Some fundamentals for Open Research Data Management in Health Sciences

Alicia Fátima Gómez Sánchez (a) and Pablo Iriarte (b)
(a) Library and Computing Services, University of Hertfordshire, Hatfield, United Kingdom
(b) Scientific Information Division, University of Geneva, Switzerland

Abstract

The open science movement is increasingly expanding and pushing researchers to embrace new ways of working. Funding agencies, stakeholders, publishers and scientific communities want them to take care of the whole research process, from planning the initial stages of research, to the publication, sharing and archiving of their data. The aim of this article is to present some fundamentals about research data and research data management plans (DMPs), particularly in the biomedical field. Some of the main points related to the publishing of research data, as well as some recommendations for choosing a suitable repository are described. Finally, reasons and advantages for health librarians to be involved in the curation and making research data open and re-usable are set out.

Key words: research data; data management plans; biomedical research; data curation; information dissemination.

Introduction

The interest in research data and research data management in the context of open science has dramatically increased in the last years. Specially funders (1-3) but also publishers, have already implemented data sharing policies (4-6), with the aim to make science more transparent and reproducible. The first step in this road was the introduction and development of Data Management Plans (DMPs), required by the EU on projects financed under the H2020 program (7) and quickly used by a lot of national funding agencies as a must have criteria for the new projects. In fact, the NIH has officially supported the concept of data sharing as an essential issue for the translation of research results into knowledge, products and procedures to improve human health since 2003 (8).
To achieve openness and transparency, research data must be not only open accessible, but also discoverable and reusable. Data need to be described using appropriate metadata, which can be defined as the structured information about data following the right standards, and deposited in trustworthy repositories that assure access and preservation (9).
The aim of this article is to describe some of the main characteristics of research data, especially in the biomedical field, and to provide an overview about how librarians could help researchers to manage research data in the context of open science.

Research data, metadata and data management plans

Research data can be a wide diversity of collected information: textual or numerical data, samples, notebooks, images, questionnaires, recorded audios or videos, models, software, reports, procedures, workflows, and many more. Formats can also vary: text files, software, websites, images, etc.
All information about the type and the format of the information needs to be described. In addition, data need to be complemented by proper metadata. Metadata describe the data, and are essential to recover and reuse research data. Moreover, there are metadata standards that allow the interoperability across systems. Metadata can be classified in 3 main types (10): descriptive, administrative, and structural:

- descriptive metadata serves to discovery and understand a resource, and refers for example to the title, author, publication date or abstract. The main standard for this is the Dublin Core Schema, which is a small set of vocabulary terms

Address for correspondence: Alicia Fátima Gómez Sánchez, Library and Computing Services, University of Hertfordshire, Hatfield, United Kingdom. E-mail: a.gomez-sanchez@herts.ac.uk.

that can be used to describe resources (11);
- for librarians supporting data management the two main types of administrative metadata are related to intellectual property, and preservation metadata. The most adopted standard for these metadata is the PREMIS Data Dictionary and all its supporting documentation, which was developed by the Preservation Metadata Implementation Strategies (PREMIS) international working group, established by the Online Computer Library Center (OCLC) and Research Libraries Group (RLG) (12);
- structural metadata help to describe the relationship between the different parts of the resources; they are important for navigation, and an example can be the sequence or the place in the hierarchy (10).

Data Management Plans (DMPs) are documents or web forms describing the data management life cycle for the data to be collected, processed and/or generated by a project, and serve to make research data findable and re-usable (7). Research data management has to be considered in the context of the research data lifecycle, including identifying, cleaning, describing, storing and preserving or sharing data (13). Support for some of these stages can be offered by biomedical and other specialized libraries, especially in the development of metadata and data standards. Furthermore, data management plans for grant applications include the description of the data, the utility, information about how to make data findable – again, through the provision of metadata – and making them openly available via deposit in open repositories. All this DMP information could be better described if an information specialist is involved in the research process.

Preparing biomedical research data to be shared

Documentation and licenses

The metadata included in the DMP is necessary but not sufficient. In order to complete the picture and add the context to the research data we need to add some material that explains how data has been created, what they mean, how their structure is, and which alterations and manipulations have been done to clean and analyse the data. "Creating this comprehensive documentation is very important because it transfers knowledge about your data to other potential users enabling researchers to discover, understand, and properly cite your data. It provides the context to the data and ensures re-use and comprehension in the long term" (14).

There are different descriptive metadata standards, used to particular needs or disciplines (15, 16). By applying a metadata standard recognized by your discipline, you can help others discover, comprehend, and evaluate data across time and distance without having to access the data itself. However the choice of the right metadata standard is not easy and often it is imposed by the repository or data archive where we publish the data. Two standards are widely used: DDI and DataCite. In the biomedical field the Minimum Information for Biological and Biomedical Investigations becomes largely used.

A "readme" file could be added to give more information about a data file and help the data to be correctly interpreted. It is very useful for the author himself (it is always difficult to understand in the future the data and the code we have applied) or by other researchers when sharing or publishing data (17).

When the research data has a DMP, is well documented, has the files converted to an open format, anonymized and clean, then it is ready to be shared in a repository, after having chosen the right license for the data publication and reuse. For example you can use one of the less restrictive Creative Commons licenses like CC0, CC-BY, CC-BY-ND or CC-BY-SA. If you decide to publish your data or database as open data then one of the Open Data Commons Licences must be used, like the Public Domain Dedication and License (PDDL), the Attribution License (ODC-By) or the Open Database License (ODC-ODbL).

Types and formats

Because of their diversity and complexity of biomedical research (fundamental, preclinical, clinical, imagery, OMICS, laboratory, nursing, public health, etc.), it is difficult to make an exhaustive list of biomedical research data types and formats. Regarding the format of research data, the problem is intimately linked to their perpetuation, their transmission and their quality. It is therefore encouraged to use non-proprietary formats, which will not depend on a software or company, but which can be read as much as possible. As for quality, the question is important for data in the form of media files (sound, image and video), since it is not uncommon to sacrifice part of the quality, and therefore information due to compression, in order to reduce the weight of the file. Regarding this point, the choice is always a matter of compromise according to the needs and the capacities of the services.

Anonymisation

In order to be accessible and interoperable, research data must be cleaned, anonymised and published in a repository. In many cases, data produced by biomedical research relates to humans and is therefore subject to strict data protection rules and laws. In addition, in most Western countries health information of individuals is considered sensitive data and must therefore be particularly protected (18).

The sharing of patient data requires the agreement of the person concerned. This agreement can be translated into three levels of consent from the patient allowing the use of his or her personal data (19):

- broad consent: data might be shared after use;
- middle consent: participants were told that their data might be shared with people working in specific research areas related to the study;
- explicit consent: participants would be contacted for an opinion whenever there was a request for sharing.

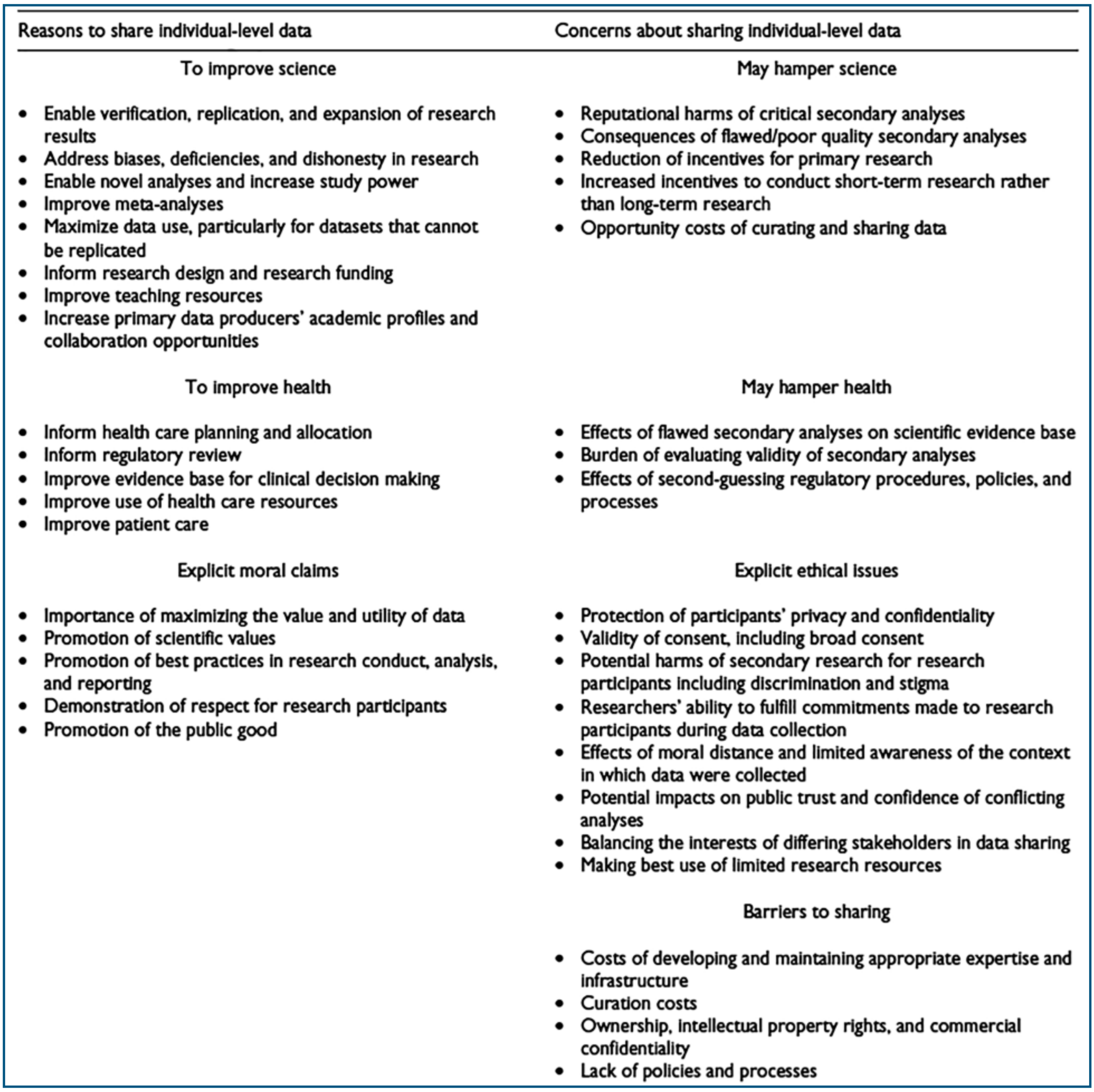
This characteristic of biomedical research remains the most important obstacle to data sharing, which can only be done on very strict rules governed by contracts between the research teams. However, anonymization and statistical disclosure control techniques have been developed from many years (20). Today, there is a software allowing to remove direct identifiers (names, email, date of birth, social security number, address, etc.) and recode indirect identifiers (information that can make it possible to identify the person by crossing the data with other public datasets, such as the dates of entry and discharge from hospital, dates of delivery, etc.) and other sensitive information in order to obtain a good balance between anonymization and loss of information (21-23). Thanks to a precise data analysis, cleaning and anonymization work, we can convert medical data that seemed impossible to share, into anonymous sets, shareable on a data repository publicly or on request (*Table 1*).

Choosing an appropriate repository

After the description and the preparation of the data, the next important step is the election of a trustworthy repository to archive and preserve the data, that may be general or limited to datasets. Of course, institutional repositories should be considered, but there are many other options that can be used to archive datasets as Zenodo, Figshare, Dryad or another data repository cited in the re3data.org registry.

Talking about Health Sciences, the must be underlined that some fields, as for Genomics or Proteomics, where data have their own structures and databases and have been storing open data for many years, particularly in the OMICS, public health or clinical trials. Some examples of very well established archives and knowledge databases are Genbank (the NIH genetic sequence database), Gene Ontology, Pfam (for protein families), UniProt (Universal Protein Resource), the European Nucleotide Archive (ENA), HealthData.gov, or the datasets included in the International Clinical Trials Registry Platform (ICTRP) of the WHO, among many others. Researchers working for instance in the fields of the OMICS are aware that sometimes there is even a requirement for some journals to store data regarding an article in the related archives. In addition, archiving datasets in specific subject repositories can improve the visibility of the research and increase the number of citations or downloads.

Health librarians should be able to recognise the most accurate repositories to give the best advice. In addition, information specialists should have some knowledge about the main certifications or audit tools for trustworthy repositories, as the Trustworthy Repositories Audit & Certification (TRAC), DRAMBORA (Digital Repository Audit Method based on risk assessment), the Nestor Catalogue of Criteria for Trusted Digital Repositories, the Data Seal of Approval, or the ISO 16363 Audit and Certification of Trustworthy Digital Repositories. This kind of certifications can assure the preservation and accessibility of data over time.



Reasons to share individual-level data	Concerns about sharing individual-level data
To improve science	May hamper science
• Enable verification, replication, and expansion of research results • Address biases, deficiencies, and dishonesty in research • Enable novel analyses and increase study power • Improve meta-analyses • Maximize data use, particularly for datasets that cannot be replicated • Inform research design and research funding • Improve teaching resources • Increase primary data producers' academic profiles and collaboration opportunities	• Reputational harms of critical secondary analyses • Consequences of flawed/poor quality secondary analyses • Reduction of incentives for primary research • Increased incentives to conduct short-term research rather than long-term research • Opportunity costs of curating and sharing data
To improve health	May hamper health
• Inform health care planning and allocation • Inform regulatory review • Improve evidence base for clinical decision making • Improve use of health care resources • Improve patient care	• Effects of flawed secondary analyses on scientific evidence base • Burden of evaluating validity of secondary analyses • Effects of second-guessing regulatory procedures, policies, and processes
Explicit moral claims	Explicit ethical issues
• Importance of maximizing the value and utility of data • Promotion of scientific values • Promotion of best practices in research conduct, analysis, and reporting • Demonstration of respect for research participants • Promotion of the public good	• Protection of participants' privacy and confidentiality • Validity of consent, including broad consent • Potential harms of secondary research for research participants including discrimination and stigma • Researchers' ability to fulfill commitments made to research participants during data collection • Effects of moral distance and limited awareness of the context in which data were collected • Potential impacts on public trust and confidence of conflicting analyses • Balancing the interests of differing stakeholders in data sharing • Making best use of limited research resources
	Barriers to sharing
	• Costs of developing and maintaining appropriate expertise and infrastructure • Curation costs • Ownership, intellectual property rights, and commercial confidentiality • Lack of policies and processes

Table 1. *Summarizing the benefits and concerns of biomedical data sharing (24).*

Conclusions

Research data has become the new "fuel" of science and the biomedical field is not an exception. Funder's or institutional mandates are one of the reasons to make data openly available, but more important is to make science transparent and reproducible. A good description of the data and the setting up of good metadata is essential to recover information in databases, and information specialists can help on their descriptions as they are aware of describing and organizing information.

Besides the different nature and formats of research data, there are also some particularities in some Health Sciences fields that should be underlined, as the importance of confidentiality or the existence of subject specific repositories, that health librarians should recognize to help researchers make the most of their data.

Biomedical librarians have to invest this new field and work on a good collaboration and integration in the research process from the beginning, to ensure that the data are compliant with the FAIR principles: findable, accessible, interoperable and re-usable. Finally, datasets should be considered as research output in addition to research publications, following some of the responsible metrics recommendations by the San Francisco Declaration on Research Assessment (DORA) or the Leiden Manifesto (25).

Acknowledgements
AFGS is supported by the University of Hertfordshire, and PI by the University of Geneva. The authors thank Bill Worthington (University of Hertfordshire) for English editing.

Received on 17 May 2017.
Accepted on 22 May 2017.

REFERENCES

1. NIH. National institutes of health genomic data sharing policy [Internet]. Available from: https://gds.nih.gov/03policy2.html
2. European Commission. Open access & Data management - H2020 Online Manual [Internet]. Available from: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm
3. Research Councils UK. RCUK Common Principles on Data Policy [Internet]. 2011. Available from: http://www.rcuk.ac.uk/research/datapolicy/
4. Pool R. Dare to share? Research Information [Internet]. 2016;(December 2016/January 2017). Available from: https://www.researchinformation.info/feature/dare-share
5. PLOS ONE. Data Availability [Internet]. Available from: http://journals.plos.org/plosone/s/data-availability
6. Nature. Data Policies [Internet]. Available from: https://www.nature.com/sdata/policies/data-policies
7. European Commission. Data management - H2020 Online Manual [Internet]. Available from: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
8. NIH. Final NIH Statement on Sharing Research Data [Internet]. 2003. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html
9. Huber V. Développement d'une offre de formation sur la gestion des données de la recherche en médecine et santé publique [Internet] [Bachelor's thesis]. Haute École de Gestion de Genève; 2016. Available from: http://doc.rero.ch/record/278064
10. Riley J. Understanding Metadata: What is metadata, and what is it for? [Internet]. National Information Standards Organization (NISO); 2017. Available from: http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf
11. Dublin Core Metadata Initiative. DCMI Metadata Terms. A one-stop source of up-to-date information on DCMI metadata terms [Internet]. Available from: http://dublincore.org/documents/dcmi-terms/
12. US Library of Congress. PREMIS: Preservation Metadata Maintenance Activity [Internet]. Available from: http://www.loc.gov/standards/premis/
13. Goben A, Raszewski R. The data life cycle applied to our own data. Journal of the Medical Library Association: JMLA [Internet]. January 2015;103(1):40-4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4279933/
14. Consortium of European Social Sciences Data Archives. CESSDA User Guide for data management: Documentation and metadata [Internet]. 2015. Available from: https://cessda.net/content/download/241/2391/file/CESSDA%20User%20Guide%20for%20data%20management_4_Documentation%20and%20metadata.pdf
15. Digital Curation Centre (DCC). List of Metadata Standards [Internet]. Available from: http://www.dcc.ac.uk/resources/metadata-standards/list
16. Research Data Alliance. RDA Metadata Standards Directory [Internet]. Available from: http://rd-alliance.github.io/metadata-directory/standards/

17. Research Data Management Service Group. Cornell University. Guide to writing readme style metadata [Internet]. Available from: https://data.research.cornell.edu/content/readme
18. European Commission. Data protection in the EU [Internet]. Public Health. Available from: https://ec.europa.eu/health/data_collection/data_protection/in_eu_en
19. Hate K, Meherally S, Shah More N, Jayaraman A, Bull S, Parker M, et al. Sweat, Skepticism, and Uncharted Territory. Journal of Empirical Research on Human Research Ethics [Internet]. July 2015;10(3):239-50. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4547203/
20. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, *et al*. Statistical Disclosure Control. 1st edition. Chichester, West Sussex, United Kingdom: Wiley; 2012.
21. Statistical Disclosure Control (SDCMicro) [Internet]. Available from: http://www.ihsn.org/home/software/disclosure-control-toolbox
22. μ-ARGUS [Internet]. Available from: http://neon.vb.cbs.nl/casc/..%5Ccasc%5Cmu.htm
23. ARX - Powerful Data Anonymization A comprehensive software for risk- and utility-based privacy-preserving microdata publishing [Internet]. Available from: http://arx.deidentifier.org/
24. Bull S, Roberts N, Parker M. Views of ethical best practices in sharing individual-level data from medical and public health research. Journal of Empirical Research on Human Research Ethics [Internet]. August 2015; Available from: http://dx.doi.org/10.1177/1556264615594767
25. Hicks D, Wouters P, Waltman L, Rijcke S de, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature News [Internet]. April 2015;520(7548):429. Available from: http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351