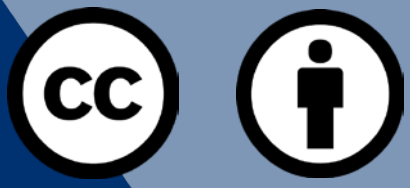
October 2016

Building and sustaining data infrastructures: putting policy into practice

Juan Bicarregui



Cite this as: Bicarregui, Juan (2016) Building and sustaining data infrastructures: putting policy into practice. https://dx.doi.org/10.6084/m9.figshare.4055538

Building and sustaining data infrastructures: putting policy into practice

A report for the Wellcome Trust
Juan Bicarregui
August 2016

Preamble

This report reviews and analyses the data infrastructure provision available to Wellcome Trust funded researchers and suggests some areas where intervention would be helpful.

As Wellcome mostly funds researchers in UK, the report focuses on the context in the UK and international global actions relevant to UK. This report is primarily relevant for life science and health but also compares with other fields.

This report is one of five being commissioned by Wellcome. The other four cover the following topics and these topics are therefore not covered in this report but some comment is necessary to put in context.
The other reports are:

A. *Embedding cultures and incentives to support open research*
C. *Developing skills for managing research data and software*
D. *Establishing data standards, metadata and interoperability*
E. *Ensuring global equity in open research*

Abstract

This report describes the personal view of the author on the topic of building and sustaining data infrastructures. In Section 1, it summarises of some relevant policies and initiatives and in Section 2, it reviews some issues around current provision of data infrastructure focusing on how data infrastructure can make data Findable, Accessible, Interoperable and Reusable. It argues for the separation of the functions of dissemination, verification and value judgement of science that have traditionally all be encompassed in the publication of articles. Section 3 discusses how the type of data affects the features required of the data infrastructure and goes on to suggest some areas where improvements in data infrastructure could yield benefits. This section also discusses some issues around the sustainability of data infrastructres. Finally, Section 4 makes some recommendations for new features that could be provided by the data infrastructre concetrating on how supports could be provided for the assurance of provenance of the data created. It ends with an outline of a possible programme of work to begin to implement this new data infrastructure.

Table of contents

1. Review of Policy and Initiatives

1.1 The policy landscape

The last twenty years[1] have seen a steady move towards openness in scientific research[2] that has in the last 5 years accelerated with some significant changes in the global policy environment in which research is conducted. Adapting research practice to fully implement these new policy requirements will require a new infrastructure to support it. This section summarises some of the most significant policy statement and initiatives that affect UK researchers.

In 2011, RCUK published some joint principles on data that were closely followed by renewed policies from each Research Council. In June 2013, the G8 London Statement[3] reaffirmed that *"scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards"* and in the US the Office of Science and Technology Policy (OSTP) required of each agency that "*data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publically accessible to search, retrieve and analyse*"[4]. The OSTP memo led to new policies from NSF[5], NIH[6] and other agencies.

In 2015, Research Councils issued guidance on data principles providing explanatory text for each of the 7 Common Principles from 2011, and G7 issued a further statement[7] calling for *"convergence and alignment of inter-operable data management that could accomplish an effective open-data science environment at the G7 level and beyond.*" Most recently, in April 2016 the European Commission issued the Communication on European Cloud Initiatives,[8] which proposes a series of measures to enable Europe to benefit from open science, and in July 2016, the UK Concordat on Research Data[9] reinforced the Principles and Guidelines with agreement across the sector.

The Wellcome Trust has for many years been a leader in this move towards open science, for example, the Trust policy on Data Management Plans was highly influential during the formulation of RCUK data principles. It is now timely for the Trust to develop the data infrastructure to fully support its innovations in open science policy.

1.2 RCUK: Principles, Guidance and Concordat

The RCUK Common Principles on Data provide a concise summary of some tensions inherent in the process of making research data open. Similar points are made in the G8 statement and OSTP memo. The 7 principles can be summarised as follows:

1) Data should be made openly available
2) Data should be managed
3) Data should be discoverable
4) There may be constraints
5) Originators may have first use
6) Reusers have responsibilities
7) Data sharing is not free

These are depicted graphically in the diagram (right) that shows how there are three dimensions of tension inherent in the principles.

The first dimension, shown on the vertical axis, is between the first two principles: although data is a public good, in that it should be available to all and is unlimited in how often it can be used, it needs to be managed and some organisation has to take responsibility for the management.

The second tension is shown in the dimension marked Access: Principle 3 proposes access and discoverability through metadata, but Principle 4 brings attention to possible constraints such as legal, ethical or security. For any particular data set, a judgement must therefore be made that balances the benefits of access against the possible reasons for constraints.

The third dimension of tension is the most relevant here as provides one of the motivations for the recommendations presented later. It concerns the recognition of intellectual contribution. Principle 5, that data originators should have first use in order to publish papers, is driven by the fact that the production of papers is currently valued more highly than the production of data. However, ideally it should be beneficial for data creators to see their data used as widely and as quickly as possible, and therefore not to make use of Principle 5's embargo periods, which delay the realisation of any added benefit that could be gained through reuse of the data. The need for embargo periods demonstrates that the proposal of Principle 6, for reusers to recognise the sources of their data, is not considered sufficient incentive for originators to make their data open. A richer and more accurate system of giving credit and value to the different roles played in research is therefore required if data is to be made available for reuse as quickly as possible. This issue is discussed further in Section 3.2.3.

1.3 The European Open Science Cloud (EOSC)

At the European level, the need to provide an infrastructure to meet the above policy requirements has been recognised recently in the EC Communication on European Cloud Initiatives[10]. The Communication sets out five reasons why Europe is not yet fully tapping into the full potential of data. These are:

a) *Data for publicly funded research is not always open and there is a lack of clear structures that incentivise and reward data sharing.*
b) *A lack of interoperability required for data sharing where data is large and complex, in varied formats and requiring complex software, noting deep-rooted walls between disciplines.*
c) *Fragmentation between data infrastructures that are split by scientific and economic domains, countries and governance models and have different access policies.*
d) *A surging demand for HPC at a scale where no single member state has the financial resources to develop the necessary HPC Ecosystem in a competitive time frame.*
e) *The ability to reuse data employing advance analysis techniques in a dependable environment that ensures adequate protection of personal data considering the forthcoming revision of Copyright legislation.*

To address these barriers, the Communication proposes that Europe develops a European Open Science Cloud (EOSC) that provides a trusted, open environment for the research community for storing, sharing and re-using scientific data and results, and a European Data Infrastructure (EDI) that provides an underpinning computing infrastructure comprising super-computing capacity, fast connectivity and high-capacity data management. The EOSC will offer a virtual environment with free at the point of use, open, and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines.

The Communication sets out that to develop the EOSC, it will be necessary to:

a) *Make all scientific data produced by the Horizon 2020 programme open by default.*
b) *Raise awareness and change incentive structures for academics, industry and public services to share their data.*

c) *Develop a specification for interoperability and data sharing across disciplines and infrastructures*
d) *Create a fit-for-purpose pan-European governance structure to federate scientific data infrastructures and overcome fragmentation.*
e) *Develop cloud based services for Open Science, supported by the necessary data infrastructure*
f) *Enlarge the scientific user base to researchers and innovators from all disciplines.*

The EDI will underpin the EOSC with data infrastructures which store and manage data, high-bandwidth networks which transport data, and ever more powerful computers to process data.

The Communication is supported by a report from a High Level Expert Group on the EOSC[11] and aligns with the Riding the Wave[12] and The Data Harvest[13] reports from previous High Level Expert Groups.

The current phase of the Horizon 2020 workprogramme contains several action lines that will contribute to the building of the EOSC and EDI. In particular INFRADEV-4-2016, The European Open Science Cloud for Research, will bring together a broad range of stakeholders to address the Fragmentation, Interoperability and Governance issues identified in the Communication and is highly relevant to the subject of this report.

1.4 OSTP memo, BD2K and NIH Commons

The "Holdren memo" from the Office of Science and Technology Policy in 2013 required US federal agencies to demonstrate how they will make their data open. As a result, the NSF has published open data policies that require Investigators "to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants" and the NIH is also making their policies consistent in this regard. For example the NIH Genomic Data Sharing Policy states that "investigators should submit the data to the relevant NIH-designated data repository (e.g., dbGaP, GEO, SRA, the Cancer Genomics Hub)".

The NIH Big Data to Knowledge initiative (BD2K), is a substantial research programme that "seeks to better define how to extract value from the data, both for the individual investigator and the overall research community, create the analytic tools needed to enhance utility of the data, provide the next generation of trained personnel, and develop data science concepts and tools that can be made available to all stakeholders."[14] It has funded 11 centres of excellence in data science, the development of a biomedical data discovery index that will enable discovery, access and citation of biomedical research data sets, and grants to enhance the training of methodologists and practitioners in data science. It is expected to have a total investment of nearly $656 million between 2014 and 2020[15].

In the NIH Commons project[16], the NIH are currently in the process of making some high impact data sets available under FAIR principles[17] in a cloud environment. This project goes further than simply releasing data through data repositories; making the data available in a cloud compute environment along with metadata will facilitate data analysis and merging with other data. The process of deciding which data will be the first to be released in this way is underway at the time of writing this report, and further datasets will be added later.

The initiative is motivated by ease of use, but should also save costs through economies of scale wherein access to cloud resources are negotiated at aggregate level. This removes the need for each grant to request its own compute resources, either private or cloud based. Access to these compute resources will then be provided through credits for access awarded along with grants (although formally awards will not be grants themselves). The cloud resources will also allow collection of metrics of use of the data, which will give an indication of which data are being used for which purposes, and give pointers of which data to keep and which to discard.

Provision of data in this way should also make it easier to cross link data and to build interfaces that provide homogenous access to heterogeneous data through APIs, and should, in the long term, lead to more standardisation and improved strategies for curation and persistence.

The NIH Commons project provides an example of how data can be made available for reuse in an effective and efficient manner through a centrally provided data infrastructure that furthers the goals of open science.

2. Current Infrastructure provision

2.1 Supporting different aspects of the communication of research

Currently, the primary route for communication of research is still through the publication of articles. In its traditional form, the publication of a paper embodies three different aspects of dissemination: firstly, the paper provides a vehicle for dissemination of the results; secondly, it describes the process undertaken and thereby attests to the correctness of the data upon which the results are based; thirdly, it presents those results for assessment of their scientific value.

Historically, responsibility for all three of these functions has fallen largely within the remit of the Journal. However, the move to electronic communication, originally motivated by convenience, has also provided an opportunity to separate responsibility for the different aspects of research communication from each other, and to build more specific infrastructure to support each of them.

Considering the first function of communication, the **dissemination** of research outputs, this is a role for repositories: repositories of papers, repositories of data and repositories of software. As well as providing access to the individual items they contain, repositories can also add significant value to their content by bringing together collections of similar items so that they can be more easily found, and so that analysis can be carried out more easily across them. These features correspond to the *Findable* and *Accessible* of the FAIR principles and are discussed further in Section 2.2 below.

The second function, the recording of the activities undertaken in order to provide reproducibility and demonstrate the **provenance** of the data, begins with the research processes and the environment that supports it. This environment is normally provided by the researcher's institution, but can also be provided externally, for example at a large-scale shared facility such as the Diamond Synchrotron. For research with a well-tried process, the recording of the activities undertaken can be built into the processes inherent in use of the environment, and the provenance of the data can therefore potentially be verified as part of its creation. In other forms of research, where the process is more bespoke to a particular study, if provenance information is to be captured at all, it might have to be recorded manually as the research progresses, and possibly verified as part of the ingest process into a repository. In all cases, demonstrating provenance requires a richness of metadata so that the data can be understood by others with a level of verifiability[18] that the research was indeed

conducted as described. These features correspond to the *Interoperable* and *Reusable* of the FAIR principles and are discussed further in Section 2.3 below.

The final function, the assessment of the **scientific value**, is quite distinct from the others. It is not about the recording and disseminating of the work done, but an inevitably subjective judgement made by peers. This is different in nature to the other two aspects, which can be carried out without any value judgement.

It is worth noting that new forms of publishing, such as that provided by Wellcome Open Research[19] through F1000Research[20] provide a significant step towards this separation of concerns. By separating the dissemination and review functions it enables much more rapid dissemination, and by opening up the review process it engenders a more collaborative approach to reviewing and community comment.

The platform also supports the deposition of data to support the finding. However, to fully support the provenance function as described requires features of the data infrastructure that cannot be provided by a publication platform.

2.2 Making Data Findable and Accessible

Where data is referred to in a paper, Findability is relatively straightforward. Whether through a general-purpose persistent id or through an accession number to a particular database, many systems exist which can turn a reference to data in a paper into access. However, without such an explicit reference, data can be difficult to find without prior knowledge of its location. In comparison, when searching for papers, journals and collections such as PMC and ArXiv provide the content for general-purpose search engines to deliver easy cross-disciplinary searching.

To some extent, data repositories, with their metadata catalogues, provide a solution. Metadata-based searching can be effective, provided that the metadata is rich enough and the query does not depend on the values in the actual data. Datacite[21], although primarily intended as a service for resolving persistent ids, is also a form of metadata repository as the landing page for each dataset provides a limited set of metadata that can be searched. However, more refined searching through metadata requires richer metadata that can be onerous to create if not produced automatically, by the instrument producing the data for example.

It would be preferable to be able to directly search the actual data itself in the way a search engine searches the content of web pages, as this would remove the need for extensive metadata, as well as enabling the user to search by the actual value of a data field. It would blur the distinction between what is data and what is metadata. However, for such a mechanism to work across disciplines, the required level of standardisation in the representation of data is still some way off. A first step along the way to this is the registration and collection of metadata schema in machine interpretable form to enable the building of tools that work across metadata schema. The Open Metadata Registry[22] based on W3C's Simple Knowledge Organization System (SKOS)[23] and the Metadata Standards Directory[24] from the RDA are steps towards this goal.

The result of a search is a pointer to the data, often in the form of a Persistent Id (PId). However, PIds do not always point to the data itself, but often to a textual description of it. This description then includes another pointer to the data itself. There are currently many forms of PIds in use, which is not problematic as the important feature is that a given PId continues to point to a particular piece of data (see also Section 3.1.5). Note that a PId need only be unique in one direction: whilst a given

PId should always resolve to the same data, it is not problematic if a given object has many PIds that resolve to it. Sometimes use of the phrase Unique Id confuses this.

For software, the situation is more ad hoc. The ability to search for software that provides a specific function by semantic metadata is a long way off. This ability should, however, be understood as the ultimate goal. In the meantime, researchers have to manage with textual descriptions of the software's functionality.

2.3 Making Data Interoperable and Reusable

Interoperability[25] has at least two distinct facets: at one level it can mean interoperability of the data infrastructure technology, and at another level it can mean interoperability of the data itself. We deal with each of these in turn.

2.3.1 Technological Interoperability

The ESOC Communication identifies fragmentation of infrastructure provision as a barrier to maximising the use of data. In the US, similar issues are being addressed by the BD2K initiative. Technological fragmentation arises for two reasons: technically different infrastructures lead to low interoperability, and separate governance and funding arrangements lead to heterogeneity of provision.

Currently, data infrastructure is provided by a mixture of horizontal and vertical services. The advantage of vertical service provision is that it can be dedicated to the needs of particular research fields. However, it can also lead to a lack of interoperability with other vertical infrastructures. Horizontal services, on the other hand, are more likely to lead to cross-disciplinary homogeneity, however it is more difficult for horizontal services to be tailored to particular researchers' needs.

Domain specific vertical infrastructures are currently serving the needs of their particular communities well. However, unless there are strong incentives to provide interoperability between infrastructures, provision will remain fragmented and opportunities for cross-disciplinary research and innovation will be lost. Furthermore, it seems that current users of the infrastructures are wary of change, as they are not the ones who will benefit from a more interdisciplinary provision of infrastructure.

It is therefore clear that some incentive in relation to the researchers and infrastructure providers will be required to bring about change. This could be in the form of supplementary funding to provide data sharing, or through mechanisms to give recognition where openness is well achieved. For example, in the US, the BD2K programme is issuing supplements to grants to enhance interoperability[26], and the GMOD project[27] is providing a collection of open source software tools for managing, visualising, storing, and disseminating genetic and genomic data. These initiatives will hopefully lead to better interoperability.

2.3.2 Data Interoperability

At the policy level, although the RCUK data principles and Concordat on research data are significant moves towards harmonisation. However, it is clear that different types of data lead to different requirements for its preservation and access. These different requirements depend on the nature of the data rather than the funder of the research. For example, whilst it is more likely that environmental data will be collected as part of research funded by NERC, policy concerning

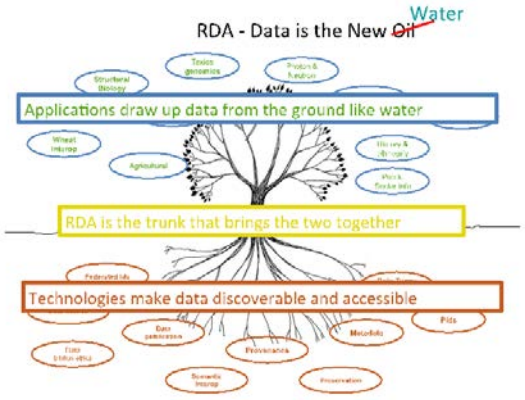
environmental data should apply to environmental data regardless of the research funder. Similarly, whilst privacy is more likely to be a consideration for research funded by MRC or the ESRC, the same policy regarding privacy should apply across all research (see section 3.1.3). Furthermore, these policies should be consistent across geographic boundaries.

Heterogeneity of data formats makes it hard to merge data from different sources, but standardisation across disciplines is a difficult social endeavour as the broader the standard, the harder it is to reach agreement. Furthermore, many standards already exist, and attempts to unify standards run the risk of instead creating more: the "Esperanto model" of standardisation, where a group agree on a new standard that unifies the existing ones can lead to adoption problems due to legacy, whereas the "Imperial model" where advocates of each standard push for it wider adoption can entrench differences.

In fact, inertia due to legacy, and the high cost of change, both mean that an effective solution is more likely to be found by mapping between standards rather than identifying a single common form. Progress is likely to be made piecemeal between similar fields, and this is underway, for example in European domain networks and ESFRI clusters such as BioMedBridges[28], SeaDataNet[29], and ENVRI[30].

2.3.3 The Research Data Alliance

The Research Data Alliance (RDA)[31] is dedicated to building social and technical bridges that enable open sharing. It addresses both the technological and data interoperability challenges described above. RDA has over 4000 members from 110 countries and provides a neutral space where its members can come together to form Working Groups that address particular problems in data-sharing. Some groups focused on domain specific issues whilst others on general technological problems. For example, there are domain-focused groups in Agrisemantics, BioSharing, Rice Data Interoperability, Wheat Data Interoperability, Agriculture Data, Biodiversity Data Integration, Global Water Information, Health Data, Marine Data Harmonization, Metabolomics Data Interoperability, Quality of Urban Life, Materials Data, Photon and Neutron Science, and Structural Biology.
Other groups are focused on technological issues for example on Data Citation, Data Description Registry Interoperability, Data Security and Trust, Empirical Humanities Metadata, Publishing Data Bibliometrics, Research Data Collections, International Materials Resource Registries, Legal Interoperability, Reproducibility and New Paradigms for Data Discovery. RDA also has some community needs focused group, for example in: Development of Cloud Computing Capacity and Education in Developing World Research, Education and Training on handling of research data, and Ethics and Social Aspects of Data.



The diagram (right) suggests how RDA bridges between technology-focussed and domain-focused activities.

2.4 Structured and unstructured repositories

There are many different types of data repository provided by different types of stakeholder. Some repositories are domain or discipline specific, sometimes quite narrowly targeted at providing data of a particular type for a specific purpose. Other, broader, structured repositories are sometimes

provided by funders or at a national level. Unstructured repositories are also available and can provide persistence through a simple deposit process, albeit sometimes at the cost of paucity of metadata. Institutions also provide data repositories typically for use by their researchers during the research process. Many data repositories are run by not for profit organisations but for profit services are also available.

Given the discussion about provision of FAIR data above, it is clear that data is best made available through repositories where aggregation can add most value. When data is underpinning a publication the choice of data repository is often recommended by the journal. For example Nature's Scientific Data journal says that "Data should be submitted to discipline-specific, community-recognized repositories where possible, or to generalist repositories if no suitable community resource is available" It later goes on to say "We are glad to support the use of institutional or project-specific repositories, if they are able to mint DataCite DOIs for hosted data, and share data under open terms of use. *In areas where well-established subject or data-type specific repositories exist, we ask authors to submit their data in parallel to the appropriate resources."*

Where a leading subject repository exists, for examples in PDB for proteins, quality journals normally give precise guidance on which repository to use. However, where there is not such a clear leading repository, there is a risk that the need for fuller metadata descriptions in structured repositories may lead depositors to use unstructured repositories with lower metadata requirements where findability and reusability may be compromised.

Partnering between a journal or funder with a particular repository can also be a way of ensuring at least some repository provision exists. For example, PLOS have partnered with Dryad[32] to streamline the data deposit and paper submission and reviewing processes. NSF have also continue to support Dryad[33] for their fundees to use where no thematic repository is available.

3. Analysis of current provision of data infrastructure

3.1 How the type of data impacts on the required infrastructure

It is difficult to generalise about how the type of data impacts the infrastructure required, as this depends on how the data is to be used and reused. However, some issues are general.

3.1.1 Issues related to scale of data

When data is large it makes little sense to copy it or even to move it and it may be best to keep data close to its source, or move it just once to the right repository. However, the kind of access required to the data can vary and this will influence the infrastructure required for the repository. Sometimes researchers will need access to just a small part of a large data archive, for example, a particular dataset from the collection of all data recorded at a synchrotron. In this case, the only searching that needs to be done at scale is to locate the data and, provided this search can be done against the metadata, it is therefore not compute intensive. The resulting data is also relatively modest in size. Other times, however, researchers require analysis over large data, for example, in particle physics or genomics, there is a need to compare data values across a wide range of data sets. In these cases, compute resource needs to be collocated with data so that the data does not have to be moved.

Additionally, the necessary software has to be available on that resource. This may well be bespoke software tailored to the specific data.

When data is large, a cost-benefit judgement must sometimes be made between keeping the data and recalculating or recreating it. This judgement depends on where the reuse and verification opportunities lie. For example, where the process of reducing raw data to derived data is straightforward but compute intensive, as in a topological reconstruction, this process should only be done once as there is little scientific value in being able to reproduce it. In this case it is the derived data, rather than the raw data, that is scientifically valuable for reuse and needs to be shared. On the other hand, when it is the data reduction itself that is interesting, for example when there may be several alternative analysis available, it can be valuable to be able to redo the reduction using different software or with different assumptions. There can, therefore, be value in sharing the raw data as well as the derived data.

Another example of when it is not cost effective to archive data is when data is created from a computer model: depending on what use is being made of it, it may be easier to recreate data from the software and inputs than it is to store the outputs. Moreover, a sharing request could have an explicit purpose of checking the software or comparing the output to a different analysis. In this case, both the input and output data must be kept. Again, there can be no general rules and it is necessary to judge on a case-by-case basis for each type of data. It is therefore necessary that the data infrastructure support all these different situations. (See also Section 4.3.3 on logging.)

3.1.2 Issues related to heterogeneity of data

As described in Section 2.3.2 above, interoperability of data across disciplines is still some way off. What requirements, then, does this heterogeneity of data place on the data infrastructure? Whilst it is possible to build heterogeneous archives, the handing these different forms of data necessarily increases the complexity of any tools or services that process that data. This complexity can be avoided by providing tools that are targeted at particular communities; however, this may lead to missed opportunities for really novel cross-disciplinary research. On the other hand, as was described in Section 2.2, making metadata standards available in machine-readable forms through registries can enable the building of tools that do handle heterogeneous data.

3.1.3 Cases where openness needs to be restricted, for example due to privacy, security, ethics

The RCUK principles say that data may be restricted for legal, ethical and commercial reasons. Privacy, confidentiality and consent are particularly relevant to health related research concerning personal data and the use of formal data-sharing agreements is normally appropriate in these cases. Ideally, these agreements would lead to machine interpretable access policies that can then be implemented directly by the data infrastructure. This is an area in which there is much research, some of which has been brought together in the RDA working group on practical polices[34]. Implementing this kind of solution in a data infrastructure would necessarily be incremental, starting with some simple aspects of policy, such as authorisation lists, before extending to more complex aspects such as role based access. It is currently difficult to imagine the entirety of funder data policies on privacy being encoded in machine interpretable ways, particularly as there are a great many different funder policies. It may be necessary, therefore, to develop policy and machine interpretation together, which would require policy change or simplification. As in other domains, one can expect that simply attempting to encode a policy in software can help clarify and simplify it.

Anonymisation of data is another way to address issues of privacy but, while anonymisation can be effective within one dataset, merging data from different sources risks deanonymisation. Safe

Havens[35] that provide a safe environment for data analysis, and Data Analysis as a Service where queries are performed remotely without direct access to the actual data are possible solutions, but again, deanonymisation of the resulting data by a combination with other data remains an issue.

3.1.4 Why have some domains developed data repositories sooner than others?

It is interesting to consider why some domains have developed a culture of data sharing more quickly than others. Is this due to different funders' policies or different scientific needs, or is it simply an accident of history brought about by the vision and innovation of particular individuals? Whatever the underlying reasons, it is clear that this has led to a current situation that is highly diverse, leaving researchers with a great deal of flexibility regarding how they should be curating and providing access to the data they are using.

In Genomics, for example, the science has driven the development, as there was a need to compare large sets of stings of ATCGs. Here, the data itself is a pattern and comparison of these patterns is a core requirement of the science. Conveniently, the problem of string comparison is a well-understood area of computer science that could to be applied to this specific domain.

Another area where pattern comparison is the basis of the science is in image interpretation. From an algorithmic point of view, this problem has some similarities string comparison, although it is considerably more complex. However, it is an area that has progressed rapidly as can be seen, for example, in the use of face recognition in some social networking platforms. This technology could perhaps be usefully deployed in research where image comparison and analysis is required for example, in neuroscience or oncology.

3.1.5 How is the challenge of data discoverability being addressed?

As described in Section 2.2 above, linking data to papers through Persistent Identifiers is relatively straightforward, nor is it really problematic if there are many different forms of PId, for example, where different databases each have their own bespoke PId scheme. However, PIds are not a solution to the issue of data discovery and metadata catalogues are only a solution where sufficient information to support the query is captured in the metadata. Until such time as standardisation or registration of data schemes has made it possible it is possible to search directly into the data itself, richness of metadata is the best way to address discoverability. The generation of rich metadata, ideally automatically, is therefore currently the key to enabling data discovery and, as happened with PubMed for publications, it would seem that strong policy in this area would lead to greater findability and accessibility.

3.2 Areas where improvements in data infrastructure could yield benefits

3.2.1 Cross linking data, papers and software

There is little precedent or infrastructure provision on how to link papers, data and software to trace the provenance of data and therefore verify its authenticity. To achieve traceability, researchers require access not only to the original data but also to intermediate and final data and to any software that produced it. Ideally, a user would be able to click on a graph in a paper and be directed to the data behind that graph and the software that generated it. The user would also be given the opportunities to rerun the analysis with the same, or an alternative, hypothesis; or to analyse the same data using different software; or to compare data from different sources. Rerunning the analysis is the easiest of these as it simply requires access to linked objects and the ability to run the software. It is more difficult to rework the analysis as this requires a deeper understanding of the

data. Merging data with other data is the most difficult of the three tasks as very precise semantics are required, so there is a risk of incorrect merging of data due to a mismatch in semantics.

The best way to assure data provenance is to keep data where it was originally created and link it in situ to other data rather than copy it to new place. The Biostudies Database[36] at EBI is aiming to provide this kind of system. Related data, that can be in different databases, spreadsheets, repositories etc. is linked via the Biostudies Database, which then provides a single place that aggregates all the relevant data regardless of its location. This is advantageous as it makes it possible to dynamically maintain the data record as it evolves independently of any static published papers.

3.2.2 Data provenance and forward traceability

The same infrastructure that provides the backward traceability for provenance described above, should also be able to provide forward traceability, from the data to the results that use it. This has obvious scientific benefits in terms of dependency tracking if and when new data becomes available but also has some other societal benefits. Firstly, it would enable subjects to ask what data about them has been used for. Noting that the data about an individual belongs to the subject, this would enable a subject to consent for their data to be used in one research study but not another. Secondly, by showing how results are derived from various data sources, it could also help with public understanding of science. This could help to engage the public and justify the expenditure of public money on research.

3.2.3 Role based attribution of credit for intellectual contribution to research

As suggested in Section 1.2, a model of credit attribution that gives more precise information about the different roles that individuals play in a piece of research would help remove the need for embargo periods on data. Attributions similar to that of the role-based credits in a film could give this information and place more emphasis on the importance and value of data production in the research community. A standard set of roles for people involved in the creation of a set of data would enable individual's contribution to be more accurately recorded. The CRediT taxonomy[37] produced by the CASRAI collaboration provides such a framework for publications defining 14 roles including, for example, Conceptualisation, Methodology, Software, Analysis, Curation, Writing, Supervision and Administration. The recording and publication of linked research objects would enable these separate contributions to be noted at a finer granularity for each component. The raw data for these attributions could in some cases be recorded automatically through the authentication system in the infrastructure as data files are created.

3.3 Sustainability of Data Infrastructure

Despite the challenges above, it is clear that the current data infrastructure available to researchers supports high quality science and that it is being funded by one means of other. The question is whether it can be more effective and whether there is additional value that can be added by changing the way it is provisioned.

3.3.1 Business models for data infrastructure

Neylon[38] provides an abstract framework for discussion about financial models for infrastructure provision based models of the provision of public goods from economics. He concludes that sustainability follows from trustworthiness of institutions with transparent community governance.

In *Comparing Approaches for the Sustainability of Scientific Data Repositories,* Downs and Chen[39] define a typology of approaches to sustainability, contrasting discrete revenue stream models, based on usage fees, subscriptions, grants, advertising, donations or subsidies, with cooperative revenue stream models, based on institutional commitments, bilateral and multilateral sharing, and long term commitments from stakeholders and funders. Based on this typology, the authors then argue[40] for a mixed revenue model as a way to reduce risk to long term stewardship.

The RDA-WDS Cost Recovery Interest Group[41] analysed income streams for a range of 25 data repositories covering a broad range of research areas and with both domain and national scope. It found a variety of different funding models with approximately half of the repositories relying on structural funding for their primary source of income, often supplemented by other sources such as R&D projects. Given the underlying motivation of openness, it is not surprising that only two of the repositories raised income from data access fees, whereas it is perhaps more surprising, given the acceptance of article processing charges for papers, that only two of the repositories were primarily funded through data deposit fees. They speculate that data deposit fees may in future gain greater stakeholder acceptance but express concern that this model of funding could for economic reasons drive down commitment to quality of curation.

Economic models for provision of data infrastructure have also been considered in several recent reports from research funders. In May 2016, Science Europe and the Knowledge Exchange published the results of an extensive survey on F*unding research data management and related infrastructures (SEKE)*[42] and the EC Communication on European Cloud Initiatives (ESOC) and High Level Expert Group on European Open Science Cloud (HLEG) referred to in Section 1 also comment on this topic. Further analysis of sustainability and coordination of data infrastructures is currently being undertaken by OECD with results expected to be available in Spring 2017 [43].

Two of the issues raised in these reports are summarised below.

3.3.2 Project based funding of data infrastructure

Current data infrastructure is being provided by a variety of means at European, National and Institutional levels. Many of these are short-term project-based initiatives. HLEG reports that

> "*The short and dispersed funding cycles of core research and e-infrastructures are not fit for the purpose of regulating and making effective use of global scientific data.*"

Likewise, SEKE concludes that:

> *"Most funding mechanisms are geared to funding research on a project basis, whereas the services and infrastructure for data management and access require a good amount of permanence"*

and that:

> *"Sustainability of RDI/RDM is at risk as long as funding is project-based."*

This is because, whilst project based funding is fine for duration of grant, curation and provision of access to outputs implies on-going cost after the grant. Although institutional level on-going costs can be funded as indirect costs, institutional provision of on going data access still leaves problems of piecemeal provision and consequential lack of coordination. Also, costs related to data curation are often seen by institutions as a costly liability rather than an investment in the maintenance of an asset (described in section 1.2 above). It is worth noting that other core research infrastructures like large facilities or networking are not provided in this way.

HLEG also gives an estimate of the scale of funding required to provide appropriate data stewardship:

"…we expect that on average about 5% of research expenditure should be spent on properly managing and stewarding data."

3.3.3 Coordination

A side effect of project-based funding is a lack of coordination. Both ESOC and HLEG identify fragmentation as a problem.

"Data infrastructures are split by scientific and economic domains, by countries and by governance models" [ESOC]

and

"…the components needed to create a first generation EOSC are largely there but they are lost in fragmentation" [HLEG].

SEKE concludes that there is a lack of coordinating or strategic approach:

"…business models for sustainable entities need to be developed, and responsibility for maintaining the data produced during projects (operations around curation, storage, archiving, sharing) needs to be defined and assigned. This requires more coordination, involving many actors, levels and disciplines."

3.4 Some examples of large scale centres for Data Infrastructure provision

One example where centralised service provision is overcoming some of these problems by providing cloud computing alongside some critical data sets is the **NIH Commons** described in Section 1.4. This kind of service also enables measurement of use and impact of datasets in long term and assessment of which are worth keeping. Such measurement should also help with attribution of credit issue described in Section 3.2.3 above.

The following sections describe briefly some other examples of large scale centres for the provision of data infrastructure.

3.4.1 DANS (Data Archiving and Networked Services)

DANS is an institute of the national research council and the national academy of the Netherlands. Dillo 2016[44] reports that about two thirds its budget come through structural funding from these two organisations and one third from R&D projects. However, compared to the growing demands placed on DANS, the structural funding is inelastic and the project funding is time consuming to acquire. Therefore DANS is exploring other income streams such as hourly fees for processing and documenting individual data deposits and fees for archiving services including institutional agreements with universities and (dark) archiving services for public or private third parties.

3.4.2 NERC JASMIN

NERC has for many years centralised much of its research data holdings through the 5 NERC data centres. In 2012, NERC went a step further by commissioning the JASMIN "super-data-cluster" facility[45] that delivers a complete computational environment for data analysis. JASMIN is half super-computer and half data-centre linked together by a high bandwidth networking and enables environmental researchers to bring their processing tasks to the data. JASMIN represents an investment of about £15M in 3 phases over 5 years.

3.4.3 ELIXIR

Established in 2013, ELIXIR is a collaboration of European Life Science Organisations building pan-European infrastructure for biological information to handle a rapidly growing volume and variety of data from high-throughput experiments such as DNA sequencing. ELIXIR also has Nodes in 20 European Countries and its central coordinating Hub is based on the Wellcome Trust Genome Campus in Hinxton alongside EMBL-EBI and the Wellcome Trust Sanger Institute. In 2014 ESRFI selected ELIXIR as one of three prioritised Research Infrastructures projects which led to an award of the €19M ELIXIR-EXCELERATE EU project to accelerate the implementation of Europe's life-science data infrastructure.

ELIXIR Nodes provide a large number of important software tools and data services[46] that are critical elements of the computing infrastructure for life science research, including PDBe[47] and EuropePMC[48], and a registry to help researchers navigate them.

ELIXIR Gateway[49] on F1000Research publishes outputs from all ELIXIR's activities and ELIXIR AAI[50] provides tools for Authentication and Authorisation of users of ELIXIR services. ELIXIR has established the ELIXIR Bridging Force Interest Group[51] to connect ELIXIR with relevant RDA Interest and Working Groups such as those on agricultural data, big data analysis, federated identity management, marine data, structural biology, toxicogenomics, and data publishing.

4 Recommendations for development of data infrastructure

By considering the three aspects of research communication identified in Section 2.1 in the light of the analysis in Section 3, we identify some key components of a data infrastructure that could be further developed to increase support for open research. We touch only briefly on the dissemination and value judgement aspects as they relate more to other reports in this series, before going on to give a more detailed recommendation on provenance.

4.1 Supporting collection and dissemination through Repositories

The communication of papers is already well served by PMC and other repositories such as OpenAire that is actively aggregating content from other repositories at the European level. Furthermore, new forms of publishing, such as those provided by Wellcome Open Research through F1000Research that separate the dissemination and value judgement aspects of communication are a significant step towards the separation of concerns described in section 2.1.

For communication of data, many domain-specific data repositories, such as those provided by EMBL-EBI and NCBI, are available and effective in many fields relevant to Wellcome researchers. The challenge here is to broaden the range of fields served and to make this data interoperable. Interoperability is the topic of Report D in this series on establishing data standards, metadata and interoperability so is not discussed further here. Breadth of coverage is partly about culture and incentives, the topic of Report A, but also about availability and ease of use of technology, that is a feature of the infrastructure. For fields where there is little or no culture of data sharing, in particular, it is particularly important that there are no technological barriers to overcome.

4.1.1 An easy start repository for data sharing

As described in Section 2.4, when depositing data in a repository there is often a trade-off between ease of deposit and richness of metadata and hence Findability and Reusability. For these areas therefore, an easy start solution, focused on simplifying as far as possible the process of deposit, is likely to be more acceptable and lead to have greater uptake. The ability to deposit relevant data along with the submission of a paper and have that seamlessly integrated in the review process, as can be done in Wellcome Open Research, F1000Research or Dryad is such a solution.

A further step in this direction can then be taken in order to support the deposition of a collection of different data sources that may underpin a result into an aggregate object. The EBI Biostudies database[52] described in Section 3.2.1 provides an example of how this can be done by providing a place to link together different sorts of objects.

As described at the end of Section 2.3.1, there is also scope for motivating uptake through the provision of grant supplements as is being done in BD2K, and development of easy to use tools as in GMOD.

The situation for software is more fluid. Although well managed collections of important software tools are provided by major data centres such as NCBI and EMBL-EBI and excellent support systems for community software development and dissemination are also available through general-purpose software repositories such as gitHub, there is little domain-specific guidance aimed at smaller software development teams on which of these repositories to use and how to use them. Provision of guidance and support to grantees on how best to use of these resources might significantly improve uptake and make a dramatic difference to the accessibility of software developed in grants.

4.2 Supporting the judgement of scientific value through peer review

Historically, this is an area in which significant added value has been provided by journals. The power of the journal lies largely in the reputation vested in its title, as this underpins a virtuous cycle wherein a respected journal can attract high quality papers and can therefore provide both a high standard of review and a high threshold for acceptance. The reputation of the title is then reified through the journal's Impact Factor, which perpetuates the cycle. However, the impact factor has been widely criticised as a blunt tool for assessing value, and article level metrics have been advocated as an alternative[53].

New modes of publishing, such as Wellcome Open Research, step away from this model and provide a platform for open article-level reviewing. By separating the judgement of scientific value from the dissemination and provenance functions described in Section 2.1, it will be possible to deliver a more open market in provision of these value adding services. Services that add value by assessing quality, or by collecting relevant articles, can be produced independently of dissemination function and publishing platform. Given that value is a domain-specific judgement, i.e. that an article of high value to one researcher may be of little or not interest to another, these new modes of publishing should lead to the development of new article discovery services tailored to the specific needs of particular research communities.

These platforms thus present a new opportunity: building on the infrastructure provided for open article-level reviewing, it should be a simple matter to set up new services which collect and promote articles that are relevant to particular communities. These new "titles" would provide an alternative to one of the functions of some conventional journals, that of aggregation of material relevant to a particular field. Support for setting up such titles could easily be provided along with

the platform with a system similar to "EasyChair" for Conferences[54]. This point relates to Report A in this series, which focuses on embedding cultures and incentives to support open research.

4.3 Supporting the assurance of Provenance.

The remaining aspect of research communication, the demonstration of provenance, provides more significant challenges for the data infrastructure. While uploading data to support a paper, as can be done for example in F1000research, is an important first step in the right direction, providing data in this way can be onerous and is only partial. Firstly, the researcher has to identify which data is necessary to underpin the paper and upload it, which is labour intensive. Secondly, the resulting data may only be a partial record of the process undertaken to produce the paper. Information about how the data was created, and about other data that was collected but is not relevant to the results in the paper, is likely to be lost.

In order to gain access to the whole provenance trail, access to the whole environment in which the research was conducted must be available. However, research institutions are, justifiably, unlikely to want to grant open access to resources within their own infrastructure. If the full provenance trail is to be available, the research itself is best done in an environment where it can later be shared.

To enable this, the funder could provide, along with the award, the entire data infrastructure for carrying out computational aspects of the research via a cloud service. Where possible, as well as the data storage and compute resources, this infrastructure would also provide the software required for the analysis (or software development support if the software is not available a priori). Researchers would then carry out their research using this platform rather than the resources of their own institution and, at the appropriate time, the resulting data and other outputs would be curated and made accessible in situ, avoiding the need to copy them over to another infrastructure which risks losing the connection between the items.

The infrastructure would need to be controlled enough to be able to provide the provenance trail, while also being flexible enough to adapt to the needs of individual researchers. To ensure such an infrastructure would be well used, it would have to be such that researchers prefer it to that of their home institution. Support would need to be at least as effective as local support would be, but could easily in fact be more effective, as the scale of the infrastructure could enable specialised support staff dedicated to each field of study.

The resulting linked research objects, comprising all data and other items created or used during the research, would be produced automatically during the research and the infrastructure would support the curation of these and persistent access to them. In this way the provenance of the data trail leading to the conclusions in the paper would not be lost, as the paper would link to the original data and its metadata in the original environment, rather than simply displaying a copy of the data alone. It would also support the role-based attribution of credit described in Section 3.2.3 by recording the contributions of different researchers. If effective, the infrastructure should also provide more cost-efficient access to resources through resource sharing and economies of scale.

Access to the infrastructure and the appropriate level of resources would be provided along with the research grant and would support not only the publication of data but also the linking of description, methodology, software, and data. This infrastructure would provide:

1) A persistent data storage space with access control mechanisms, allowing data to move from private to public accessibility during the course of the research.
2) A compute environment that tracks provenance of data, for example, by logging which data files were created from which other files and by which software.

3) Provision of key software and an environment to install and develop other software.
4) Logging of which researchers carried out which tasks during the use of the environment, providing the basis for role-based attribution of credit for those involved in the computational/data tasks.
5) Time stamping of activities for demonstration of precedence in attribution.
6) The ability to automatically identify and collate the set of files (and their log entries) that support a given set of results in a paper into an identified package.
7) A means to link these packages to particular points in a paper in an open archive.

We expand on each of these in turn below noting that, as much of this infrastructure needs to be targeted to specific types of research and research communities, implementation via pilots for specific fields of research is perhaps the most effective way to start its development.

4.3.1 Storage

Storage can easily be provided "in the cloud" however a persistent mechanism to identify and timestamp each file would be required (see below). It would be advantageous also not to distinguish between scratch space and persistent space, but rather to have one type of space, and the ability to promote scratch files to persistent ones as required as the research progresses.

It will be necessary to tailor the data services to the requirements of the particular methodology of research being pursued. However, this should be straightforward the Data Management Plans (DMPs) for the research should state what the resource requirements are. This may require additional guidance on DMP content.

Where the volume of data is not an issue, write-once storage could be adequate. Where volume is an issue and there is a need to delete data in order to reuse storage, metadata should still be kept in the log so as to be able to know what was done, redo it where possible, and experiment with alternatives (see paragraph on logging below).

Note that sometimes data created within the infrastructure may need to link to data held elsewhere, for example, where data was created on a large facility like a synchrotron. (See paragraph on Linking below).

4.3.2 Computation

Compute resource should be provided alongside the storage. The reason for this is not only to eliminate the need to move data, but also to allow logging of the data creation trail. Where it is expected that significant compute resource will be required, this should be described and justified in the DMP, along with any specific software requirements so that consideration can be given to these requirements as part of the evaluation of the cost of the research.

It may initially be possible to give unlimited time and storage on this compute infrastructure, while reserving the right to later introduce a credit based system for its use. However, it would probably be preferable to introduce quotas immediately in order to instil a culture of careful use. These quotas should be sufficiently generous that it makes no sense for a researcher to use an alternative infrastructure.

Except for cases where new software is being developed as part of the research, software requirements should also be detailed in the DMP so that access to the required software can also be provided. Where new software is being developed, an environment for this development could be

provided as part of infrastructure so that the software is developed in a maintainable form and support in the use of the infrastructure can be given.

4.3.3 Logging

Once computing resources for the research are provided in a controlled environment, it is possible to create logs of what is done, and therefore, provenance trails. A careful balance must be found between prescription and flexibility, so as not to constrain the researcher while enabling the logging.

One way in which this could be done would be to wrap the software components in shells that log the input and outputs (both file based and interactive IO). Which researchers carry out which tasks would also be logged to enable the recording of role based attribution. All of the resulting pieces of information would require PIds.

4.3.4 Linking and Packaging

The Logs described above provide the raw information for the creation of links that enable provenance trails and the rerunning of analysis. The processing of log files to create trace trees could be done incrementally or retrospectively at the point at which the researcher asks for it. The identification of the resulting package would be through a PId that would be referred to in the paper.

Although PIds provide the basis for identifying packages, a classification of PIds would assist the analysis of what took place during the research. An RDA Working group on PId Information Types has defined Metadata schema and API for the classification of PIds[55]. These can begin simply and build in richness over time.

4.4 Implementation of Recommendations

This section describes a possible approach to implementing the recommendations in Sections 4.1.1 and 4.3. It attempts to give some indication of timescale and budget, however it is important to note that this is very speculative and significantly more detail would have to be worked out to give a reliable estimate.

4.4.1 Implementing an easy start repository for data sharing

The provision of the easy start repository described in Section 4.1.1 is itself a simple thing to deliver. Several providers of such services exist and provision would be best achieved through an agreement with one of these to tailor services and support for particular targeted communities. This would be reinforced by communications and policy guidance from the funder. The cost of this provision would thus scale according to usage and it therefore difficult to estimate. It is unlikely to require more than a relatively modest investment.

The rest of this section relates to the Recommendation in Section 4.3.

4.4.2 Provenance assured infrastructure

For the provenance assured infrastructure described in Section 4.3, it would be sensible to begin with a number of pilot studies using this form of infrastructure provision as this would be an effective way to test if this is a feasible way of working and developing the ideas further. These pilots should be in some already well-advanced areas such as biomedical or large facilities. The infrastructure would then be built incrementally as pilot projects are established that will use it.

As motivated in Section 3.1.4, choosing a pilot that involves images would be helpful as image processing across different sets of images is a key emerging area for added value.

4.4.3 Resourcing

Storage, Compute and Software

To begin, storage and compute resources would not need to be large, but could grow as required by DMPs in approved pilot projects as these resources can be procured relatively quickly. Similarly, the choice of which software to install on the infrastructure could be done on an as-needed basis. Resource requirements would be sized as part of the choice of which pilots to undertake. In the first phase of development, selecting a few pilots for which scale of data is not an issue would reduce risk as data and compute resource requirements would be modest.

Logging, packaging and linking

A solution for these functions must be designed and implemented as part of the development of the infrastructure above. A small team of two or three senior developers, working in an organisation with experience in the domain, should be able to produce an initial minimally functioned system within two years. This initial system would then require continued development over several more years to reach full functionality.

Operation

A small operational team would be required to run the service and assist its users. This might be an individual or a small team, depending on the number of pilots. Initially, a highly collaborative model of cooperation between the infrastructure's first users and the development and support team would be essential in order to define and implement the service.

4.4.4 Timescale and budget

Timescale

An initial commitment for five years should be sufficient to develop the infrastructure to a level at which it could be evaluated and its value assessed. This would consist of a two year set up phase, followed by the running of pilots for a further three years, with assessment in the fourth and fifth years. An outline of some key steps in this development is given in the table below.

Year	Activities	Outputs
Year 1	• Set up the project team • Define the specifications for the system • Begin implementation of the infrastructure • Announce an invitation for pilots to begin in Year 2 • Perhaps solicit some applications where an easy start can be expected	Detailed specification of system Partial Implementation Criteria for selection of pilots
Year 2	• Continue implementation • Procure and set up data and compute resources • Receive and assess the applications for use of the infrastructure • Work with the applicants to select pilots (the pilots can be new projects or add-ons to existing projects).	Complete implementation Availability of data and compute resources Selection of pilot studies
Year 3	• Set up and run the first round of pilots. • Continue to develop the infrastructure depending	Infrastructure working for pilots

	on the needs of the pilots. • Request a second round of pilot applications	Results of first usage trials
Year 4	• Refine or adapt the model of provision • Initial assessment of the success, or otherwise, of the first round pilots. • Select the second round of pilots	First fully provenanced outputs of research Initial assessment of added value with suggestions for improvements.
Year 5	• Further assessment of success, with a decision to continue or stop the project. • If the decision is to continue, start the second round of pilots • Continue development if considered successful, or close down if not	Results of the pilots Availability of research outputs with full provenance trail Assessment of added value Decision on whether to continue

Speculative Budget

The budget required to develop and assess such a fully featured infrastructure depends crucially on the number and scale of the pilots to be supported. However a certain minimum size and duration of activity would be needed for viable project with a reasonable likelihood of a successful outcome. At the lowest end of the scale, a project to support a few pilots with modest data requirements could perhaps be run with a budget of about £250,000 per year for the first 5 years. A broader project supporting more varied pilots, some of which could have more significant data requirements, would require perhaps double or treble that amount. At the other end of the scale, an initiative to provide an infrastructure that would support a whole research area or range of research areas would require a profiled investment of some tens of millions of pounds over a five to ten year period.

References

[1] In 1997, the US National Research Council argued that "full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research." US National Research Council, *Bits of power: Issues in global access to scientific data* (US National Research Council : Washington, 1997).

[2] The Royal Society's 2012 report on Science as an Open Exercise, <https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf>.

[3] G8 London, 2013 <https://www.gov.uk/government/publications/g8-science-ministers-statement-london-12-june-2013>

[4] Many of these sets of principles are the same captured in the FAIR principles. Wilkinson, M. D. *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship* (*Scientific Data* 3: Article 160018, 2016, doi: 10.1038/sdata.2016.18). <http://www.nature.com/articles/sdata201618>.

[5] NSF policy: <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>.

[6] NIH policies for various fields of research: <http://grants.nih.gov/policy/sharing.htm>.

[7] G7 2015 statement <http://www.g8.utoronto.ca/science/2015-berlin.html>.

[8] European Commission issued the Communication on European Cloud Initiatives <http://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/1-2016-178-EN-F1-1.PDF> (April 2016).

[9] UK Concordat on Research Data <http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf> (July 2016).

[10] EC Communication on European Science Cloud Initiatives, 19th April 2016, https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe. A concise summary of the Communication is provided in the UK Parliamentary Explanatory Memorandum EM 8099 http://europeanmemoranda.cabinetoffice.gov.uk/files/2016/05/EM_8099-16.pdf which is quoted here.

[11] A Cloud on the 2020 Horizon, Commission High Level Expert Group on the European Open Science Cloud - Realising the European Open Science Cloud: first report and recommendations, 20 June 2016, https://www.eudat.eu/sites/default/files/HLEG%20EOSC%20first%20Report.pdf

[12] Riding the Wave <http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707>.

[13] The Data Harvest <https://ec.europa.eu/digital-single-market/en/news/data-harvest-report>.

[14] Margolis R, et al. *The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data*, J Am Med Inform Assoc 2014;21:957–958. doi:10.1136/amiajnl-2014-002974

[15] NIH news release, October 9, 2014

https://www.nih.gov/news-events/news-releases/nih-invests-almost-32-million-increase-utility-biomedical-research-data

[16] NIH Commons project <https://datascience.nih.gov/commons>.

[17] Wilkinson, M. D. *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship* (*Scientific Data* 3: Article 160018, 2016, doi: 10.1038/sdata.2016.18) <http://www.nature.com/articles/sdata201618>.

[18] Note that verifiability is not synonymous with reproducibility. It is clear that some research is not reproducible, for example, where observations are made of a changing environment, or where access to a particular unique reagent is required. For such research, whilst the raw data may be irreproducible, the analysis can still be verifiable.

[19] Wellcome Open Research <http://wellcomeopenresearch.org>.

[20] F1000Research <http://f1000research.com>.

[21] Datacite <https://www.datacite.org>.

[22] The Open Metadata Registry <http://metadataregistry.org>.

[23] Simple Knowledge Organization System (SKOS), W3C <https://www.w3.org/TR/skos-primer>.

[24] Metadata Standards Agency, see <http://rd-alliance.github.io/metadata-directory/> and <https://rd-alliance.org/group/metadata-standards-catalog-wg/outcomes/metadata-standards-directory-wg-recommedations.html>.

[25] Following the G8 definition quoted in Section 1.1 above, this includes “assessable, intelligible, useable, and interoperable to specific quality standards”.

[26] Supplements to Support Interoperability of NIH Funded Biomedical Data Repositories (Admin Supp), <http://grants.nih.gov/grants/guide/pa-files/PA-15-144.html>.

[27] Generic Model Organism Database project <http://www.gmod.org/wiki/Main_Page>.

[28] BioMedBridges <www.**biomedbridges**.eu>.

[29] SeaDataNet <www.seadatanet.org>.

[30] ENVRI <http://envri.eu>.

[31] Research Data Alliance <rd-alliance.org>.

[32] Make data sharing easy: PLOS launches its Data Repository Integration Partner Program http://blogs.plos.org/tech/make-data-sharing-easy-plos-launches-its-data-repository-integration-partner-program/ Retrieved 8 September 2016.

[33] NSF sustaining their support for open data, 2016/09/06, https://blog.datadryad.org/2016/09/06/nsf-sustaining-their-support-for-open-data/, Retrieved 8 September 2016

[34] RDA Practical Policy WG, **https://rd-alliance.org/groups/practical-policy-wg.html**

dx.doi.org/10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC

[35] For example, see <http://www.acmedsci.ac.uk/viewFile/53c7d8a7567db.pdf>.

[36] The EBI Biostudies Database <http://www.ebi.ac.uk/biostudies/>.

[37] CRediT role taxonomy, http://casrai.org/credit

[38] Cameron Neylon, *Squaring Circles: Economics and Governance of Scholarly Infrastructures*, SCIDATACON Session on Sustainable Business Models for Data Repositories, Sept 13 2016. http://www.scidatacon.org/2016/sessions/45/paper/190/

[39] Robert S. Chen, Robert R. Downs, 2013, Comparing Approaches for the Sustainability of Scientific Data Repositories, Columbia University Academic Commons, http://hdl.handle.net/10022/AC:P:19169.

[40] Robert R Downs, Robert S. Chen, A Portfolio Approach to a Sustainable Business Model for Scientific Data Stewardship, SCIDATACON Session on Sustainable Business Models for Data Repositories, Sept 13 2016, http://www.scidatacon.org/2016/sessions/45/paper/273/

[41] Income Streams for Data Repositories, Final report, V. 1.00, 10 February 2016

https://rd-alliance.org/sites/default/files/attachment/Income_Streams_for_Data_Repositories-FINAL-160210.pdf

[42] Funding research data management and related infrastructures, Knowledge Exchange and Science Europe briefing paper, May 2016. http://www.scienceeurope.org/uploads/PublicDocumentsAndSpeeches/SE-KE_Briefing_Paper_Funding_RDM.pdf

[43] OECD has two projects running at the time of writing:

SUSTAINABLE BUSINESS MODELS FOR DATA REPOSITORIES:

https://innovationpolicyplatform.org/system/files/SUSTAINABLE%20BUSINESS%20MODELS%20FOR%20DATA%20REPOSITORIES_0.pdf

and INTERNATIONAL CO-ORDINATION OF CYBER-INFRASTRUCTURES FOR OPEN SCIENCE

https://innovationpolicyplatform.org/system/files/INTERNATIONAL%20CO-ORDINATION%20OF%20CYBER-INFRASTRUCTURES%20FOR%20OPEN%20SCIENCE_0.pdf

[44] Ingrid Dillo, The challenge of a business model with diverse income streams, SCIDATACON Session on Sustainable Business Models for Data Repositories, Sept. 13, 2016, http://www.scidatacon.org/2016/sessions/45/paper/59/

[45] JASMIN, http://www.jasmin.ac.uk/

[46] ELIXIR, https://www.elixir-europe.org/services

[47] Protein Data Bank Europe, http://wwwdev.ebi.ac.uk/pdbe/

[48] Europe PMC, http://europepmc.org/

[49] F1000research, http://f1000research.com/channels/elixir

[50] Elixir Services, https://www.elixir-europe.org/services/compute/aai-overview

[51] Elixir Bridging Force, https://rd-alliance.org/groups/elixir-bridging-force-ig.html

[52] The EBI Biostudies Database <http://www.ebi.ac.uk/biostudies/>.

[53] A Comprehensive Assessment of Impact with Article-Level Metrics, PLOS <https://plos.org/article-level-metrics>.

[54] EasyChair <http://easychair.org/>.

[55] RDA PId INfomratino Types Working group output. https://rd-alliance.org/groups/pid-information-types-wg.html

dx.doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786

October 2016

Version 1

Wellcome exists to improve health for everyone by helping great ideas to thrive. We're a global charitable foundation, both politically and financially independent. We support scientists and researchers, take on big problems, fuel imaginations and spark debate.

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