Data Storage

Christopher Frederick Isambard Blumzon
and Adrian-Tudor Pănescu

Contents

Abstract

While research data has become integral to the scholarly endeavour, a number of challenges hinder its development, management and dissemination. This chapter follows the life cycle of research data, by considering aspects ranging from

C. F. I. Blumzon (✉)
Figshare, London, UK
e-mail: c.george@digital-science.com; chris@figshare.com

A.-T. Pănescu
Figshare, London, UK

"Gheorghe Asachi" Technical University of Iași, Iași, Romania
e-mail: tudor@figshare.com

A. Bespalov et al. (eds.), *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*, Handbook of Experimental Pharmacology 257,
https://doi.org/10.1007/164_2019_288

storage and preservation to sharing and legal factors. While it provides a wide overview of the current ecosystem, it also pinpoints the elements comprising the modern research sharing practices such as metadata creation, the FAIR principles, identifiers, Creative Commons licencing and the various repository options. Furthermore, the chapter discusses the mandates and regulations that influence data sharing and the possible technological means of overcoming their complexity, such as blockchain systems.

Keywords
Blockchain · FAIR · Identifier · Licence · Metadata · Preservation · Repository · Reproducibility · Storage

1 Introduction

The evolution of scientific research has been recently shaped by the so-called reproducibility crisis, a phenomenon brought to light by a number of studies that failed to replicate previous results (see, e.g. Eklund et al. 2016; Phillips 2017). This highlighted the necessity of making available the research data underlying studies published in more traditional mediums, such as journals, articles and conference papers, practice which was promptly mandated by both funding bodies (European Commission 2017; National Institutes of Health 2018) and the publishing industry (Elsevier 2018; Springer Nature 2018).

This has left the main actors of scholarly communication, researchers, in an interesting but also possibly difficult position. While the necessity of making data available, especially when it is generated by publicly funded research or it presents high-impact consequences, is uncontested, a number of challenges remain as neither the technical, legal or societal environments were fully prepared for this prospect.

In a 2017 study across Springer Nature and Wiley authors, less than half of the respondents reported sharing research data frequently, with more than 30% rarely or never sharing (Digital Science 2017). Various reasons for the lack of sharing have been identified, ranging from the lack of experience, technical knowledge or time, to fear of shaming (in case errors are discovered in the published dataset), or competition to published results (Federer et al. 2015; Youngseek and Zhang 2015). While the latter are more difficult to overcome, requiring profound changes across the whole scholarly communication spectrum, the more practical aspects can be solved by either technical means or specialised guidance.

This chapter attempts to provide a high level overview of all components encompassing research data sharing while detailing some of the more important aspects, such as storage, metadata, or data anonymisation.

2 Data Storage Systems

While data storage has a history closely linked to that of computing, research data presents a handful of new and unique challenges, especially when it comes to persistence and privacy. Usually most technical details will be handled by specialised professionals, but gathering a basic understanding of the inner workings, advantages and limitations of the various options can help devise data management plans customised to the needs of each research project, community or subject area.

2.1 Types of Storage

The first aspects of storage that need to be considered are the actual medium and employed technology; currently, the most prevalent options are:

- Magnetic (hard disk drives (HDD), magnetic tapes): data is stored using the magnetization patterns of a special surface.
- Optical (compact disks, Blu-ray): data is stored in deformities on a circular surface which can be read when being illuminated by a laser diode.
- Semiconductor: data is stored using semiconductor-based integrated circuits. While traditionally this type of technology was used for *volatile* storage (data is lost if electric power is not supplied, as opposed to magnetic or optical storage), so-called solid-state drives (SSD) are now included in consumer computers, offering a non-volatile option with superior access speeds to their magnetic counterparts.

When considering these options, various aspects need to be accounted for, such as convenience, costs and reliability. For example, while tape drives tend to be cheaper than hard disks (a 2016 analysis determined that 1 gigabyte of tape storage costed $0.02 opposed to $0.033 for HDD (Coughlin 2016)), they also exhibit slow data retrieval rates and require specialised hardware.

Reliability is one of the most important aspects when considering scientific data, as loss of information can lead to delays or even experiment failures. While in the early days of solid-state drives these encountered higher failure rates than HDD counterparts, a 2016 study puts them at comparable or even lower rates; under 2% of SSDs fail in their first usage year (Schroeder et al. 2016). Reliability can also be determined by brand and models; Klein (2017) determined an average 1.94% annual failure rate, but with a maximum at over 14% for a certain model. As no technology can offer absolute guarantees regarding reliability, other protection mechanisms, such as backups, need to be considered, these being discussed in the next section.

2.2 Features of Storage Systems

Most often the underlying technology for storing data can be of less relevance for the scholarly and scientific pursuit, but other characteristics can play an important role when choosing a solution.

The location of the data storage is the first considered aspect. Storing data on a local machine is advantageous as it allows the researcher to quickly access it, but might place obstacles when attempting to share it with a larger team, and also requires that the owner of the machine is fully responsible for its preservation.

Storage facilities managed at the institutional level, such as storage area network (SAN) systems, move the burden of managing data storage from the individual researcher to specialised personnel, providing higher reliability and enhanced possibilities for sharing data among peers.

Finally, data can be stored off-site in specialised facilities; this model became prominent with the advent of *cloud* systems, such as Amazon Web Services, Microsoft Azure or Google Cloud Platform, and has benefits in terms of reliability, scalability and accessibility. This might be preferred when the individual researcher or institution does not possess the required resources for managing a storage solution, when large quantities of data need to be stored, or when data needs to be shared across a large network of collaborators. At the same time, the privacy and legal implications need to be considered, given that a third party usually handles the storage; for a more in-depth discussion on this, see Sect. 4. It is worth noting that *cloud* deployments can also be managed by governmental bodies or similar official entities, this alleviating some of the legal issues (for example, the Australian National Research Data Storage Services provides such facilities to researchers in Australia, including storage of sensitive data, such as clinical trial sets (Australian Research Data Commons 2018)).

From a technical point of view, the choice of a storage solution needs to account for the following:

- Redundancy: as noted previously, no storage system can be guaranteed to function without faults and thus it is important that data is copied and stored on different systems simultaneously. The higher the number of copies and the broader their distribution, the higher the guarantee for their persistence is.
- Persistence and preservation: simply having data stored on a medium does not provide guarantees that, over time, it would not become inaccessible. For example, both tape drives and hard disks can become demagnetised, hence corrupting the stored data. This phenomenon is frequently described as *bit rot*. Hence, data needs to be periodically tested and, if problems arise, fixed. A common method for detecting issues employs so-called checksums, fingerprints of data files which change when even 1 byte switches value. If a file is detected to have changed, it is usually replaced with a redundant copy.

- Transformation: as technology evolves, so do the methods for storing data, this also leading to deprecation; for example, floppy disks are rarely used nowadays, despite being ubiquitous just a few years back. Storage and archival systems need to account for this and migrate data to current technological requirements, while ensuring that its contents are not semantically modified.

Of course, the emphasis on each of these requirements depends on the characteristics of the underlying data; for example, for raw data the transformation aspect might be less relevant, as that is not the final considered form of the data, but redundancy could play a more important role due to its sole existence as a research artefact.

2.3 Data File Formats

The file formats and even the structure and organisation of research data will most often be enforced by various laboratory instruments and software used for producing it. Nevertheless, it might be beneficial to apply transformations to the raw outputs in order to ensure their persistence over time, their ease of use, and the possibility for others to work with them.

While no silver bullet exists for file formats, a number of considerations can help choosing the right option. A first question regards the choice between proprietary and open-source file formats. Proprietary formats place a barrier for other collaborators that need to access the data file, as they might require special software or even hardware licences in order to read and modify these; for example, the Microsoft Office formats for storing tabular data (Excel files such as XLS and XLSX) can be opened only by certain applications, while comma-separated value (CSV) files can be manipulated by any text editor.

Another point to consider regards the standardisation of formats; file formats which are backed up by an established standard provide higher guarantees in terms of accessibility and preservation over time, as clear rules on how data is encapsulated and structured are defined. For example, the Digital Imaging and Communications in Medicine (DICOM) format is the de facto method for storing and transmitting medical information; using it guarantees that any systems that implements the standard can fully read the data files. Similarly, the Clinical Data Interchange Standards Consortium (CDISC) (2018) has developed a number of standards encompassing the whole research workflow, such as the Clinical Data Acquisition Standards Harmonization (CDASH), which establishes data collection procedures, or the Study Data Tabulation Model (STDM), a standard for clinical data organisation and formatting. As a counterexample, the CSV format does not have complete standards behind it, and thus, particularities can arise at both the structural and semantic levels.

Finally, the effects on the research life cycle need to be considered. Producing the data files is most often only the first step in the research workflow. Data files will need to be processed, shared with peers and even published alongside more

traditional outputs, such as journal articles. While it is highly improbable that data will maintain the same format along the whole cycle (published data rarely includes the whole initial raw dataset), informed choices can aid the process.

2.4 Dataset Structure and Organisation

Another important aspect regards the organisation and splitting of datasets. While a single file might seem a simpler method for storing a dataset, issues will arise when it grows in size, and it needs to be processed or transferred to other systems. Similarly, a large number of files can pose issues with navigating and understanding the structure; making a choice needs again to be an informed process.

The first point to consider is the dimension of the dataset. Large instances, especially those exceeding 1 terabyte, might prove difficult to handle. Consider for example the need to transfer such a large file over the network[1] and the possibility that the connection will drop during the operation; most often the transfer will need to be restarted, wasting the effort. In such cases, splitting the dataset might prove to be a better solution, as each smaller file can be transferred individually and independently of the others, network issues requiring only the retransfer of unsent files. A number of storage systems include an option for so-called chunked file transfers, where the system automatically splits larger files in smaller blocks, allowing these to be transferred independently and at any point in time.

In cases where a large number of files constitute a dataset, it is important to describe the overall structure such that other applications or human users can understand it. Traditionally, *folders* are used for categorising and structuring content, but these can prove ineffective in describing the full organisation, and certain systems might not even implement this facility. A common solution to this issue is including separate file(s) describing the structure, usually called *manifests*, along with the datasets. Preferably these would follow a standard structure and semantics, and for this purpose standards such as BagIt[2] and Metadata Encoding and Transmission Standard (METS)[3] have been established. Along with the structural description, these files can also contain technical information (e.g. checksums) that might ease other processes along the scholarly workflow, such as preservation.

3 Metadata: Data Describing Data

Storing research data can have many purposes, from facilitating study replication to allowing further hypotheses to be tested. Nevertheless, archiving only the data points, with no information regarding their purpose, provenance or collection

[1] Even over an optical fibre network connection 1 terabyte of data will require over 1 h to transfer.
[2] https://tools.ietf.org/html/draft-kunze-bagit-16.
[3] https://www.loc.gov/standards/mets/.

method, will exponentially decrease their value over time, as both other researchers and the authors of the data will find impossible to reuse them without further information about the syntax and semantics.

Metadata, *data about data*, is the information created, stored and shared in order to describe objects (either physical or digital), facilitating the interaction with said objects for obtaining knowledge (Riley 2017). Metadata can describe various aspects of the underlying dataset, and it is often useful to split the various attributes based on their purpose.

Descriptive metadata, such as the title, creators of the dataset or description, is useful for allowing others to find and achieve a basic understanding of the dataset. Often linked to this is the *licencing and rights* metadata that describes the legal ways in which the data can be shared and reused; it includes the copyright statement, rights holder and reuse terms.

Technical metadata, which most often includes information on the data files, such as their formats and size, is useful for transferring and processing data across systems and its general management. *Preservation* metadata will often enhance the technical attributes by including information useful for ensuring that data remains accessible and usable, such as the checksum or replica replacement events (see Sect. 2.2). Finally, *structural* metadata describes the way in which data files are organised and their formats.

Given its complexity, producing metadata can become a significant undertaking, its complexity exceeding even that of the underlying data in certain cases. This is one of the reasons for which the standardisation of metadata has become mandatory, this happening at three main levels.

At the structural level, standardisation ensures that, on one hand, a minimum set of attributes is always attached to datasets and, on the other, that enough attributes are present for ensuring proper description of any possible artefact, no matter its origin, subject area or geographical location. Multiple such standards have been developed, from more succinct ones, such as the DataCite Schema[4] or the Dublin Core Metadata Element Set,[5] to more extensive, such as MARC21[6] or the Common European Research Information Format (CERIF).[7] The usage of these standards might vary across subject areas (e.g. the *Document, Discover and Interoperate (DDI)* standard is targeted at survey data in social, behavioural and health sciences (DDI Alliance 2018)) or the main purpose of the metadata (e.g. the METS standard emphasises technical, preservation and structural metadata more than the DataCite schema).

At the semantic level, the focus is on ensuring that the language used for describing data observes a controlled variability both inside a research area and across domains. For example, the CRediT vocabulary (CASRAI 2012) defines

[4] https://schema.datacite.org/meta/kernel-4.1/.

[5] http://dublincore.org/documents/dces/.

[6] https://www.loc.gov/marc/bibliographic/.

[7] See https://www.eurocris.org/cerif/main-features-cerif.

various roles involved in research activities, Friend of a Friend (FOAF) establishes the terminology for describing and linking persons, institutions and other entities (Brickley and Miller 2014), and the Multipurpose Internet Mail Extensions (MIME) standard defines the various file formats (Freed Innosoft and Borenstein 1996).

The third point considered from a standardisation point of view involves the formats used for storing and transferring metadata. The Extensible Markup Language (XML)[8] is one of the most prevalent formats, almost all standards providing a schema and guidance on employing it. The JavaScript Object Notation (JSON)[9] format is also starting to gain traction, both due to its pervasiveness in web services nowadays and also due to certain initiatives, such as schema.org which use it as the de facto output format.

3.1 Unique and Persistent Identifiers

One important aspect of metadata, considered by most standards, vocabularies and formats relates to the usage of identifiers. Similar to social security numbers for humans or serial numbers for devices, when it comes to research data, the aim of identifiers is to uniquely and persistently describe it. This has become a stringent necessity in the age of Internet, both due to the requirement to maintain resources accessible for periods of times of the order of years or even decades, no matter the status or location of the system preserving them at any discrete moment,[10] and also due to the necessity of linking various resources across systems. Thus, various elements of research data started receiving identifiers, various initiatives and standards becoming available.

Even before the prevalence of research data sharing, bibliographic records received identifiers, such as International Standard Book Numbers (ISBN) for books and International Standard Serial Numbers (ISSN) for periodicals. For research data, Archival Resource Keys (ARK) and Handles[11] are more prevalent, as these mechanisms facilitate issuing new identifiers and, thus, are more suited for the larger volume of produced records.

The Digital Object Identifier (DOI) system[12] is quickly emerging as the industry standard; it builds upon the Handle infrastructure, but adds an additional dimension over it, namely, *persistence* (DOI Foundation 2017). At a practical level, this is implemented using a number of processes that ensure that an identified object will remain available online (possibly only at the metadata level) even when the original

[8] https://www.w3.org/XML/.

[9] https://www.json.org/.

[10] Similar to bit rot, link rot describes the phenomenon of web addresses becoming unavailable over time, for example, due to servers going offline. This can pose significant issues for research artefacts, which need to remain available for longer periods of time due to their societal importance; nevertheless, link rot was proven to be pervasive across scholarly resources (Sanderson et al. 2011).

[11] http://handle.net/.

[12] https://doi.org.

holding server becomes unavailable and the resource needs to be transferred elsewhere. A DOI is linked to the metadata of the object and is usually assigned when the object becomes public. The metadata of the object can be updated at any time and, for example, the online address where the object resides, could be updated when the object's location changes; so-called resolver applications are in charge of redirecting accesses of the DOI to the actual address of the underlying object.

A second important dimension of research outputs relates to persons and institutions. ORCiD is currently the most widespread identifier for researchers, with over 5 million registrations (ORCID 2018), while the International Standard Name Identifier (ISNI)[13] and the Global Research Identifier Database (GRID)[14] provide identifiers for research institutions, groups and funding bodies.

Identifiers have been developed for other entities of significant importance in terms of sharing and interoperability. For example, the Protein Data Bank provides identifiers for the proteins, nucleic acids and other complex assemblies (RCSB PDB 2018), while GenBank indexes genetic sequences using so-called accession numbers (National Center for Biotechnology Information 2017).

Research Resource Identifiers (RRID) (FORCE11 2018) aim to cover a wider area, providing identifiers for any type of asset used in the scientific pursuit; the current registry includes entities ranging from organisms, cells and antibodies to software applications, databases and even institutions. Research Resource Identifiers have been adopted by a considerable number of publishing institutions and are quickly converging towards a community standard.

The main takeaway here is that, it is in general better to use a standard unique and possibly persistent identifier for describing and citing a research-related entity, as this will ensure both its common understanding and accessibility over time.

3.2 The FAIR Principles

As outlined, producing quality metadata for research data can prove to be an overwhelming effort, due to the wide array of choices in terms of standards and formats, the broad target audience or the high number of requirements. To overcome this, the community has devised the FAIR principles (Wilkinson et al. 2016), a concise set of recommendations for scientific data management and stewardship which focuses on the *aims* of metadata.

The FAIR principles are one of the first attempts to systematically address the issues around data management and stewardship; they were formulated by a large consortium of research individuals and organisations and are intended for both data producers and data publishers, targeting the promotion of maximum use and reuse of data. The acronym *FAIR* stands for the four properties research data should present.

[13] http://www.isni.org/.

[14] https://grid.ac.

Findability relates to the possibility of coming across the resource using one of the many Internet facilities. This requires that the attached metadata is rich enough (e.g. description and keywords are crucial for this), that a persistent identifier is associated and included in the metadata and that all this information is made publicly available on the Internet.

Accessibility mostly considers the methods through which data can be retrieved. As such, a standard and open protocol, like the ones used over the Internet, should be employed. Moreover, metadata should always remain available, even when the object ceases to be, in order to provide the continuity of the record.

Interoperability considers the ways in which data can be used, processed and analysed across systems, both by human operators and machines. For this, metadata should both "use a formal, accessible, shared, and broadly applicable language for knowledge representation" (FORCE11 2015) and standardised vocabularies.[15]

Moreover, the interoperability guideline requires that metadata contains *qualified* references to other metadata. This links both the persistent and unique identifiers described earlier, but also to the relations between them, the foundation of *Linked Data*, a concept introduced by the inventor of the World Wide Web, Tim Berners-Lee. This concept relies heavily on the Resource Description Framework (RDF) specification which allows describing a *graph* linking pieces of information (W3C RDF Working Group 2014). The linked data concept is of utmost importance to the scholarly workflow, as it can provide a wider image over scientific research, as proven by projects such as SciGraph, which defines over one billion relationships between entities such as journals, institutions, funders or clinical trials (Springer Nature 2017), or SCHOLIX[16] which links research data to the outputs that reference it.

Finally, the FAIR principles mandate that research data should be *reusable*, thus allowing for study replicability and reproducibility. For this, it requires that metadata contains accurate and relevant attributes (e.g. it describes the columns of tabular data) and information about its provenance. Moreover, it touches on certain legal aspects, such as the need for clear and accessible licencing and adherence to "*domain-relevant community standards*", such as, for example, the requirements on patient data protection.

4 Legal Aspects of Data Storage

There are a number of legal aspects to consider regarding the storage and sharing of research data; certain elements will differ depending on the geographic location. This section outlines the main points to consider.

[15]Here the principles become *recursive*, mandating that vocabularies describing FAIR datasets should themselves follow the same principles, see FORCE11 (2015).

[16]https://scholix.org.

4.1 Anonymisation of Research Data

Broadly, anonymisation allows data to be shared while preserving privacy. Anonymity is not be confused with confidentiality, although the two are linked. Anonymity is the process of not disclosing the identity of a research participant or the author of a particular view or opinion. Confidentiality is the process of not disclosing to other parties opinions or information gathered in the research process (Clark 2006).

The process of anonymising research data requires that key identifiers are changed or masked. An individual's identity can be disclosed from *direct identifiers* such as names or geographic information or *indirect identifiers* which, when linked with other available data, could identify someone, like occupation or age. Anonymisation should be planned in the early stages of research, or costs can become burdensome later. Anonymisation considerations should be built in when gaining consent for data sharing.

One of the challenges of anonymisation is balance. Going too far could result in important information being missed or incorrect conclusions being drawn, all the while balancing the potential of reidentification. If the research data is for public release, the probability of potential reidentification needs to be low. It may be acceptable for this probability to be higher for private or semi-public as other controls can be put in place (El Emam et al. 2015).

For example, in the USA the Health Insurance Portability and Accountability Act of 1996 (HIPAA) directly addresses anonymisation concerns (U.S. Department of Health and Human Services 2017); it requires that systems and repositories that handle such information need to ensure physical, technical and administrative safeguards that meet the obligations laid out in the Act.

4.2 Legal Frameworks to Consider

As mentioned earlier, the legal frameworks that need to be considered will vary dependent on geography. Three important frameworks to consider are the General Data Protection Regulation (GDPR)[17] in the EU, the UK Data Protection Act 1998/2018[18] and the Patriot ACT in the USA.[19]

The EU General Data Protection Regulation (GDPR) along with the new UK Data Protection Act came into force on May 25, 2018, and governs the processing (holding or using) of personal data in the UK. Although not specifically aimed at research, some changes will still need to be considered. GDPR has a clearer definition of personal data which is that personal data is about living people from which

[17] https://www.eugdpr.org/.

[18] https://www.gov.uk/government/collections/data-protection-act-2018.

[19] https://www.justice.gov/archive/ll/highlights.htm.

they can be identified. As well as data containing obvious *identifiers*, such as name and date of birth, this includes some genetic, biometric and online data if unique to an individual. Data that has been pseudonymised (with identifiers separated), where the dataset and identifiers are held by the same organisation, is still personal data.

The UK Data Protection Act 1998 and its update in 2018 applies in Scotland, England, Wales and Northern Ireland. The Act gives individuals rights of access to request copies of their personal data collected by a researcher. It requires that any processors of personal data must comply with eight principles, which make sure that personal data are:

1. Fairly and lawfully processed
2. Processed for limited purposes
3. Adequate, relevant and not excessive
4. Accurate and up to date
5. Not kept for longer than is necessary
6. Processed in line with your rights
7. Secure
8. Not transferred to other countries without adequate protection

There are exceptions for personal data collected as part of research. It can be retained indefinitely if needed and can be used for other purposes in some circumstances. People should still be informed if the above applies.

Sensitive data also falls under UK Data Protection rules. Sensitive data includes but is not limited to race or ethnic origin, political opinion, religious beliefs or physical or mental health. Sensitive data can only be processed for research purposes if explicit consent (ideally in writing) has been obtained, the data is in substantial public interest and not causing substantial damage and distress, or if the analysis of racial/ethnic origins is for purpose of equal opportunities monitoring.

The legal definition of personal data is complex and is affected by the act's subsequent update in 2018 and GDPR, but the simplest and safest definition is of any information about a living, identifiable individual. This is relevant to anonymisation, as if research data is appropriately anonymised, then the UK Data Protection act will no longer apply. Institutions generally have a Data Protection Officer which should be utilised to address any specific concerns about research outputs.

The PATRIOT Act was signed into law in the USA in 2001. This legislation, again, not specifically aimed at research, has impact when it comes to data storage and grants the potential for the US government to have access to data stored by US cloud servers providers. A common misconception is that avoidance of US-located servers solves the problem, which is only partially accurate. This act would, in theory, allow US judicial authorities and intelligence agencies to request data stored in cloud services outside of the USA. The police, the judiciary and intelligence agencies are able in one way or another to request information from higher education and research institutions and any other parties concerned (van Hoboken et al. 2012).

From a legal perspective, access to cloud data cannot be denied and "cloud service providers can give no guarantees in this respect" (van Hoboken et al. 2012). In practice, access can take place in one of two ways:

- If the cloud service provider is subject to US jurisdiction, a request for release of the data can be made directly to the service provider company in the USA.
- If the cloud service provider is not subject to US jurisdiction, data may be retrieved from the service provider or the institution or with the assistance of relevant local judicial authorities or intelligence agencies.

4.3 Licencing

As is the case with any type of scientific output, research data requires a framework upon which sharing, along with proper attribution, can be achieved. What sets data apart from say, journal papers, is that it can be reused in more ways and such copyright protocols suited for citing research can prove insufficient when considering, for example, the extraction of new hypothesis from existing datasets. This is why new means of licencing and enforcing copyright have either been devised or borrowed from other domains where reuse is common. When data is shared, the original copyright owner usually retains the copyright (UK Data Service 2012), but a licence can be applied in order to describe how the data can be reused. It is important to note that when no proper licencing terms are applied, content cannot be redistributed or reused (Brock 2018).

The Creative Commons (CC) suite of licencing options[20] is one of the most popular for research data; the model consists of a modular set of clauses which can be aggregated for obtaining licences with varying degrees of freedom in terms of reuse. As such, the CC BY licence allows unrestricted reuse as long as attribution is given to the original authors, while more restrictive options such as CC BY-NC-SA or CC BY-NC-ND disallow either using a different licence for derivative work (SA, *share-alike*) or no derivatives (ND) at all, respectively, with a supplementary clause forbidding the usage of the data for any commercial interest (NC, *no commercial*).

Apart from these, other licencing options have been devised for more specific use cases. For example, research software can use one of the deeds commonly employed across the software development ecosystem, such as the MIT licence[21] or the GNU General Public License (GPL).[22] Another example relates to licences developed by national bodies, in order to ensure better compliance with the regional laws and regulations; such instances include the European Union Public License (EUPL)[23] or the Open Government License (OGL).[24] Finally, data can be placed in the public domain, forgoing any copyright or reuse terms; such content can be associated with a notice such as the one provided by Creative Commons as CC0.[25]

[20] https://creativecommons.org/.

[21] https://opensource.org/licenses/MIT.

[22] https://www.gnu.org/licenses/gpl-3.0.en.html.

[23] https://joinup.ec.europa.eu/collection/eupl.

[24] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

[25] https://creativecommons.org/share-your-work/public-domain/cc0/.

While in general researchers should aim for allowing unrestricted use of their data, as also stipulated by the FAIR principles, this is of course not always possible or desirable due to various ethical, regulatory or even technical reasons. In such cases, consulting with personnel specialised in licencing and copyright issues, such as a librarian or lawyer, might be desirable, in order to avoid issues with certain deeds that might place undesirable obstacles on reuse. For example, the no commercial (NC) clause in the CC suite can disallow the use of data not only by commercial corporations but also by research institutions generating minimal amounts of revenue (Klimpel 2013).

4.4 Blockchain: A Technical Solution for Legal Requirements

As the legal requirements around the research area become more complex, technical solutions for alleviating them are currently being researched. The aim of these is to simplify the workflows and maintain a low entry point for users lacking legal expertise while also ensuring the required level of compliance.

In recent years, the blockchain technology has been examined as a potential mean of solving some of these issues, the movement in this direction being fuelled by the increase in interest due to usage in the financial domain (e.g. Bitcoin). At a very high level, the blockchain allows recording data and verifying its authenticity without the need for a central authority. In the most popular implementations, each participant in a blockchain holds a copy of the whole record set, and each record is linked to a previous one by a cryptographic value, distilled using the records' content; thus, even the most trivial change would propagate across the whole chain, signalling the modification.

The basic idea behind blockchains can prove useful in areas where authenticity, provenance and anonymization are important, research being one of them. In Digital Science and van Rossum (2017), the authors have identified a number of ways in which blockchains could be implemented across the scholarly workflow, such as:

- Hypothesis registration: allow researchers to signal a discovery, proof or new dataset while providing evidence of ownership in any future instance.
- Study preregistration: while committing research plans before executing them is already a practice (see, e.g. the "Preregistration Challenge"[26]), it can be difficult to ensure that these plans are not modified while the experiments are ongoing, in order to mask potential discrepancies with the actual results; a blockchain system could easily detect such a change.
- Digital rights management: a blockchain system can easily record ownership, and a related technology, *smart contracts*, a system for defining, actioning and enforcing a set of clauses, can be used to ensure that future usage of data respects and attributes ownership along with any associated stipulations (Panescu and

[26] https://cos.io/prereg/.

Manta 2018). This could prove useful also in terms of counting citations and understanding how data is further reused.
- Data anonymisation and provenance: this can prove to be of utmost importance for medical and pharmacological research, where, on one hand, stringent requirements on patient privacy and data anonymisation are in place, and, on the other hand, the origin of the data should be verifiable. The use of cryptographic controls and the distributed architecture of blockchain systems can help with these challenges.

The application of blockchain technology to various real-world problems is still in its infancy, with many challenges still to overcome. Nevertheless this space is one to closely follow, as it can provide glimpses onto the future of research data sharing.

5 Overview of Research Data Repositories and Tools

What data sharing means is rarely explicitly defined. Questions arise such as:

- How *raw* should the data be?
- Where and how should data be shared?
- How should data be standardised and structured to make it useful?

To complicate matters, these answers will differ dependent on the particular discipline of the person asking the question. Successes in this area like the sharing of DNA sequences via Genbank (National Center for Biotechnology Information 2017) and functional magnetic resonance imaging (fMRI) scans via the Human Connectome Project[27] are research outputs that lend themselves to standardisation. These successes may be difficult to replicate in other disciplines.

Pharmacology presents some particular challenges to data sharing, with common research outputs like raw traces of electrophysiological measurements and large imaging files potentially too unwieldy for existing solutions. For this very reason, The British Journal of Pharmacology (BJP) has a recommended but not mandated policy of data sharing (George et al. 2017).

Look at the wider world of repositories for research data, there are many available options. This section will look at some of the reasons to share research data and how this may influence the discovery path to choosing a repository for publication. It will also look at some of the key features to consider which will have different levels of importance dependent on usage requirements.

There are a number of other considerations to take into account when looking at where to share research data outside of a feature analysis only. There are a variety of certifications to look at, starting with the Data Seal of Approval/CoreTrustSeal. CoreTrustSeal is a non-profit whose stated aim is to promote sustainable and

[27] http://www.humanconnectomeproject.org/.

trustworthy data infrastructures. The 16 questions asked on assessment are an excellent set of questions to think about when deciding where and how to store data (Data Seal of Approval, ICSU World Data System 2016).

Additional certifications to look out for are the Nestor seal[28] and ISO16363[29]; their aim is to provide stricter guidelines concerning the processes data repositories should follow in order to ensure higher guarantees in terms of data and metadata preservation and accessibility.

5.1 Repositories

As alluded to earlier, there are more general repositories for data sharing and numerous very specific repositories. Dependent on the reasons for data sharing, there are numerous methods to choosing where and how to share data. A good starting point are tools such as FAIRsharing[30] from the University of Oxford e-Research centre which has additional resources outside of repositories including data policies and standards, and r3data,[31] an extensive index which can be browsed by subject and country.

Publisher mandates are one of the most common cases for data sharing when publishing an article. Some journals require and others may only recommend that data behind the paper is shared alongside publication. In these instances, the journal will often have a list of approved repositories that they either directly work and integrate with or simply recommend. Examples can be found from Nature (2018, Recommended Repositories) and Elsevier (2018, Link with data repositories).

Funder mandates are an ever-increasing force (Digital Science 2017) in open research data sharing, with some funders opting for a policy that all research data that can be shared must be. In certain cases, funders will have even more stringent requirements, demanding so-called data management plans (Netherlands Organisation for Scientific Research 2015), which detail the whole life cycle of research data, from collection to publication.

Research data sharing in life sciences is commonly scrutinised in terms of adherence to the ALCOA principles, a series of guidelines adopted by a number of large funding and regulatory institutions such as the US Food and Drug Administration, National Institutes of Health or the World Health Organisation. These principles establish a number of properties quality data and, by transitivity, the systems generating, processing and holding it should exhibit:

- Attributable: the generator of the data should be clearly identified.
- Legible: records should remain readable, especially by human users, over time.

[28] http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html.

[29] https://www.iso.org/standard/56510.html.

[30] https://fairsharing.org/.

[31] https://doi.org/10.17616/R3D.

- Contemporaneous: data points should be recorded as near as possible to the event that generated them, thus avoiding any loss of information.
- Original: analyses should be performed using the initially collected data, or a strictly controlled copy, to avoid, for example, errors generated from the transcription of handwritten data to an electronic record.
- Accurate: data should be complete, correct and truthful of the event being recorded.
- Complete[32]: data should not be deleted at any point.
- Consistent: research data should be recorded in a chronological manner.
- Enduring: research data should be preserved for an extended period of time.
- Available: research data should remain accessible for the lifetime of either the final research artefact (e.g. research paper) or physical product (e.g. a drug).

Of course, data may be shared outside of any mandate to do so. This may be to contribute to the increasing success and importance of Open Science movement (Ali-Khan et al. 2018), increase the visibility and reach of the work or to safely ensure the long-term availability of the data or myriad other reasons. In this situation, the wider world of repositories is available. One of the first decisions would be to choose whether to opt for a subject-specific repository or one of the more general repositories available.

Subject-specific repositories have the advantage of potentially having functionality and metadata schemas directly related to your field as well as the advantage of having a targeted audience of potential viewers. Examples of these include:

- Clinical Trials[33] is a database of privately and publicly funded clinical studies conducted around the world.
- The SICAS Medical Image Repository[34] host medical research data and images.
- The TCIA[35] is a service which de-identifies and hosts a large archive of medical images of cancer accessible for public download.

While there are numerous comparisons of generalist data repositories available (see Amorim et al. 2015; Stockholm University Library 2018; Dataverse Project 2017), the rapid pace of development seen on these platforms can mean these are difficult to maintain, and thus repositories should be evaluated at the time of use.

Figshare[36] was launched in January 2011 and hosts millions of outputs from across the research spectrum. It is free to use, upload and access. It has a number of different methods of framing data from single items to themed collections. Figshare

[32]The last four points correspond to a later addition to the ALCOA principles, ALCOA Plus.

[33]https://clinicaltrials.gov.

[34]https://www.smir.ch.

[35]http://www.cancerimagingarchive.net.

[36]https://figshare.com/.

features a range of file visualisation options which may be attractive dependent on the particular files used.

Zenodo[37] was launched in March 2013 is non-profit general repository from OpenAIRE and CERN. It is free to use and access. It is a mature, well-featured product and of particular interest may be the Communities functionality. These curated groups allow for elements of subject-specific repositories to be catered for.

DataDryad[38] was launched in January 2008 and is a non-profit repository which is more tied to the traditional publication process. While all types of data are accepted, they must be linked to a published or to-be-published paper. Outputs on DataDryad are free to download and reuse; uploads incur a submission fee as opposed to the examples above. There are significant integrations with journals, and this option is worth considering if this is of significant importance.

Other options to consider are institutional or national data repositories. Data repositories at institutions are becoming increasingly prevalent and should be investigated as either the sole method of publication or as reference record to the destination of choosing. Examples of these include:

- Cambridge University Data Repository[39]
- Sheffield University Data Repository[40]
- Monash University Data Repository[41]

National and even international repositories are again an area that is still in its infancy but are under active development in many countries around the world. Examples of these include:

- The Norwegian Centre for Research Data[42]
- The UK Data Service[43]
- Swedish National Data Service[44]
- European Open Science Cloud (EOSC)[45]

The repository chosen should be one that works well with FAIR principles outlined previously. While there are many factors to consider that will have different weightings dependent on use, some good general areas to consider include:

[37] https://zenodo.org/.
[38] https://datadryad.org.
[39] https://www.repository.cam.ac.uk.
[40] https://www.sheffield.ac.uk/library/rdm/orda.
[41] https://monash.figshare.com/.
[42] http://www.nsd.uib.no/nsd/english/index.html.
[43] https://www.ukdataservice.ac.uk.
[44] https://snd.gu.se/en.
[45] https://www.eoscpilot.eu/.

- Embargo and access options: does the repository allow the ability to grant different levels of access conditions to the files and/or metadata? Can data be shared privately? Can the files, metadata or both be embargoed?
- Licences: does the repository provide access to the necessary licence options needed for publication of the data? Can you add your own licence?
- Metrics and citations: what type of metrics does the repository track? Do they report the metrics to any tracking bodies? Do they track citations of the data? Do they track nontraditional online attention, e.g. altmetrics?
- Availability: what is the sustainability model of the repository? What guarantees do they provide about the continued availability of the data? Is the data easily accessible programmatically to allow for ease of export?

References

Ali-Khan SE, Jean A, MacDonald E, Gold ER (2018) Defining success in open science. MNI Open Res 2:2. Gates Found Author Manuscript. https://doi.org/10.12688/mniopenres.12780.1

Amorim RC, Castro JA, da Silva JR, Ribeiro C (2015) A comparative study of platforms for research data management: interoperability, metadata capabilities and integration potential. In: Rocha A, Correia A, Costanzo S, Reis L (eds) New contributions in information systems and technologies, Advances in intelligent systems and computing, vol 353. Springer, Cham

Australian Research Data Commons (2018) Australian National Data Service – what we do. https://web.archive.org/web/20180319112156/https://www.ands.org.au/about-us/what-we-do. Accessed 11 Sept 2018

Brickley D, Miller L (2014) FOAF vocabulary specification 0.99. http://xmlns.com/foaf/spec/ https://web.archive.org/web/20180906035208/http://xmlns.com/foaf/spec/. Accessed 6 Sept 2018

Brock J (2018) 'Bronze' open access supersedes green and gold. Nature Index. https://web.archive.org/web/20180313152023/https://www.natureindex.com/news-blog/bronze-open-access-supersedes-green-and-gold. Accessed 13 Sept 2018

CASRAI (2012) CRediT. https://casrai.org/credit/. Accessed 6 Sept 2018

Clark A (2006) Anonymising research data (NCRM working paper). ESRC National Centre for Research Methods. Available via NCRM. http://eprints.ncrm.ac.uk/480/

Clinical Data Interchange Standards Consortium (2018) CDISC standards in the clinical research process. https://web.archive.org/web/20180503030958/https://www.cdisc.org/standards. Accessed 21 Oct 2018

Coughlin T (2016) The costs of storage. Forbes. https://web.archive.org/web/20170126160617/https://www.forbes.com/sites/tomcoughlin/2016/07/24/the-costs-of-storage/. Accessed 19 Aug 2018

Data Seal of Approval, ICSU World Data System (2016) Core trustworthy data repositories requirements. https://web.archive.org/web/20180906181720/https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf. Accessed 6 Sept 2018

Dataverse Project (2017) A comparative review of various data repositories. https://web.archive.org/web/20180828161907/https://dataverse.org/blog/comparative-review-various-data-repositories. Accessed 6 Sept 2018

DDI Alliance (2018) Document, discover and interoperate. https://web.archive.org/web/20180828151047/https://www.ddialliance.org/. Accessed 6 Sept 2018

Digital Science (2017) The state of open data 2017 – a selection of analyses and articles about open data, curated by Figshare. Digital Science, London

Digital Science, van Rossum J (2017) Blockchain for research. Digital Science, London

DOI Foundation (2017) DOI system and the handle system. https://web.archive.org/web/20180112115303/https://www.doi.org/factsheets/DOIHandle.html. Accessed 6 Sept 2018
Eklund A, Thomas EN, Knutsson H (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. PNAS 113:7900. https://doi.org/10.1073/pnas.1602413113
El Emam K, Rodgers S, Malin B (2015) Anonymising and sharing individual patient data. BMJ 350:h1139. https://doi.org/10.1136/bmj.h1139
Elsevier (2018) Sharing research data. https://web.archive.org/web/20180528101029/https://www.elsevier.com/authors/author-services/research-data/. Accessed 6 Sept 2018
European Commission, Directorate-General for Research & Innovation (2017) 2020 programme – guidelines to the rules on open access to scientific publications and open access to research data in horizon 2020, version 3.2. https://web.archive.org/web/20180826235248/http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Accessed 15 July 2018
Federer L, Lu Y-L, Joubert DJ, Welsh J, Brandy B (2015) Biomedical data sharing and reuse: attitudes and practices of clinical and scientific research staff. PLoS One 10:e0129506. https://doi.org/10.1371/journal.pone.0129506
FORCE11 (2015) The FAIR data principles. https://web.archive.org/web/20180831010426/https://www.force11.org/group/fairgroup/fairprinciples. Accessed 6 Sept 2018
FORCE11 (2018) Resource identification initiative. https://web.archive.org/web/20181005133039/https://www.force11.org/group/resource-identification-initiative. Accessed 21 Oct 2018
Freed Innosoft N, Borenstein N (1996) Multipurpose internet mail extensions (MIME) part two: media types. https://web.archive.org/web/20180819062818/https://tools.ietf.org/html/rfc2046. Accessed 19 Aug 2018
George C et al (2017) Updating the guidelines for data transparency in the British Journal of Pharmacology – data sharing and the use of scatter plots instead of bar charts. Br J Pharm 174:2801. https://doi.org/10.1111/bph.13925
Klein A (2017) Backblaze hard drive stats for 2016. https://web.archive.org/web/20180611041208/https://www.backblaze.com/blog/hard-drive-benchmark-stats-2016/. Accessed 19 Aug 2018
Klimpel P (2013) Consequences, risks and side-effects of the license modules "non-commercial use only – NC". https://web.archive.org/web/20180402122740/https://openglam.org/files/2013/01/iRights_CC-NC_Guide_English.pdf. Accessed 6 Sept 2018
National Center for Biotechnology Information (2017) GenBank. https://web.archive.org/web/20180906065340/https://www.ncbi.nlm.nih.gov/genbank/. Accessed 6 Sept 2018
National Institutes of Health (2018) NIH public access policy details. https://web.archive.org/web/20180421191423/https://publicaccess.nih.gov/policy.htm. Accessed 6 Sept 2018
Netherlands Organisation for Scientific Research (2015) Data management protocol. https://web.archive.org/web/20170410095202/http://www.nwo.nl/en/policies/open+science/data+management. Accessed 21 Oct 2018
ORCID (2018) ORCID. https://web.archive.org/web/20180904023120/https://orcid.org/. Accessed 6 Sept 2018
Panescu A-T, Manta V (2018) Smart contracts for research data rights management over the Ethereum blockchain network. Sci Technol Libr 37:235. https://doi.org/10.1080/0194262X.2018.1474838
Phillips N (2017) Online software spots genetic errors in cancer papers. Nature 551:422. https://doi.org/10.1038/nature.2017.23003
RCSB PDB (2018) Protein data bank. https://www.rcsb.org/. Accessed 15 July 2018
Riley J (2017) Understanding metadata – what is metadata, and what is it for? NISO, Baltimore
Sanderson R, Phillips M, Van de Sompel H (2011) Analyzing the persistence of referenced web resources with memento. https://arxiv.org/abs/1105.3459
Schroeder B, Lagisetty R, Merchant A (2016) Flash reliability in production: the expected and the unexpected. In: Proc of 14th USENIX Conference on File and Storage Technology (FAST 16)
Springer Nature (2017) SN SciGraph. https://web.archive.org/web/20180823083626/https://www.springernature.com/gp/researchers/scigraph. Accessed 6 Sept 2018

Springer Nature (2018) Research data support. https://web.archive.org/web/20180309193834/https://www.springernature.com/gp/authors/research-data-policy. Accessed 6 Sept 2018
Stockholm University Library (2018) Data repositories. https://web.archive.org/web/20180828161808/https://www.su.se/english/library/publish/research-data/data-repositories. Accessed 6 Sept 2018
U.S. Department of Health and Human Services (2017) Research. https://web.archive.org/web/20180828205916/https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html. Accessed 6 Sept 2018
UK Data Service (2012) Copyright for data sharing and fair dealing. https://web.archive.org/web/20180828160153/https://www.ukdataservice.ac.uk/manage-data/rights/sharing. Accessed 28 Aug 2018
van Hoboken J, Arnback A, van Ejik NANM (2012) Cloud computing in higher education and research institutions and the USA Patriot Act. SSRN. https://doi.org/10.2139/ssrn.2181534
W3C RDF Working Group (2014) Resource description framework (RDF). https://www.w3.org/RDF/. Accessed 6 Sept 2018
Wilkinson MD et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18
Youngseek K, Zhang P (2015) Understanding data sharing behaviors of STEM researchers: the roles of attitudes, norms, and data repositories. Libr Inf Sci Res 37:189. https://doi.org/10.1016/j.lisr.2015.04.006