describes what the data is, how it was collected and what the author thinks it means. Researchers are encouraged to use ADS, which is well known, and ideally have factored into their grant application that they will deposit with the service. Where to store the data may depend on direction from funding bodies, for example links between AHRC funding and deposition in ADS, or NERC funding archiving at the British Geological Survey.

For smaller projects, such as those undertaken as part of normal arts and humanities duties without the need for further project funding, there is still data created, but often no funding. ADS sometimes issues a waiver if they think the data is at risk, otherwise it may be placed into the institutional repository. However, some researchers report that institutional repositories are generally not taking the range of data archaeologists create such as CAD, laser scanning, GIS data, and only take text and images or small data sizes. In contrast ADS takes all data.

Submitted data may be in a spreadsheet, generally published as a supplementary data set - the publication with a supplementary information file. Researchers generally "hope" it has sufficient metadata to make it useful for others.

Chemistry and crystallography data storage

In the first instance data is quite frequently stored locally on an instrument, hard drive or institutional server with very little external access.

Some research data is held on institutional repositories but there are a small number of internationally recognised databanks such as the Cambridge Structural Database (CSD) which are essential to researchers. In some cases, funders insist that data is held in these repositories, for example 3D atomic protein structures.

A few rival data banks are being established which attempt to scrape structures from websites to set up competing collections but these are not as comprehensive.
There is an established relationship with journals where instructions to authors mention that they "should" deposit with these established databases. It can depend upon the publisher and journal as to how rigorously it is checked and this varies with crystallography being more specific and less so with general chemistry.

Social sciences data storage

The UK Data Service[68] is seen as a primary destination for most general data, and deposition there was thought to be a requirement for ESRC funded projects.

Exceptions to this included data that could identify individuals, and in particular an example was provided of a study using diaries and cameras, where the visual image data was considered to be more sensitive, and data sets kept in a highly protected data repository within the university. Data was generally anonymised where possible, and a separate file detailing the respondents was kept independently, on a computer not connected online and stored in a locked drawer. Where surveys included open ended questions in addition to quantitative data, the former was often not made public, but on request due to confidentiality concerns.

Some large scale studies involved data from other sources (including other countries) in addition to the primary data collection by the researchers. In these cases, such data was not deposited with the UK Data Service, instead a dedicated archive for the entire collection of data was maintained, with the research group acting as a repository (and a data disseminator). This was considered to be secondary to the research role of the group but nevertheless important, providing a more complete set of data, including raw data and derived data sets, and harmonised versions of the data (using the same variables) from across different data sets. As an illustration of the importance, it was reported that national statistical institutes from other

countries sometimes call on the archive to ask for data they have lost.

Some archives are stored on university maintained and protected file storage, others in their own universities or national statistical institutes. In some cases registered uses require a password to access the data which is held in processed excel spreadsheets.

Ownership of data

Biological sciences data ownership

Data ownership across the biological sciences was described as a "complex landscape" with "a lot of challenges". Often researchers were said to be "blithely unaware" on data ownership and it was difficult to train them without swamping people. The 2016 Concordat on Open Research Data[69] was seen to be a very positive development, but was yet to be implemented and interpreted in research council guidelines. If guidelines became more harmonised there could be a huge benefit.

Distinctions were drawn between PhD students and staff, due to different contractual arrangements. Funders have different policies on ownership and it can also depend on the research centre that manages the funds. Funders in some cases have ownership policies that students sign up to but these can be varied, and people can unwittingly move from one data ownership regime to another because they have included another funder. The Engineering and Physical Sciences Research Council (EPSRC)[70] are reported to have much more stringent data dissemination guidelines than other research councils. As an example, an evolutionary biologist funded by The Natural Environment Research Council (NERC)[71] could collaborate with someone at another institution who is an EPSRC PhD student. The published paper would acknowledge EPSRC funding and so follow EPSRC policies.

Staff are generally subject to terms on ownership through their employment contract, so the position is legally clearer. In plant sciences and systems biology, the range of understanding on data ownership is very large, some groups are very sophisticated, working with their knowledge transfer office, others less so. There was reported to be limited commercial interest in most cases and data ownership was not thought to be a big consideration. Groups with commercial interests are said to be much more aware of data ownership and undertake training to sort out their position at an early stage.

Some data services providers view the data as being held in trust on behalf of the researcher, expecting the researchers to be in a position to submit that data legally.

Data may be made freely available from data providers under creative commons licences, even where this data is derived from other data. Such data is viewed to be not owned by the institution or funders, in fact funders are said to insist it is made available to the public in some cases. A distinction was drawn between data and algorithms, the latter or the software needed, are sometimes available as charged services.

When working on drug trials for pharmaceutical companies there are often strict contractual arrangements with respect to data ownership.

Digital humanities data ownership

Ownership in the digital humanities was seen as equally complex or misunderstood. Interviewees referred to a wide variety of arrangements and concepts including the legal deposit legislation, copyright laws, and ownership of general public archives. Some research was said to involve "mostly third parties" or "nested ownership". Other researchers

[1]

68 ukdataservice.ac.uk
69 rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf
70 epsrc.ac.uk
71 nerc.ac.uk

openly admitted to have "never thought about" data ownership, but had a vague notion that it was owned by the researcher. It was questioned whether their institution would know of the existence of much of the data, or the scale of it. In some cases the data was considered to be owned by the data source for example records and printed material held by the British Library[72] with their institutions having a licence to use it. Where images have been created the ownership was thought to lie with the creator but this is not always clear to those we interviewed.

When asked about ownership, funding councils such as the Arts and Humanities Research Council (AHRC)[73] were reported to have a general commitment to make data open, in the absence of other restrictive permissions. Data deposition sometimes expected a description of how researchers planned to make it open as long as possible, but a formal commitment to five years after the end of the project was common for some, and ten years for Jisc projects.

One internet journal reported having switched to creative commons licensing[74], using CC-BY by default but with "ownership retained by authors". The Archaeology Data Service (ADS)[75] was thought to operate in a similar way and others reported common use of creative commons licences.

Ownership was said to depend partly on the nature of the data - this was generally not commercially sensitive but sometimes other factors were considered. Some archaeology data may be generated by commercial companies but not used for commercial gain. In other examples, the archaeological units were funded by property developers, such as where archaeological remains have been discovered during construction and an archaeological dig is required by law before continuing.

Chemistry and crystallography data ownership

Some researchers were not completely clear about the ownership of data but most felt that it belonged to their institution, along with any resulting intellectual property (IP). If the data resides in the institution repository it is often assigned a creative commons licence with attribution. In the case of software a GNU license[76] may be applied. With charitable or publicly funded research the ownership can be particularly unclear. Some claim ownership for three years then the data is open. Often they have no interest in ownership.

Ownership was also reported to be restricted in some cases until after papers are published.

It was noted that data from student's work may be handled differently since there is no employment contract with the institution. This means that students can retain ownership of their data.

Data held on national or international repositories such as the Cambridge Crystallographic Data Centre[77] often isn't assigned a formal licence and is made available for anyone to use freely on an individual basis. Deposited data typically has no statement about rights and the data centre is considered to be not in a position to assign licences but it is thought this should be done by the person depositing. There are some value-added datasets and licences over the software and the enriched database, which are licenced commercially to academia.

Some free services are made available such as look up of datasets, but are sometimes used differently and restrictions are applied on the services rather than the data.

Ownership is an area where it is acknowledged further work and clarity is needed.

Social sciences data ownership

The researcher's institution was generally thought to be the data owner, although ownership was described as a "relative term". This in part is related to the involvement of third parties for data collection, and funders. The data collectors were said to designate the research group as data owners, passing confidential material about identities

to them to maintain. ESRC as the funder were thought to hold a share of the IP by some researchers, but others were aware of clear contracts that IP was retained by the institution even where data was shared with the funder. The research group in one case considered themselves to be "custodians" of the data on behalf of the research community, and this was "public owned". Where international data was collated the ownership issue was very complicated, for example, data from the USA was completely open while Australian data was very restricted in terms of ownership.

Formats and standards

Biological sciences formats and standards

FAIRsharing.org is a *"curated, informative and educational resource on data and metadata standards (inter-related to databases and data policies)."* Although covering several disciplines, its main content is for the life, environmental and biomedical. There are a large number of community standards in these domains.

Across biosciences as a whole there are many different community standards, specific to the data types. Examples include the PDB file as a standard representation for macromolecular structure data; protein structures commonly use the Stockholm format;

Global resources such as EBI refer to community standard libraries to ensure data they are processing adheres to "sanity checks" in those libraries with curators involved in this adherence.

Different communities are observed to be at different levels of maturity of organisation with their data - where data types are easy and consistent, it has been easy for communities to organise to form standards. Others with diverse data types have found it hard to organise standards.

Some data types are very well supported for example nucleotide and protein; for example, the HUPO Proteomics Standards Initiative has been very active in developing and using standards for data and metadata sharing[78].

Here data types are relatively narrow. The challenge in biology is reported to be heterogeneity with the wider variety of data types, where there is only really large volumes in a couple of areas such as DNA sequences.

Only rarely there are conversion tools between standards. Database developers and maintainers, as well as data users and creators are often the drivers behind community-driven standards initiatives, providing use cases for the development process. They ultimately are also consumers of the produced standards, for example, variation data using the VCF format. The community-efforts also include other stakeholders, such as vendors of equipment - where the equipment generates a certain data type, software manufacturers and knowledge engineers. The success of a standard depends upon its adoption, and clearly if major data repositories implements it then data annotated to that standard becomes available.

It was noted that this is difficult to achieve the high standards expected to adhere to FAIR. Especially getting good metadata depends on a range of context dependent factors , for example the availability of annotation tools that hide the complexity of the standards to the users, and the presence of a curation team assisting users to provide all necessary metadata elements during submission to a data repository.

Automation is desired by many researchers, aspiring to have user friendly annotation and submission tools to pipeline richly described data to data repositories. When dealing with large volumes of data, adherence to standards can be

[1]

72 bl.uk
73 ahrc.ac.uk
74 https://creativecommons.org/licenses
75 http://archaeologydataservice.ac.uk
76 gnu.org/licenses/licenses.en.html
77 ccdc.cam.ac.uk
78 https://fairsharing.org/collection/HUPOPSI

challenging even with suitable tooling, as manually checking for compliance is impractical. In some cases there is a lot of nuance in the metadata that is needed to fully understand the nature of the data and results. For example, in studying disease, there is a need to know enough about the patient, the progression of the disease, diet or other factors.

Digital humanities formats and standards

Standards were reportedly a challenging issue for some areas of the humanities, in particular history, where sometimes it isn't even recognised by researchers that they are using "data". Interviewees reported a lack of relevant standards and some opposition to standardising, viewing this as constraining the way they work. More broadly in humanities there is usage of RDF, XML, WARC, .csv and .xls formats, and the choice of format varies depending on the research methodology, the individual project and in some cases personal preferences. Guidelines used include the Text Encoding Initiative (TEI)[79] on XML encoding. Images may use an uncompressed Tiff formatting, Jpeg, Jpeg2000 or .png, with imaging standards provided by the Federal Agencies Digitisation Guidelines Initiative (FADGI)[80] at the Library of Congress.

In addition to formal standards such as TEI and FAGDI, in archaeology, guides from the ADS provide good practice and there are some de facto standards accepted by the community. These are typically developed in high profile articles about good practice and then become widely cited when describing data management methodology. Rather than a formal standards process it is more of a recognition reflecting best practice to make sure the wider community is following it. There are other more technical standards used to calibrate equipment against. This is important in terms of the performance of analytical instruments, and impacts on repeatability. The samples need to be of good quality and standards are used to screen the data and judge the quality of the wider dataset.

Standards can sometimes be expressed in free text in terms of acceptable levels of quality eg *"I'm unhappy if the errors are over 0.2"*. Where mass spectrometry is used, guidelines are published in the primary journal as best practice articles. Isotypes are used in other fields in different journals, but the mass spectrometry journal is considered the home for good practice with this data.

Chemistry and crystallography formats and standards

The use of standards is variable across chemistry and crystallography.

Some standards are based upon legacy formats from Fortran work. The 1996 XML (CML) schema is also common use as are NetCDF and ASCII.

There are many proprietary formats produced and licensed by instrument manufacturers.

The field of cheminformatics tends to use standard open formats which allows pharmaceutical companies to compare data. The UK company Medchemica specialises in this area.

The International Union of Crystallography maintains a file format - the Crystallographic Information File (CIF) as the internationally agreed standard file format for information exchange in crystallography. This is used across crystallography, representing the results of the experiment but also information about the experiment, considered "metadata in the broadest sense".

The CheckCIF service is used prior to publication or at review to do some integrity checks of the data such as minimal level of reporting. However, it has quite low minimal expectations of data so if data could be useful it will be given a home.

In some cases journal publishers will request standard formats for data and apply automatic checking for these formats.

Social sciences formats and standards

Commonly used formats in sociology include SPSS and Stata, or the underlying data such as a symmetrical structured data set. Also the Data Documentation Initiative (DDI)

standard describes data that result from observational methods in the social, behavioural, economic, and health sciences[81]. It is not thought to be important which of the common types are used for submission as technology now allows easy translation between for example Excel, SPSS or Stata. Software often uses proprietary formats, even though some of these are in wide usage, and some offer more advanced database analysis such as SIR (Scientific Information Retrieval) database, but may require specialist programming skills.

In one case much of the research involves converting data from various countries into a international standard format. In another the data from interviews is held in Microsoft Word and processed using NVIVO. Also standard image, video and audio formats are used for example, Jpeg and Mp3.

Where data is sourced from other studies, the formats can vary by country and by study, including for example raw ASCII .txt files, structured ALTO XML files, or TEI encoded textual data sets. Large scale studies using other data often apply harmonisation to the data for example to use standard classification systems for variables using a "lowest common denominator" system to allow comparison across studies, or metadata and schema cross-walking to allow automated comparison.

Guidance on data management

Biological sciences guidance

Institutional support was seen to be good for some researchers, said to be specific enough for their work. Sometimes the expertise already exists within the research group, and is shared with other groups or by responding to specific requests for help. Posts such as data manager and specialist bioinformatician exist in some, but not all, research groups. As an example the proteomic facilities manager would provide advice on depositing to the Pride database[82]. If the data was not from a facility then the university research data services would assist. In many cases, research projects use more than one type of data, so some would come from a facility and some not. It was noted that there was a challenge in knowing what data goes where, and established research groups try to help researchers, or signpost the university services, sometimes seen as a "backstop".

Providers of national data services typically have help desks and offer training courses and guidance documents in topics such as accessing or analysing data. "Biocurators" may help by monitoring such services, and tools also help by providing structure to the way data is managed.

National consortia and open community initiatives such as Bioschemas[83], which helps in tagging web based datasets to improve findability by search engines, provide networks nationally to help discuss some of the issues around managing data. Resources like FAIRsharing help to set up and, make discoverable, community standards.

Digital humanities guidance

For some humanities researchers, central university services were considered a good source of guidance, however for one this was in the form of a digital preservation unit that was "more commercial". The researchers felt that they were "low down the pecking order" for access to this service and tend to have to pay for it.

The Digital Preservation Coalition[84] was also seen as having useful resources such as a handbook[85] providing "digital preservation 101" instruction on how to prepare data.

[1]

79 tei-c.org/index.xml
80 digitizationguidelines.gov
81 https://fairsharing.org/bsg-s000605
82 ebi.ac.uk/pride/archive
83 http://bioschemas.org
84 dpconline.org
85 http://dpconline.org/handbook

The deposition process for submitting to an institutional repository often specifies the file format but otherwise provides no guidance on for example what you should think about to make your data more easily reusable. National data services such as ADS provide a helpdesk service and documentation, and require metadata and use of templates to help with data preparation. Some felt that researchers turned straight to such services rather than looking to their institution for support, but this depends on the nature of the project and institutional involvement, or the role of the individual.

In some research areas there is a lot of debate going on about how data should be stored with the huge amounts of data and associated cost. The IT support and research computing officer was seen as one key contact point, others referred to peers providing guidance.

Some researchers referred to the lack of tools available to help guide them through the process of effective data management, and how the vast majority of people in the field generate data in excel spreadsheets initially.

The Jisc web site[86] was mentioned as a source of support by one interviewee.

Chemistry and crystallography data guidance

Most researchers have some access to support from their institutions generally though specialist library roles or research support services, and it was normally considered of a high quality. In one case a post was created in the run up to the research excellence framework (REF)[87] which encompasses FAIR support. One interviewee mentioned that institutional support was particularly useful with respect to intellectual property and licensing issues. One interviewee mentioned that their institution's spin-out company service has been helpful.

Some of the interviewees saw themselves as the main source of support for their group or department, their in depth knowledge and interest in the field means they were happy to provide this support even though it was not part of their formal role.

Online support was considered important as were international groups such as the International Union of Crystallography[88].

The Cambridge Crystallographic Data Centre (CCDC), which provides the Cambridge Structural Database[89] (the world's repository for small-molecule organic and metal-organic crystal structures), provides some implicit guidance within the deposition process. This includes syntax checking, checking for process data and to seek justification where it is absent. Additional metadata is also asked for along with validation checks.

Social sciences data guidance

One primary source of guidance for social science researchers was the UK Data Service[90], providing information on technical reports and variable lists for example. Another source was data management course of the data team in their own institution. Research groups also tended to liaise with other studies doing similar things to share good practice. In other cases, for large scale studies, it was felt there were no institutional services that covered what they did, instead best practice was learnt via annual meetings of specialist sector research groups, sharing with colleagues across the world in similar work. Some researchers felt that interaction with local library and research services created too much of an admin overhead to be worthwhile and some turned to Google searches for information online. More recently, participation in Eurostat[91] was seen as useful.

[1]

86 jisc.ac.uk
87 ref.ac.uk
88 iucr.org
89 ccdc.cam.ac.uk/solutions/csd-system/components/csd
90 ukdataservice.ac.uk
91 http://ec.europa.eu/eurostat

Awareness and understanding of FAIR

Biological sciences awareness and understanding

Awareness of FAIR was from involvement in working groups and committees (such as ELIXIR[92] and EPSRC[93]) and in some cases through colleagues or more widely on Twitter.

FAIR was understood as set of principles (a "meme"), to enable data sharing and make data easily available for reuse (but recognising ethical and legal constraints), and to implement open research in practice. It was felt that "if you are trying to do good science, you need to do all these things", in one case a researcher felt that it was almost immoral not to follow FAIR although the issue of interpretation was also discussed. It might be seen as a researcher's legacy.

However, not all researchers are aware of the FAIR principles. Some only have knowledge of open access to data which takes place two years after the first research paper is published.

Digital humanities awareness and understanding

Awareness of FAIR has tended to be quite recently from colleagues/research collaborators or from national events or workshops including those provided by the Software Sustainability Institute (SSI)[94]. The events have usually been around topics such as good practice in archiving or big data, rather than humanities or history more generally.

As a set of principles FAIR was thought to be useful and sensible, helping promote reuse and making data fully open or findable, making sure the data is there, people can find it, know what to do with it, and know it's there to be reused, including being machine readable.

Whilst individually these aspects have been discussed previously, FAIR was seen as the first "whole process" approach.

Chemistry and crystallography awareness and understanding

The level of awareness of FAIR in the chemistry research field seems to be variable. Some chemists have been there from the start and had an interest in the principles long before the FAIR acronym was created. Indeed some chemists helped in creating the principles. However, this is not true across the whole discipline. For one interviewee our contact was their first introduction to FAIR. One description of FAIR was that it was seen as a recognition that just making data open on the web wasn't enough to enable people to reuse in a meaningful way, so allowed a layer of services and processes around it that made it genuinely reusable.

There was a high level of understanding in general with a focus upon data needing to be reusable (exploitable) with a minimum of effort. Crystallographers in particular felt they were at the leading edge with standard format data being available from known repositories.

International repositories used by researchers have been available for a long time, for example the Protein Data Bank[95] was established in 1971, and is well known and highly respected by those in the field looking for research data.

Those who were new to FAIR were, none-the-less, familiar with the broad concepts.

Social sciences awareness and understanding

There was less reported awareness of FAIR in sociology than in other disciplines, although the underlying elements had some well-established good practice. Where it was known, FAIR was understood to be principles about data access. It was thought to be a complex area, and FAIR was an attempt at "simplifying and boiling down the key things in a way that is understandable". Some felt that FAIR simply encapsulated their existing and well established practice.

[1]

92 elixir-europe.org
93 epsrc.ac.uk
94 software.ac.uk
95 rcsb.org/pdb/home/home.do

Motivations and challenges for FAIR adherence

Biological sciences motivations and challenges

Motivations and practices were described as hugely variable within biological sciences. However, motivations and challenges reported included three broad areas of: impact, policy and culture/skills.

Impact and altruism

- Re-use of data was thought to lead to increased citations and dissemination of scientific work so generally desirable
- Open research as an end in itself, also characterised as altruism, love or doing the right thing in giving the best data, was seen as motivating some researchers
- However, it was also suggested that there was "almost no immediate return to any researcher for following FAIR principles". There was thought to be a small amount of kudos from evangelists in open data community. Some institutions were working to provide institutional mechanisms for recognition
- In general researchers were thought to be willing to share if they can get their publication. In some fields, such as epidemiology, they were reportedly not so keen
- The motivation of altruism was said to be limited in comparison to the imperative to publish, and the argument that research isn't finished until the data is openly disseminated was said to be an arguable position but rare and depends on the definition of "good" research
- However, an alternative view was presented that good data management was a necessary overhead, and to cut corners leads to significant problems later on

Funders, publishers and policy

- Requirements from funders and publishers were seen as a driver but different strategies and funding mechanisms across the funders led to challenges. Mandates alone were seen as insufficient, need to be supported by systems, for which there is now some investment
- The concordat on open research data was seen as a positive development but needed continued momentum, and cultural changes at the level of the e-infrastructure leadership council[96] needed coordinating
- A number of practical issues, such as the level of heterogeneity in biological data, were identified, and communities were thought to avoid difficult issues unless it was clear there is a consistent long term funded direction of travel
- Groups such as experimental researchers were reported to be struggling with these issues, lacking resources in their grants and personnel with the right skills and training. It was suggested that research councils need to allow costs to be requested for researchers to make data available and disseminate
- The lack of a consistent set of FAIR metrics was seen as a barrier, needing the process and metrics that determine what is FAIR

Skills and culture

- Some described a big overhead to ensure data is handled properly, but recognised a push towards making it standard. Overheads in storing and making available increasingly large amounts of data were also identified, with difficulties getting funding for maintenance of data services in some cases, rather than new features
- Some networks and groups were collaborating to bring data together and ensure it is publicly available and disseminated, eg Genome 3d[97]

- There were references to data management not being part of the culture in some areas (eg experimental biology), and that researchers needed to be trained, and universities "needed to get better at this"

- It was reported to be really difficult to manage good data practice in schools of 130 PIs who work on everything from regional scale ecology to atomic scale. The variety was highlighted in the differences of reproducibility between someone who can replicate an experiment in a couple of days for trivial costs compared to someone studying a 40 year pedigree of sheep, where each season is a completely unique dataset never to be repeated

- It was felt that to be a successful scientist you need to know how to analyse and manage your data, and have some awareness that if someone sends you something, know that the quality is good

- In most cases datasets are said to be made available in reasonable time, however, sometimes part of the data is held back, with the expectation of a further paper to be written

- Researchers were seen to not always have the necessary skills in terms of how or why sharing is important or done, but a tension was identified between training researchers and investing in core facilities and services to submit to journals in standard formats

Systems and infrastructure

- It was suggested that RDM systems must give immediate benefit to that researcher doing the work to avoid the perception of "wasting their time on data management". Some benefits are seen to accrue over a long period of time, but sometimes to the principal investigator (PI) and the benefits needed to be more immediate

- However, it was proposed that with an appropriately designed system, you can get data management as a side effect of day to day research practice

- Research computing infrastructure was felt to be more prevalent in physical sciences - coordinated projects that provide both tools and data management eg collaborative computational projects (CCP)[98]

- Systems such as Biodare[99] (Biological Data Repository) entice users to manage their data by providing tools they have to use to get their research done. However, individuals still attempt to subvert the good intentions of the systems designers by finding ways to avoid appropriately entering data (such as entering "xxxxxxxx..." to appear to have provided a description)

- It was felt that people who provide repositories are typically central and cannot provide the domain specific tools that would offer immediate value, but this can be seen to be an organisational problem, not a software one

[1]

96 gov.uk/government/groups/e-infrastructure-leadership-council
97 http://genome3d.eu
98 ccp.ac.uk
99 biodare.ed.ac.uk/robust/About2.action

Digital humanities motivations and challenges

Motivations and challenges for researchers to adhere to FAIR in the digital humanities fell into several areas:

Impact

- Impact - it was considered to be a "fantastic result" if others were using your data, and some researchers liked the idea that others used their data. Impact was said to be being severely limited if data is not findable. Furthermore the question was posed "what's the result of research if it's not findable?"
- Reproducibility was not thought to be something that people generally think about, as they don't anticipate anyone re-running their research. Re-use was more of a driver
- However, making data available or it's reuse was not the primary form of impact people think about, for historians it was usually about publishing a book, or running an exhibition

Using good data

- Benefits were identified from open data that could be used to integrate with their own data. Researchers that were both 'producer' and 'consumer' of data want other people's data accessible "in the best way that suits me" and it was felt that it would "make my life a lot easier if it were FAIR"
- However, other researchers were not always asking the same questions and data may not have been collected with that research question in mind
- Manually finding and adding metadata was seen as very labour intensive, with big data sets across the world
- There was some reluctance to spend time on making data FAIR as it was not easy and doesn't bring any apparent benefit to their individual career
- The lack of adherence to shared vocabularies was seen as a key barrier
- A particular difficulty with archaeology was that by necessity data has gaps, rather than a slick dataset that will answer all the questions
- Other specific issues reported for example with art historians or literary researchers were problems around individual estates not allow them to re-use or publish data as it is not owned or created by researchers

Skills and culture versus tools and services

- The importance of metadata was recognised, to properly document data, and that the awareness was higher but usage not yet widespread. Students as researchers of the future were thought to be not as aware as they could or should be
- Researchers were said to not be quite ready to fully embrace digital culture, although awareness of issues of findability and accessibility had improved. Interoperability was seen to be valued when people can see it works
- In one archaeology journal the switch back to open access was reported to be well received, and a positive attitude to data that is open and publications being open access. However, this was not necessarily accompanied by an awareness of what that means technically, in how things would be interoperable
- There were expectations on data services and journals to "do it for them", such that individuals were less concerned and left it up to others. It was felt that ADS for example had more templates and was getting more strict on what is submitted, but people were "not thinking about what the end product would look like... not valuing the things needed"

- It was suggested researchers should contextualise data to make it accessible, but that this needed technical support through the research process

- Some institutional level advice and services were seen to be of limited value in some specific disciplines, for example around copyright issues for humanities data that was often not owned by the researcher. However, local advice seemed to be geared towards an approach that minimised risk at the expense of sharing

- Researchers were generally thought to agree in general on the value of openness and sharing, but historians were said to be very cautious sharing research before they've finished with it. They were reported to not want to put it in the public domain in case it may feed into another article or a book later. There was a sense of letting it out of their control and not wanting others to use it

Chemistry and crystallography motivations and challenges

Various motivations emerged across chemistry and crystallography.

Firstly the attitude of the funding bodies was identified as influential. However there was some scepticism about seeing this through to implementation. Some funders ask for adherence to FAIR like principles but don't follow-up or police these principles.

Some are motivated by the requirements of repositories whether they are international or institutional repositories. One researcher helped create their institutional policy for submitting data to their repository.

The need to be able to access and reuse others' research is important to many researchers. This is for some a motivation to make their own research available. One researcher was very keen to reduce the need to search every institutional repository by having a joined up cross institution search facility.

Views were expressed by some that researchers needed to be aware of the need for their datasets to be made available, however, the work in preparing the data should be enabled or done by experts. These could be within the institution or at the domain repository. It was felt that researchers weren't motivated to do it themselves, and perhaps shouldn't need to be.

The competitive nature of research is an issue for openness and sharing. The protection of IP often inhibits openness particularly where researchers feel they are in a race to solve a structure. This means there will not be any sharing before publication. Even after publication some data may be withheld.

However, some individuals are much more motivated to share, perhaps those towards the end of their career or who are very secure in their position. Also it was implied that naivety about the value of their research plays a part.

A general consensus emerged around cultural attitudes being a major challenge and the need to change researchers' attitudes. Some suggestions were made in relation to early career researchers, particularly PhD students and the need to win their hearts and minds. One suggestion was that the registry might withhold a PhD if the data is not demonstrated to be FAIR, or the PhD review process could include a FAIR assessment. Peer reviewing could check all the data for reusability or one interviewee suggested there was a role for publishers in ensuring FAIRness.

The technical challenges were also discussed. Some data sets are very large making them difficult to be accessible. The issues of the cost and resources to handle these large data sets were also mentioned. It was stated that levels of investment to develop infrastructure were needed that enables people to make the data usefully available. This was thought to be not a trivial job and that it was unwise to sit back and expect researchers to take care of it.

One researcher felt that there would be a major challenge in encouraging researchers to spend more time "in front of a screen" in order to make data FAIR.

Social sciences motivations and challenges

The concept of "good science" was suggested to be well embedded in sociology, including effective preparation and management of data. Sharing and openness was reported to be core to research, but even deeper than that, reproducibility was considered an "essential characteristic of science activity". It was stated that social science cannot be reproduced unless the data on which it is based can be shared and sharing the data for any survey activity is fundamental. Others referred to the motivation as being a "moral" one, linked to tax payer's funding, the time investment from research participants, and increasing the body of knowledge in social science. There was also a recognition that their own research group can "never do enough" with the data and by making it available to others extended the opportunities for research.

There was evidence in some places of protectionism, but this was said to lead to "bad science" and thought to be counterproductive, so discouraged. There was an expectation to get first publication out of the data that was collected, and have first access to it, but otherwise it was assumed that release would happen within a few months.

Some strongly critical language was used to describe practices that didn't aim to prepare good data and share it, including "laziness" and "shameful". It was felt that there was a culture of teaching students good practice including making available survey data as soon as the work is published.

In some particular cases, where the nature of the data means that individuals can be identified, confidentiality is a real barrier to sharing. This remains true even where the data is anonymised, but there is fear that by combining other pieces of data it might be possible to identify an individual.

The time to properly curate the data was thought to be a potential barrier, needing to be included throughout the research processes. However, for well-established longitudinal studies and annual surveys, repeated methodology had led to a streamlined process that incorporated good data management with minimal additional effort.

Researchers reported that designing questions for surveys involved wide consultation with the research community to make the data multipurpose. This includes what questions to ask and what data to collect, not just driven by research interest of the immediate team but to collect data that would have broad scientific utility. Sharing and open access is often assumed, with a long standing tradition in the field, and reported to be a funder requirement.

The approach was contrasted with biomedicine, which was felt to be more investigator led, with a tradition of publishing papers and using multiple authors.

A key challenge identified for major longitudinal studies was sharing practices across countries. Other agencies such as national statistical institutes reportedly don't share the same principles, with some not making data available. Some of this is attributed to national policy-level attitudes and regulations. Examples given include:

- German law restricting access to data sets to 80% anonymised sample, giving German researchers the ability to produce more precise results
- Danish researchers stopped sharing data, arguing restrictions on release of data from the national population register, even where it is anonymous
- Cycles of withholding data from other Nordic countries such as Sweden and Finland due to changes in commissioners

» Japanese research groups collecting and refusing to release data to any non-Japanese researchers

Solutions to overcome some of these restrictions include passing the processing of the data to the host country, however, this is found to be quite inflexible so is resisted. In other cases a third party country processes the data and returns the results, but this is also felt to be cumbersome. Again, the data is doubly anonymised so the practice is felt to be restrictive.

Non-data research artefacts

The application of FAIR to non-data items was considered to be an important topic to explore. Many disciplines produce other non-data artefacts and materials, this could cause barriers to FAIR, or at least perceived barriers. The issue of non-data research artefacts was discussed in all of the interviews. It is clearly an important issue and most of the experts felt that the FAIR principles could and should apply to non-data material. In fact it was felt that FAIR provided an opportunity to improve on the currently situation, particularly since these artefacts were often ignored.

A wide variety of types of material was described and discussed, although the relationship to FAIR was not always apparent. Materials mentioned included artefacts that support the research life cycle and can be used to help validate and reproduce research. Indeed they may be needed to allow the research to be built upon. Examples mentioned include: laboratory notebooks, software, reports, workflows, publications, documents, manuscripts, photographs, physical objects, samples, mathematical models, performance, equipment types and settings, video and images of many kinds. It was acknowledged by many interviewees that this is an area of weakness in terms of true reproducibility, due in part to the lack of complete data, insufficient process description or context, or poor metadata capture - either human or equipment. In some areas, such as sociological longitudinal studies, it was felt that complete process information was included including questionnaires and user guides, to provide a "large and comprehensive description of the survey".

The examples above such as data papers also start to show how other elements of the research process can be documented and published to aid reusability in a FAIR way as well as gaining academic credit. In data papers for example, descriptions of the data purpose, context, collection and methodology are covered, and a DOI to the raw data. In biological sciences, the analysis process may be documented within the paper or provided in a supplemental file in publications, or scientific workflows deposited or code shared using eg Github.

Guidance from national tools and services helps researchers to identify how to document other aspects of the research process to aid reusability and interoperability, for example the Digital Curation Centre provides guidance for data management planning that includes suggestions on what to consider:

> *Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.*

Another issue mentioned was that much of the non-data research material was produced through non-funded, or alternatively funded, research. This means that it is more likely to be poorly curated, with little in the way of metadata, and may not make it into a repository.

Findability, accessibility, interoperability and reusability are all challenges for non-data research artefacts to an even greater extent than for traditional data outputs.

Appendix H: Implementing FAIR - Section 1

Section 1 of appendix H summarises the views of the project contributors from the identified disciplines regarding the implementation of the 4 FAIR principles. The full results of the interviews and focus groups is available in section 2.

Findability

When trying to find data most contributors used one, or a combination of, the following routes:

- **Known repositories** – within research disciplines the major well known and well run databanks, search facilities and repositories are the first place most researchers go. A wide range of these services are mentioned in the full responses report in appendix G these include Figshare[100], Zenodo[101], UK Data Service[102] and Github[103]
- **Google and other search engines** – all disciplines mentioned the Google and search engines were used in searching for data
- **Published papers and literature** – the discipline journals often provide a location to find datasets associated with the published research. Also the use of DOIs was encouraged by publishers
- **Conferences/word of mouth** – most disciplines depend to some extent upon conferences and colleagues to locate relevant research data. In a few cases social media was mentioned ie Twitter feeds and Blogs

Other sources mentioned, particularly with respect to digital humanities were institutional web pages, Web of Science/Web of Knowledge[104] and ORCID[105] IDs.

Within the bio-sciences there are many data type specific repositories that are aim at particular research fields. Some of these are findable through Google but the lack of rich metadata limits this findability. This was also common in other research disciplines where the overhead of metadata creation means metadata is often, at best, patchy. The uncertainty about which metadata standard to use, poor documentation and lack of good automated tools were also cited as barriers. However, the importance of metadata to enhance findability means that metadata schemes supported by repositories are often well developed. For example groups such as the Bioschemas ELIXIR group[106] add in specific metadata to encode data such as protein entities.

In some fields the concept of findability was not a high priority. One digital humanities researcher, when asked "how can people find your data" replied "mostly they don't". Also data in some medical and social science areas was considered confidential so was deliberately not findable.

The issue of whether raw data was findable, or even if it was desirable to find, was discussed at length. Even when raw data is available it is often in a flat structure and not easily searchable. This was another example of the need for improved metadata to be available to aid findability. Other gaps in findability provision that were identified include the need to build catalogues, the requirement for closer links between researchers and those building tools, and the *"long tail"* of research data that is not discoverable.

Accessibility

In biosciences data is often made openly available using community repositories that are well supported, examples given include protein structures and sequences. A variety of web based interfaces, APIs and downloads (sometimes simple spreadsheets) are available.

However there are challenges in *"long tail"* data, which often has poor quality metadata, making access more difficult. One contributor suggested that journal editors could enforce data policies more rigorously to improve accessibility.

There are data exchange agreements between some database and repository services, operating as a network of databases with creative commons licences.

Additional tools, software and documentation are often needed and some efforts are made to make software available on Github, allowing people to manipulate the available data in different ways. Some feel the key is to use longstanding formats including plain text and .csv files. However these are error prone with poor semantics.

Embargos are seen as an issue in some fields, for example this is common in ecology. In a minority of fields there are other IP/ethical/legal restrictions. In some cases the restrictions are managed at the point of submission, asserting the right of the researcher to submit the data and the terms of access specified.

In the digital humanities data is often accessed as .txt, .csv, XML or .xls making it accessible to many other researchers. Some projects mandate the data to be made open and this can often be through the institutional repository.

The accessibility of data is closely tied to how well the data is described and the software, tools or documentation that is made available with it. In the case of images, imaging software can be very expensive or instrument specific making access to data difficult. In particular cases the specialist software used is made available with the data. For example, the Transcribe Bentham Project at UCL[107] produced software plug-ins which were added to Mediawiki[108]

Ethnography was noted to have particular ethical issues. Data is often held back until further clearance from the community being studied is obtained. Other examples included data on the ivory trade that could be sensitive and open to misuse. Data size issues were also noted in some research, where the data was too big to download and work on using the researcher's own resources, with implications for university infrastructure.

In other cases there are commercial barriers, such as ancestry.com[109], who own historical data.

In some fields confidentiality is a major issue with a number of levels of confidentiality dictating usage terms failing into three broad options:

1. Sign a simple undertaking not to disclose the microdata to any third party and not misrepresent its content

2. Hold, harmonise and distribute the data but only release on the agreement of the data originator

3. Data received and harmonised, with no permission to pass on

Interoperability

A majority of contributors found interoperability to be the most challenging of the four principles. This is illustrated in the biosciences where there is a need for better standards to describe the processed data to make it interoperable.

Also, in some cases interoperability was not well understood, for example, the use of XML was considered to mean data was fully interoperable.

Researchers in some fields of bioscience reported good levels of interoperability, and are integrating interoperability of data into their workflows. One project mentioned was a five year initiative looking at interoperability of web services, bringing data providers together as a precursor to ELIXIR.

[1]

100 https://figshare.com
101 https://zenodo.org
102 ukdataservice.ac.uk
103 https://github.com
104 https://webofknowledge.com
105 https://orcid.org
106 elixir-europe.org/platforms/interoperability/projects/bioschemas
107 http://blogs.ucl.ac.uk/transcribe-bentham
108 mediawiki.org/wiki/MediaWiki
109 ancestry.com

Some data services provide a set of tools to map between datasets. Their issues include:

- Content - is it there, is there a standard?
- Syntax - what format, can it be read, is it valid, can it be transformed?
- Semantic - have they used the same language to describe the data?

The issue of differing terminology and the need for common vocabularies was noted in several disciplines and some services are being provided to match terminology and map standards landscapes, for example: FAIRsharing[110], Gene Ontology[111] and Bioschemas[112]. The Ontology Lookup Service[113] allows searching for synonyms for terms, where a different research context may have led to different terminology. However, mapping protein structures was identified as a challenge. PDB[114] have tried to address this, and researchers reported using their mapping tools.

FAIR is seen as a benefit to help improve standard vocabularies and mapping tools but there was also a challenge identified in incompleteness of data. For example, when studying a rare genetic disease, to compare patients the way they are described must be very easily comparable.

Generally the digital humanities field reported some good levels of interoperability because the research mainly uses processed rather than raw data, with good documentation and standards available. They use standard formats including XML, for example historians may work with newspaper archives using XML format (TEI guidelines[115].

Archaeology researchers tend to use standard terms for primary data using international standards. However, it was also reported that the community was not very aware of the benefits of linked open data.

The principle of interoperability is very important to chemistry researchers. They are keen to access other researchers' data. The use of XML (CML) is quite developed in some fields and aids interoperability. However, much of the data produced from instruments, proprietary format or software specific. This means there is considerable human intervention needed to facilitate interoperability.

Social scientists mentioned that some survey data on individuals is combined with organisational level data from other sources, however, the combination of the data is currently done manually. Vocabularies were different across countries, made more challenging by competition across international agencies. Again translation between them was currently manual. Also specialist commercial software, such as NVIVO[116], may be required to process some research data.

Reusability

Reuse of data was acknowledged to be beneficial to research across all disciplines.

However, there were variations, depending upon the specific research field and data type. Data is most reusable where data types are simple and easy to describe, and when the community is organized and collaborative. Also easy to create, high quality tools are a factor.

Reuse is often considered a result of findability, accessibility and interoperability.

Metadata, availability and standard formats are all necessary to allow successful reuse even if this means converting from historical formats into standard formats.

Many of the established data services consider facilitating reuse to be an important function so provide high quality metrics, check for errors and clean up semantics. Some data from data services is highly accessed, as determined from download statistics.

Considerable reuse is reported to be primary data, and software is often needed to reuse this data. Process information is crucial for much reusability and to aid this, lab notebooks are useful, but maybe embargoed for up to ten years or not provided at all. An additional issue is versioning of data and software which is often poor. Providing adequate metadata describing the whole research process was identified as a key challenge to reusability. Sometimes this was limited by data capture and laboratory information systems.

In some fields the advance of technology now means that instead of reusing data it is being regenerated with improved, more sensitive, instruments at a fraction of the cost.

The trend towards standard licences and the use of creative commons and open source software was mentioned, but in some cases licensing was not explicit, well understood or highly visible. There was reported to be a lack of clarity around licensing, or "dodging licensing completely" in some areas. Elsewhere, such as the Internet Archaeology Journal the licence is clear. Some researchers are unaware of what the different levels of licence mean. When using creative commons the most restrictive version is often chosen by researchers, out of fear that "someone might use it for something I don't want them to". There is general feeling that non-commercial use is preferable, where researchers don't want others to make financially gain from the data, even if this is unlikely. Advice from learned societies is also, reportedly, to use the most restrictive licensing.

For some researchers data reusability was seen as "an extra thing they have to do" and was therefore a low priority with little in the way of incentives. However, some researchers produce "data papers" - short articles that describes the data, how it was collected, what it is for, what the data is, the methodology, the context of collection and a DOI to the raw data. In some fields data papers are becoming more established and seen as a different type of publication output.

Formats can be a barrier to reuse, for example data presented in a PDF rather than .csv is more difficult to reuse, particularly if high quality metadata is not available.

Some researchers have a strong emotional, ownership attachment to their research papers. The data can be thought of in the same terms and typically non final data isn't released. One interviewee felt that "99% would not make working data available. Most wouldn't even think of it."

The size of data can be a problem for reusability, for example, raw refraction data can be very, very large creating storage and bandwidth issues.

Finally, in making data more reusable it was felt that peer reviewers should need to include reusability in their review.

[1]

110 https://fairsharing.org
111 geneontology.org
112 http://bioschemas.org
113 ebi.ac.uk/ols/index
114 rcsb.org/pdb/home/home.do
115 tei-c.org/Guidelines
116 qsrinternational.com/nvivo/nvivo-products

Appendix H: Implementing FAIR - Section 2

Findability

Biological sciences findability

Biological sciences is characterised by many data specific repositories, each of which only cover a small number of data types. However, each are well known within the research field, have a good proportion of that data and some are said to work very well. Some are findable by search engines including Google although discoverability was thought to be somewhat limited by the current lack of richer metadata sets that could be shared. Literature and research papers was also noted as an important means of finding data, with links provided to datasets, sometimes live links (eg PubMed). DOIs are used in some places, in addition to other identifiers and handles. Publications typically use DOIs and some databases, but it was noted that there are cost implications for using DOIs and the scale and granularity of datasets vary so there is some debate over what should be assigned a DOI.

There are many different metadata standards, over a thousand, typically centred around specific sub disciplines and data types, (see: fairsharing.org for a full list) and are often around the agreed metadata for publication (eg Dublin Core). This can be a minimal set and not enough to describe an experimental dataset. Tooling is also said to collect minimal metadata which is hard to make consistent across all data sets. It was reported to be a challenge is to get researchers to capture the metadata in the first place and it was desirable to have tools that support them in metadata as part of the process of data collection, currently this is often a manual process. Mechanisms that exist for OA publications it was suggested could be extended to data. Many issues are not technological, but the adoption of standards - at the repository end, and then supported by tools at the data capture end.

Discoverability and accessibility of datasets is the goal of several metadata indexes, like the USA NIH funded DataMed[117] and the EBI OmicsDI[118]. Recently the NIH has launched the FAIR Data Commons programme to address the piloting of an ecosystem of resources that implement the FAIR principles[119].

Enhanced metadata schemes that repositories can support are in different stages of development. In some cases groups such as the Bioschemas ELIXIR group add in specific metadata to encode data such as protein entities. Each entity has own URI, but doesn't have independent DOIs.

The analysis process may be documented within the paper with other information provided in the supplemental file in publications. In some cases scientific workflows are deposited or code shared using eg Github.

Digital humanities findability

Findability appeared to be a somewhat vague concept in the digital humanities, with one response to the question "how can people find your data" of "mostly they don't". The Archaeology Data Service was noted as a source for specific datasets, with institutional repositories being a second port of call. One history specialist found the Zenodo and UK Data Archive services particularly useful, and Github was reported to be widely used. Whilst much data was thought to be open and accessible through APIs, in practice "no-one knows how to", with little guidance being available. Where data was in a repository it was said that only a dedicated searcher on Google would be able to come across it.

Alternative means of finding data included talking about it at conferences with similar interested people, or via an article. In some cases the citation through a paper was considered a good source. A researcher who knows about a piece of research might locate the data through the author's Orchid ID or institutional webpage. However, a general search on a topic without knowing what was available would most likely be a keyword search for the publication through for example Web of Science, Web of Knowledge or Google/Google Scholar, but the latter was thought to be not very well structured. Locating the data would follow from accessing the article and following the references to the dataset, although increasingly data in repositories can also be found on Web of Science. Data from some sources were not available online such as that from the historic environment record office.

Particular challenges existed where data was within a book, typically this meant there was no indexing or keywords, or in other words it isn't in machine processable form and difficult to retro-convert it. An example was isotopic data from an archaeological survey being a chapter in a book. Grey literature was said to be even worse, for example a developer-funded excavation report could sit on a council shelf. ADS is trying to develop a list of the grey literature and sites relating to archaeological data.

Documentation and metadata to assist discoverability was thought to be quite limited by interviewees.

In principle it was known that documentation should be done as you go along but often it was left until the last minute. Some examples of good practice include documenting on a blog as the research proceeds, and linking this to data at the end. The necessary information was stated to include a description of the process, how/where data was collected, ownership, formats, structure, abbreviations and terminology. In short, anything that anyone would need to work out what you've done. The information could form part of the submitted article or in the supplementary file.

Some researchers rely on ADS initiating a record for a published article on behalf of the researcher, but ADS would prefer the researcher to do it. ADS has an automated ingest tool, ADSeasy, for simple archives. This could include an archaeological report for a housing development if it was straightforward.

ADS uses DOIs and Zenedo and institutional repositories can mint DOIs, but in other cases DOIs tended to be used for other outputs such as articles or book chapters rather than data. Data would typically be linked to the article rather than have its own identifier. Other identifiers were from institutional repositories and Orchid IDs were also used. There has been some experimentation with issuing DOIs to subsets of data, to link direct to sections of the data. This is all manual beyond the top level of the archive and is manageable at the moment but automation may be needed in the future.

Archival metadata standards were known to exist such as UK Archival Thesaurus (UKAT) Encoded Archival Description (EAD), but were not widely used by interviewees. Researchers are more likely to think about what XML standard they are using rather than metadata. Some do fill in forms on their content management systems, where the metadata is mostly Dublin Core.

ADS also has records from Historic England and other heritage organisations, including records of excavations. It was noted that some records may not be related to a digital archive, but a record to a physical artefact.

Some use of SPARQL, the Resource Description Framework (RDF) query language, was noted, as a semantic query language for databases, for retrieving and manipulating data stored in RDF format. Examples of usage included museum collections.

Chemistry and crystallography findability

Data is generally very findable through well-known specialist databases or search engines with considerable publication data available. Much of the data is tagged or has metadata associated with it. Specific locations that were mentioned included Figshare, Zenodo, the National Chemical Database and web of knowledge as well as Google searching. One researcher mentioned Twitter as mechanism for finding data.

All the researchers were familiar with DOIs and their use in publications and repositories.

[1]

117 nature.com/articles/ng.3864
118 nature.com/nbt/journal/v35/n5/full/nbt.3790.html
119 https://commonfund.nih.gov/bd2k/commons

There was also a feeling that some data was superficially findable but not in a format that would make it useful. The issue of whether raw data was findable, or even if it was desirable to find, was discussed at length.

At times it was felt that data held in institutional repositories was difficult to find, residing on a range of systems without standard search facilities. Also useful resources like laboratory note books were often not findable.

The use of standard metadata was strongly supported and in some areas, such as crystallography coordinates, standards were being used widely.

Social sciences findability

Data that is made available on web sites is generally found through Google, although some researchers thought others would not know how to find their data even when it was well known and deposited in a national repository. In specialist research areas data is found through word of mouth at conferences and journal articles. Some raw data, based upon individual interviews, is not findable because of confidentiality issues. However, anonymised data is available in published papers.

The UK Data Service is widely used to deposit and find data. Websites from institutions and major studies direct users to the data service and actively promote the data's existence through publicity campaigns, training workshops and seminars when new data is deposited.

Data for some major studies is also provided through the IPUMS statistical service in the US, part of the Minnesota Population Center at the University of Minnesota and the Maryland Population Research Centre at the University of Maryland. There is a gradual move towards transferring as much data as can be released into an automatically subsetting and download system. The use of IPUMS is partly to reach a larger audience, but also advantages of the automatic download system that allows you to select subsets of data in a more flexible way, choosing which variables and sometimes combining them.

Through the UK Data Service a user guide, questionnaire, technical reports and a derived variable guide are reported to be provided. Their Nesstar system allows for preparing and publishing data including metadata, although some researchers were unaware of how and if metadata were added and assumed metadata and DOIs were handled by the repository. Also used is CLOSER Discovery, a metadata-based search engine for questionnaires and data from eight leading UK longitudinal studies. Users are able to search by keyword, or browse by topic and has a range of features to improve their searches and find the data they are interested in. It is reported to have more advanced functionality and searchability, and all metadata from questionnaires is entered.

Additional information is provided on how the harmonised data is produced and a lot of software examples that would enable someone who had access to the raw data to generate the variables. However, this is in a flat and not easily searchable form. Other documentation provided when the data is deposited includes a large and comprehensive description of the survey.

There is some use of persistent and unique identifiers, through assigning own identifiers and maintaining those provided by data producers, although the usage of these is not clearly understood, and some believed the UK Data Service handled this. Some reported that DOIs are assigned for journal articles but not considered for datasets.

There is some initial use of the Data Documentation Initiative (DDI) metadata standard, but in other areas there are no standards used.

Some challenges were reported in using other data, for example from administrative data holders such as the HMRC. Applications are submitted to access the data but data holders are thought to not have established procedures for handling these requests. In other cases, awareness of new datasets was through publications, conferences or individual contacts, followed up by direct contact with the data holder.

Accessibility

Biological sciences accessibility

In many cases data is made openly available using community repositories that are well supported, examples given include protein structures and sequences, where there are also well established community norms. A variety of web based interfaces, APIs and downloads (sometimes as spreadsheets) plus streaming services are available. The methods vary, to get data in different ways and formats to suit the researcher.

There are challenges in long tail data, where the capture and metadata is reportedly poor, therefore access is variable. It was suggested that journal editors could enforce data policies more rigorously, the policies themselves were seen as suitable but there was a lack of enforcement, or even awareness amongst academic editors. An example of the PLOS journals was given with a good data sharing policy[120] but is a lack of awareness of this example.

There are data exchange agreements between some database and repository services, operating as a network of databases, and creative commons licences are well established. Often additional tools, software and documentation are needed and efforts are made to put software on Github, allowing people to manipulate the data in different ways. Some feel the key is to use longstanding formats including plain text and CSV files. However these are error prone with no or poor semantics.

There are embargos in some fields, for example it is common in ecology. Ecologists were said to have come to data sharing later, and are moving quickly as a field, but starting from a lower base. In a small minority of fields there are a mixture of other IP/ethical/legal restrictions. In some cases the restrictions are said to be managed at the point of submission, asserting the right of the researcher to submit the data and the terms of access specified. There are attempts to make this process computable.

Digital humanities accessibility

Data is often accessed as either as .txt, .csv, XML or .xls so anyone can download and access it, however, it is not all adequately described. Some projects mandate the data to be made open and this can often be through the institutional repository. Open access journals are used, and the ADS where the majority of data is to download and use offline.

Where research uses existing data it is said in some cases to be “notionally accessible”, and an API is provided but no-one knows how to use it. The accessibility of data was thought to be closely tied to how well the data is described.

The need for software, tools or documentation often depends on the type of data. Sometimes this is just an application to read an excel file, or a web browser. Where data is integrated into an online journal or the ADS it is searchable online. In the case of images, imaging software can be very expensive or instrument specific making access to the raw data difficult. Normally a Tiff version is the most accessible format. In particular cases specialist software was used and made available. With the Transcribe Bentham project at UCL plug-ins were added to Mediawiki.

Restrictions include embargos in some cases, when a clear case is made for its need. In ADS data was open access but some archives have embargo periods. Some ethical/legal issues arise for example studying Twitter data, where the tweet ids can be shared but nothing else, which was said to hinder reproducibility. Where the data relates to living people, for example, relating to contemporary history or survey data, decisions are required on how appropriate it is to share, and reference to the terms under which the data was collected.

Ethnography was noted to have some ethical issues. Data can be held back until further clearance from the community being studied is obtained. Other examples included data

[1] 120 https://fairsharing.org/recommendation/PLOS

on the ivory trade that could be sensitive and open to misuse. Size issues were also noted in some research, where the data was too big to download and work on using the researcher's own computer, with implications for university infrastructure.

In some cases there are commercial barriers such as ancestry.com, who own a lot of data, or newspaper collections which require institutional subscriptions.

Chemistry and crystallography accessibility

Research in chemistry frequently requires specialist commercial software licences making data inaccessible. In some cases the data has a licence embedded within it. This is a recognised issue and there is movement towards more open source systems. The Chemistry Abstracts Service (CAS) provides a unique CAS number that acts rather like a DOI but this requires a licence.

Frequently only processed data is easily accessible with raw data not openly available.

The need to protect IP often leads to restrictions on data accessibility which include embargos of up to five years on data release. An example of an unusual embargo is in the area of climate change where ship based temperature readings could identify the locations of ships to pirates.

Within crystallography most structures are intended for, and end up being associated with, an article. The publication of the article is the trigger to release the data, which in some cases is through a notification mechanism from the publisher. In other cases researchers have to scan articles to look for the data. Data is generally kept private for a year in the CDCC unless it becomes known it is published, then researchers are asked about releasing the data after a year.

Social sciences accessibility

Data is generally thought to be open access with an end user licence for non-commercial use, although the understanding of licence types was limited. Access to some data is possible on personal application from individuals within a small specialist research field.

Some data has more secure access including administrative data, available through a safe room at the UK Data Service, or remote secure access. Some researchers report licensing the UK Data Service to act as a data processor.

In some cases there were differing levels of confidentiality, and end users followed three key patterns of usage terms:

1. Sign a simple undertaking not to disclose the microdata to any third party and not misrepresent its content
2. Hold, harmonise and distribute the data but only release on the agreement of the data originator
3. Data received and harmonised, with no permission to pass on

In some cases, only quantitative data was made openly available from surveys, with the open text responses withheld due to concerns about possible identification of individuals. This additional data was made available on request from other researchers.

Collaborations with the University of Maryland were thought to be provide greater accessibility but there is a move towards improving standards generally.

SPSS is a standard tool for analysing the data, which requires a licence, however, there is a tendency to avoid specific software development, and find solutions to display and analysis problems using standard software.

Embargos are not common, but some restrictions exist for non-academic researchers such as government and think tanks, in the UK and abroad. However, people generally want their data to be available and used.

Interoperability

Biological sciences interoperability

Interoperability was seen by some as a big challenge in biological sciences, due to not having the standards needed to describe very accurately the stage of processing that the data is in. Often not the most raw form of data is being shared and it was considered important to know what has happened up to that point. The PDB format is used for reporting and sharing experimentally determined three-dimensional structures of biological macromolecules, and it is implemented by over 50 data repositories[121]. Also a lot of people access data by the structures that the public restful APIs allow.

Researchers in some fields are reported to be good at interoperability, and are integrating data into their workflows. Some are involved in national and European initiatives. One project mentioned was a five year initiative looking at interoperability of web services, involving EBI, bringing data providers together, a precursor to ELIXIR.

Some database services provide a set of tools to map between datasets. Questions addressed include:

- Content - is it there, is there a standard?
- Syntax - what format, can it be read, is it valid, can it be transformed?
- Semantic - have they used the same language to describe the data?

An example of differing terminology is carcinoma vs cancer. Services are being provided to do such matching, but is described as the most challenging aspect, including how to make it computable.

The wide variety of data types in biological sciences was matched by the varying usage or even existence of standard vocabularies. Some areas were said to be well supported, and groups and networks existed that were moving this agenda forward through projects, initiatives, tools and services. Examples include FAIRsharing (mapping the landscape of interoperability standards), Gene Ontology (to harmonize gene nomenclature), Bioschemas, and the EMBL-EBI Ontology Lookup Service.

The Ontology Lookup Service for example allows searching for synonyms for terms, where a different research context may have led to different terminology. Originally using a clinical coding system, data may have been abstracted from the original clinical record, and this may vary depending on the country, hospital system etc. SnomedCT is a structured clinical vocabulary used in electronic health records, however, this is licensed which is said to limit its use and lead to alternative mappings.

Mappings in protein structures was identified as a challenge - moving between coordinate frames such as an amino acid sequence to a 3d structure, where the numbering for different coordinate schemes varies. PDB have tried to address that, and researchers reported using their mapping tools.

FAIR is seen as a benefit to help improve standard vocabularies and mapping tools but there was also a challenge identified in incompleteness of data. For example when studying a rare genetic disease, to compare patients you would want the way they are described to be very clean and easily comparable. The Human Phenotype Ontology (HPO)[122], was seen to be a good framework, providing a standardised vocabulary of phenotypic abnormalities encountered in human disease.

[1]

121 source: https://fairsharing.org/bsg-s000255
122 https://fairsharing.org/bsg-s000131

Digital humanities interoperability
Generally the digital humanities field is reported to be good at interoperability because the research is mainly related to processed rather than raw data, with good documentation and standards available. Standard formats in use to support interoperability include XML although researchers need to be good at using it properly for it to be validated. Historians work well with newspaper archives using XML format (using TEI guidelines) and RDF was also in use but said to be time intensive. Many formats are open and included csv and relationships diagrams. In some cases there was a need to "stitch it together yourself offline".

In archaeology there was said to be good agreement on standard terms for primary data using an international standard. Other examples include a thesaurus of historical terms mapped onto Library of Congress subject headings and the use of name authorities such as VIAF. ADS are involved in developing a series of controlled vocabularies to initiate the record from the journal.

When using existing data, factors that facilitate easy interoperability include that the data has been prepared using a recognisable standard, is properly described and easily downloadable. However, it was also reported that the community was not aware yet of linked open data, although they were aware of aspects of it such as Orchid IDs and DOIs, where they can see it in action.

Chemistry and crystallography interoperability
The principle of interoperability is very important to chemistry researchers. They are keen to access other researcher's data. The use of XML (CML) is quite developed in some fields and helps with interoperability.

However, much of the data produced is instrument, proprietary format or software specific. This means there is considerable human intervention needed to facilitate interoperability.

Social sciences interoperability
Data was reported to be produced in standard statistical database formats that are widely understood. The data held by the data team for the research group can be outputted to suit a range of different software.

Some annual survey data on individuals was combined with organisational level data from other sources. In general this added new variables rather than merging data directly, but issues were noted on consistent use of identification matching. The combination of the data is currently done manually.

Vocabularies were noted to be different across countries, made more challenging by competition across international agencies for example the United Nations Economic Commission for Europe were promoting different vocabularies to the harmonised European time use surveys, HETUS and the American time study. Translation between them was currently manual.

In one case the research project discussed was aimed mainly at increasing interoperability in their field and harmonising data by applying a data framework that maps raw data formats.

Specialist commercial software, such as NVIVO, may be required to make sense of some research data.

Reusability

Biological sciences reusability
Re-use of data in biological sciences varies, partly due to the field and data type. With genomic sequence data for example reusability is reported to be very good, with data often reanalysed. In one example data produced 25 years ago is being reused. The relatively simple data types generated are generally well organised. This appeared to be a theme, where data types were simple and easy to describe, and when the community is well organised and collaborative, it was easy to create tools and the tools were high quality.

Re-use was seen to follow findability, accessibility and interoperability. Metadata, availability and standard formats to name a few were all necessary to allow reuse. In some cases this involved converting data from old formats or into standard formats to allow reusability. For database services, reusability was seen to be the main purpose, although additional work was needed before the data could be used for analysis, such as running quality metrics, checking for errors and cleaning up semantics. Some data from data services is highly accessed, as determined from download statistics. Re-use was said to be more likely where datasets were published and all primary data shared, otherwise it was difficult to do quality metrics and re-analyse. A lot of reuse was reported to be primary data, and software is often needed to read that data. It is typically generated from equipment, with processing as part of data generation. Knowledge of what has happened during that process is also important for reusability. Lab notebooks are useful but often embargoed for up to ten years.

One challenge identified was versioning of data and software, with attempts from data services to provide backwards compatibility. Changes to database schemas was noted to affect software and thus compatibility. In general however, researchers were downloading data alone to analyse themselves and not using the software services.

It was reported that data usage and scale has expanded rapidly, and the data landscape changing all the time. Where data used to be single structures, now it is assemblies of structures - for example, how the proteins are interacting in the cell. This is much more useful data but there were implications for the maintenance of infrastructure to keep up with this increase. There was a lot of effort put into consistent systems in generating data that is robust and with integrity. There was also a challenge in balancing the need to provide consistent data between releases, and being flexible enough to go and do new research to provide new datasets.

The useful lifetime of data was said to depend on many factors including how hard was it to get, how long did it take, how robustly was it generated and how costly was the generation? A paradigm shift in technology has been observed, with huge changes in the ability to generate and analyse data. In some cases it made more sense to re-generate data, resulting in better data, for example, more sensitive detection from improved instruments.

There was a general trend towards standard licences and the use of creative commons and open source software was mentioned, but in some cases licensing was not explicit or highly visible, for example, being included in general terms and conditions. Some very specific licences were attached to added value data.

Providing adequate metadata describing the whole research process was identified as a key challenge to reusability. Sometimes this was limited by data capture and laboratory information systems for experimental groups. Some metadata is known to be not captured for example cells are not in the same condition, so the data can't be compared with other cells.

Core facilities were said to be key to improving this situation, with some advocating a central facility to train people in standard technologies, and cost it in their grants. Another key issue was the lack of a complete dataset being made available, for example, negative data that doesn't support the hypothesis may not be included in the dataset.

Digital humanities reusability

The extent of reuse in digital humanities was reported to be difficult to know as it wasn't tracked, although downloads usually were. It typically only comes to light when new data or research refers to it. However, with more people using Zenodo and Github as well as institutional repositories this may change.

In fields such as history when curating a corpus from larger data set, it is thought to be of enormous interest to others. In other areas making data reusable was seen as "an extra thing they have to do" - to write up a data paper in journal.

A data paper is described as a short article that describes the data, how it was collected, what it is for, what the data is, the methodology, the context of collection and a DOI to the raw data. Data papers are reported to be a recent introduction, and was thought to have originated in the physical sciences.

They are seen to be a different kind of publication output, giving people with a well-structured archive academic credit for it. In five or ten years' time it was envisaged that data papers wouldn't be produced separately, but credit would be given to properly submitted data. It was said to involve a huge amount of work to submit an archive and describe it properly and this was a barrier to submission. A lot of things get stalled at the description needed to describe the data they have collected, which was described as quite onerous and repetitive, causing delays. In ADS an introduction and overview section is completed, requiring real life language of what the data is.

Examples were described where some data was left out of published data sets, because the data was considered irrelevant to the research. However, in terms of reuse this data may be of use in the future for other research but has no-where to be stored or published.

When reusing other data, it was said to be important to be able to trust the data you are using. One researcher stated "If you don't know how it is being prepared, what methods, decisions etc. then the temptation is to start again yourself."

Formats can be a barrier to reuse, for example data presented in PDF rather than CSV. Other difficulties were described in comparing with data from earlier periods that had no metadata, requiring conversations with the original researcher or not including the older data.

In principle much humanities data can be thought to be usable infinitely as it is historical although some interviewees reported challenges with openness. Where it is released it tends to be at the end of the project, but this might be withheld if there is an expectation of being "not finished with it".

Few researchers release data before publication, mostly because there is no mechanism to share before publishing. In specific fields there is more protectiveness such as archaeogenetics where data is said to be only released once in press.

There was reported to be a lack of clarity around licensing, or "dodging licensing completely" in some areas. Elsewhere, such as the Internet Archaeology Journal the licence is clear. In ADS people are said to be more cautious with datasets, and are not sure it is ok to use them, even when stated as CC-BY. Some people are unaware of what the different levels of licence mean.

Creative commons is typically used, but the most restrictive version is usually chosen by researchers, out of fear that "someone might use it for something I don't want them to". There is general feeling that non-commercial use is preferable, where researchers don't want others to make money out of it, even if this is unlikely. Advice from learned societies is also reportedly to use the most restrictive licensing.

For some researchers it is not about data but the end is the article, the text and the writing, with which there is a strong emotional attachment. Data is talked about in the same terms and typically non final data isn't released. One interviewee felt that "99% would not make working data available. Most wouldn't even think of it."

Chemistry and crystallography reusability
It is clear that considerable reuse of data is taking place but this tends to be the secondary or processed rather than raw data. In some cases the data may be quite old, over 20 years. Many funders ask that data is available for at least ten years but many institutional repositories keep data for much longer than this. Some researchers use the number of citations their data and software receives to demonstrate reusability.

The size of data can be a problem for reusability, for example the raw refraction data can be very large creating storage and bandwidth issues. Sometimes a triage arrangement is needed to decide which data to archive in a repository. Standard formats are also crucially important in raising reusability levels. Software, often developed by the researcher, it was thought should be available with the data whether through a paper or repository.

There was a feeling that the availability of laboratory note books would be very beneficial but they are often withheld. Similarly the use of electronic lab notebooks is not wide. One researcher mentioned that he would be attending, PIDapalooza, the festival of Persistent IDs.

In making data more reusable it was felt that peer reviewers should need to include reusability in their review.

Social sciences reusability
Re-use was considered core to the research in the areas studied, such that the data was designed for others to use.

Typically data was released within three-six months with an aim to keep it indefinitely. As many of the studies were trying to understand social change, it was a case of "the longer the better" for data availability. In some cases data was being retrieved from old physical sources and typing it up to add to the dataset.

With longitudinal studies, data was often released after collection of each stage of a three-five year sweep of data collection. As an example interviews might be collected over three-nine months, with three months receiving data, six months preparing it before release, including some data cleaning and preparing for deposit.

Documentation, including diaries, is particularly important for some research to provide context, however, this may be in foreign languages. The use of services such as Google Translate is increasing as the quality of these services improves.
Different licences were seen to exist, particularly across national boundaries and varying over time, such that a licence agreed may be overturned several years later when new staff took charge of an agency.

The data was thought to be comprehensive, in providing all necessary documentation needed to re-use the data, including questionnaires and user guides, technical reports and scale variables. There were some restrictions on standardised questionnaires that were licensed from the questionnaire company. Permission was granted for the making the individual questions available, but not the syntax for creating the scale from the combined questions. This meant that other researchers would not know how to produce the scale main score from the individual questions.

It was suggested reuse would benefit from search platforms across different studies or multiple sweeps of a single study to improve findability.

Appendix I: Interviews with funders and publishers

The main focus of this report has been practice within the research community, however, various references to publishers and funders have been made throughout interviews and other conversations, so it is fitting that a brief coverage of their perspective is included.

For the purposes of this study, representatives from the Economic and Social Research Council[123] (ESRC) and the Wellcome Trust[124] were interviewed to explore the funder role, and Elsevier[125] and Springer Nature[126] on behalf of publishers. However, the views reported should not be taken to represent all of the funding bodies or publishers. In particular it was noted by other contributors to this study that small publishers are involved in innovative practice and that some of the bigger publishers are supporting F and A from FAIR but not I and R. Details and comparisons of publishers' data policies are covered elsewhere[127].

Economic and Social Research Council (ESRC): Research funder

ESRC view data as the most important asset in research and focus heavily on the production of data that can be used and reused, alongside high quality research. In their research fields, data is about people and includes sensitive data. The data includes personal data about individuals, households, families, and small areas, covering all aspects of people's life, health, education, life events etc.

ESRC have a data infrastructures team and fund major longitudinal studies in addition to an administrative research network, and the UK Data Service[128]. They describe their role as including providing guidance to researchers and improving the policy environment. In terms of FAIR they see their role as working with sister councils, to align themselves with best practice. This includes to update guidelines to researchers and to make sure they have the services in place for researchers to be able to deposit data.

They publish a research data policy that has been recently updated and includes explicit reference to the FAIR data principles. The policy covers the responsibilities and what is expected from researchers, data services, providers and the funding council. The RCUK[129] position statement is a key influence and the council benchmark against that.

The policy contains nine principles within which are implementation notes on how to follow these principles for example in the production of metadata and use of standards.

The grant holder is responsible for making their data available according to FAIR. This not monitored by ESRC but it is a prerequisite of funding to produce a plan. It then becomes the responsibility of the grant holder and their university to ensure the principles are adhered to.

FAIR was seen as a set of high level principles, of best practice, for data to be more accessible, be able to be shared, and to ensure the conditions are in place for this to happen. This means a safe place to store, to publish findings, to cite where data is, opening data, opening outcomes of research and to change the culture, with people taking more responsibility. It is thought that FAIR could bring together things that have already been done and flesh out top level principles. It was felt that there would be no sudden change, it was more about encapsulating what was best practice.

With data infrastructures and longitudinal studies there is a closer investment management process, although again it is only at the point of commissioning, with no mechanism to ensure they comply. This partly a practical issue of

scale and resources, but also of trust, with the approach taken of creating best practice and influencing culture rather than forcing compliance. Funding longitudinal studies and data infrastructures was seen as key, and a lot of money is about preparing data for reuse, providing staff and technical infrastructures.

FAIR references are also included in their research guide, researchers use the funding guide when they want to apply for funding. On application, researchers are required to produce a data management plan to explain what are their data sources, analysis of gaps of what they are using and what is required, information on the data that will be produced and accessed eg the type, quality, were collected, and consent. If third parties are used the DMP should refer to quality assurance. It was noted that there is a big dilemma in balancing openness with care about confidentiality and privacy.

Researchers are expected to put their data in the repository after three months. More recently they have added flexibility to allow repositories other than the UK Data Service, as long as the principles are met. This change is in part to allow for non-academic users and data owners such as government and the private sector.

ESRC explicitly do not own the data in most cases, it is owned by the research organisation, the legal entity that researchers are members of. For some specific longitudinal studies such as the understanding society study, ESRC are the data owners, and this position is reported to be similar to the Medical Research Council[130] (MRC) on very specific studies. In terms of challenges for FAIR, implementation of the principles was not straightforward, in particular with interoperability. The aspiration was to create a place where researchers can go and find everything, that points you to the data you want. However, it was recognised that the more you try to bring together different data infrastructure and different data, the more difficult it became to align standards, access, metadata etc. and different cultures and infrastructure. Other challenges notes were public perceptions and attitudes on how they want their data to be used and skills for researchers in knowing how to work with data, to train future generations.

Wellcome: Charitable foundation

As a major funder of health research, mainly biomedical and social sciences, Wellcome provides researchers with a range of policies and guides relating to their management of data. These guides are, in part, aimed at ensuring data can be shared and reused by other researchers as well as maximising the value of the data. For example. a recent guide encourages the identification and use of recognised repositories. Wellcome is an advocate for openness and sharing of research data to maximise its impact in benefiting health, while recognising that attention must be paid to IP and ethical issues.

Although the use of standard formats is not specified they are considered important.

Wellcome does not generally own, or manage, the data produced by its funding programmes but does stipulate that it should be open, sharable and reusable.

Most of the guidance provided is at a high level with operational and specific practice left to the research groups and their institutions.

[1]

123 esrc.ac.uk
124 https://wellcome.ac.uk
125 elsevier.com
126 springernature.com/gb
127 dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
128 ukdataservice.ac.uk
129 rcuk.ac.uk
130 mrc.ac.uk

Similarly the ownership of research data from Wellcome funding resides with the institutions. However, they are aware of the emerging concept of the individual data contributor having ownership rights.

When Wellcome requires advice or support relating to data management issues they tend to turn to the DCC and Jisc, often through individual relationships. Wellcome also contributes and provides input to a range of national and international data management groups which can, in turn, provide help and guidance.

The FAIR Principles have been on Wellcome's radar for over a year and they consider the FAIR "brand" to be very useful. While it is not yet specified in policies, Wellcome aims to follow the principles and is particularly interested in the cultural barriers. Other related issues mentioned include:

- The need for incentives for researchers
- The lack of available repositories making data harder to find
- Evidence based research and whether FAIR has a role to play
- The need for better uptake of DOIs
- The use of proprietary software
- Sharing personal and sensitive data

Wellcome generally specifies that data produced as a result of its funding should be available for at least ten years. It also provides high level guidance on documentation but currently provides little in terms of licensing advice.

It is likely that the FAIR Principles will become part of future Wellcome policies.

Elsevier: Research publisher

The Research Data Management Solutions team at Elsevier are *"committed to making it easier to share research data"*. As part of this commitment they are working with authors and institutions, in particular they are working with the University of Manchester to build a repository based upon their Mendeley platform.

They work with *"a broad range of data including raw and processed data, and covering all digital formats"*. However, as a publisher Elsevier does not own the data, the ownership remains with the researcher or institution.

Elsevier work with a range of international data management groups, including the Force 11 group, they have *"been involved from the beginning in the development of the FAIR Principles and were co-authors on the first FAIR paper"* [ref]. They feel this illustrates their commitment to openness, sharing and reusability. This commitment has extended to collaborating with competitor publishers.

Individual journals from their portfolio choose from a range of Elsevier data management policies. Selecting the ones which are most appropriate to their field of research. These do not currently refer to FAIR, however, Elsevier are piloting the DANS FAIR validation tool.

The main challenges to FAIR are seen as:

- The lack of incentives or credit available to researchers to adhere to FAIR
- Variable interpretations of FAIR eg specifying what machine readable means
- Improved use of metadata and DOIs
- The need for standard file formats
- Clarity on embargos and restrictions

- Poor levels of interoperability
- Increased awareness of licensing arrangements, particularly open licenses

Springer Nature: Research publisher

Springer Nature are involved in several roles relating to data, including developing policies, supporting researchers and some curation of data. They are involved in outreach, participation in, and collaboration with, external groups and other agencies such as the Research Data Alliance[131], CODATA[132], FORCE11[133] and ELIXIR[134]. They provide a helpdesk service for researchers, dealing with questions such as finding repositories, complying with journal policies and submitting data files.

The types of data are highly variable, reflecting the variety of disciplines covered by their journals, from humanities to high energy physics. This includes data that supports the results in a publication but could be raw or processed data, images, text, code, spreadsheets etc.

The type of data researchers make available is seen to reflect the skills, culture and infrastructure available in their discipline, for example some have the tools and services to share raw data, others do not. In many cases it is felt that *"researchers are only just getting used to have a data sharing policy"*, so some of the *"implementation details within FAIR are not yet within the reach of researchers"*.

Formats vary between disciplines, but the use of repositories and established data standards is encouraged, using FAIRsharing.org[135] as a source of standards. Standards are *"enforced only in disciplines where there is a mandate and the infrastructure and culture exists to support them"*, including discipline specific repositories. Some journals take a stronger stance, requiring a data access statement and specific formatting standards, often where this is matched by a requirement by the funding council. This is said to need buy in from editors and peer reviewers and the publisher's role is to work in collaboration with editors, and through appropriate editorial policies. This can include helping researchers find discipline specific repositories, of which a list is maintained of about 80-90 that meet published criteria. Otherwise, support is provided to use more general solutions such as institutional repositories. A pragmatic approach is taken, recognising that some journals and editors are only just starting to think about data policies. A series of four levels of policy are available, starting at the very simple, covering how to find repositories and cite data. These move up to including more rigour with discipline-specific mandates and *"discussion with editors on the process for implementation so they have the guidance and resources they need"*, but also recognising this is a shared responsibility with the researchers.

Some broad discipline-specific differences are recognised, supported by their own analysis of usage, such as life sciences data sharing being generally more established than social sciences. However, this is not necessary true in sub disciplines. For example, in economics, data replication data must be available with the final publication, but it is uncommon for data sets to be part of peer review. Equally, in neuroscience, a lot of data is only available on request. It was felt that humanities generally had the least uptake of data sharing policies so far.

Researchers were seen to be *"confused"* over data ownership and licensing issues, for example not knowing if they *"have the right to apply creative commons licences"* to their data. The Springer Nature helpdesk advises researchers for example on the implications of certain licences. As a publisher they do not take ownership of data.

[1]

131 rd-alliance.org
132 codata.org
133 force11.org
134 elixir-europe.org
135 https://fairsharing.org

FAIR is not used explicitly in current policies and guidance, and this is seen to reflect current practice of researchers, who are generally not asking about it. They believe they are *"leading the way and being innovative in terms of data sharing more generally"*, including standardised data policies and publishing "Scientific Data" as a high quality data journal with data peer review. They are *"keen to share and learn with the community and have created an interest group with the Research Data Alliance"*.

FAIR is seen as a useful concept of guiding principles, giving direction and recognising that open access isn't always achievable. It is a *"useful acronym to help explain the importance of data sharing, but isn't necessarily practical"*. There are concerns that it is being overused without consideration of what it takes to achieve FAIR data across disciplines. It is thought that *"when the usage of FAIR begins to help at a practical implementation level, it will become more important to integrate the terminology into policies and guidance"*.

The main challenges to adoption of FAIR are seen to be awareness and usage of appropriate infrastructure, and awareness and investment in data curation. It is recognised that data curation skills may exist in different places - *"the researchers themselves, institutional services, independent services such as the DCC, and also publisher services"*. Curation is seen as a necessary and important part of FAIR implementation as is providing access to funding, skills and infrastructure for researchers.

Bibliography

The British Library (2017). British Library Data Strategy 2017 25 August 2017 [Online].
Available at
http://blogs.bl.uk/files/britishlibrarydatastrategyoutline.pdf
(Accessed 1 September 2017).

Data Archiving and Networked Services (DANS) (2016) FAIR data sharing in Trusted Data Repositories.Powerpoint presentation 29 September 2016 [Online].
Available at https://f.hypotheses.org/wp-content/blogs.dir/78/files/2016/11/P_Doorn.pdf
(Accessed 1 August 2017).

ELIXIR (2017). ELIXIR position paper on FAIR Data Management in the life sciences. 7 September 2017.
Available at elixir-europe.org/system/files/elixir_statement_on_fair_data_management.pdf
(Accessed 10 September 2017).

The European Commission (2016). Guidelines on FAIR Data Management in Horizon 2020.
Version 3.0 26 July 2016 [Online].
Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
(Accessed 1 August 2017).

Open Research Data Taskforce (2017). Research Data Infrastructures in the UK. Universities UK. June 2017.
Available at universitiesuk.ac.uk/policy-and-analysis/research-policy/open-science/Documents/ORDTF%20report%20nr%201%20final%2030%2006%202017.pdf
(Accessed 1 August 2017).

Research Councils UK (2015). RCUK Common Principles on Data Policy. Revised July 2015 [Online].
Available at rcuk.ac.uk/research/datapolicy
(Accessed 1 September 2017).

The Royal Society (2012). Science as an open enterprise.
The Royal Society Science Policy Centre report 02/12 (2012) [Online].
Available at https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf
(Accessed 1 August 2017).

Sipos, Gergely (2017). What is FAIR? Inspired 28. The EGI Foundation. September 2017 [Online].
Available at egi.eu/about/newsletters/what-is-fair
(Accessed 1 September 2017).

Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci.
Doi: https://Doi.org/10.1038/sdata.2016.18

Wilkinson M.D. et al. (2017) Interoperability and FAIRness through a novel combination of Web technologies.
PeerJ Computer Science 3:e110
https://doi.org/10.7717/peerj-cs.110

Wilkinson, M. D et al (2017a) A design framework and exemplar metrics for FAIRness.
27 November , 2017. bioRxiv 225490;
Doi: https://doi.org/10.1101/225490
Available at
biorxiv.org/content/early/2017/11/27/225490?rss=1
(Accessed 28 November 2017).

In addition, online information was used from the following groups:

Dutch TechCentre for Life sciences - FAIR Data and the GOFAIR initiative
dtls.nl/fair-data

The FAIR metrics group
http://fairmetrics.org

Share our vision to make the UK the most digitally advanced education and research nation in the world

jisc.ac.uk

Jisc
One Castlepark
Tower Hill
Bristol, BS2 0JA
0203 697 5800
info@jisc.ac.uk