




OPINION ARTICLE

Long-term preservation of biomedical research data [version 1; referees: 4 approved, 1 approved with reservations]

Vivek Navale , Matthew McAuliffe

Center for Information Technology, National Institutes of Health, Bethesda, Maryland, 20892, USA

v1 First published: 29 Aug 2018, 7:1353 (doi: [10.12688/f1000research.16015.1](https://doi.org/10.12688/f1000research.16015.1))
 Latest published: 29 Aug 2018, 7:1353 (doi: [10.12688/f1000research.16015.1](https://doi.org/10.12688/f1000research.16015.1))

Abstract

Genomics and molecular imaging, along with clinical and translational research have transformed biomedical science into a data-intensive scientific endeavor. For researchers to benefit from Big Data sets, developing long-term biomedical digital data preservation strategy is very important. In this opinion article, we discuss specific actions that researchers and institutions can take to make research data a continued resource even after research projects have reached the end of their lifecycle. The actions involve utilizing an Open Archival Information System model comprised of six functional entities: Ingest, Access, Data Management, Archival Storage, Administration and Preservation Planning.

We believe that involvement of data stewards early in the digital data life-cycle management process can significantly contribute towards long term preservation of biomedical data. Developing data collection strategies consistent with institutional policies, and encouraging the use of common data elements in clinical research, patient registries and other human subject research can be advantageous for data sharing and integration purposes. Specifically, data stewards at the onset of research program should engage with established repositories and curators to develop data sustainability plans for research data. Placing equal importance on the requirements for initial activities (e.g., collection, processing, storage) with subsequent activities (data analysis, sharing) can improve data quality, provide traceability and support reproducibility. Preparing and tracking data provenance, using common data elements and biomedical ontologies are important for standardizing the data description, making the interpretation and reuse of data easier.

The Big Data biomedical community requires scalable platform that can support the diversity and complexity of data ingest modes (e.g. machine, software or human entry modes). Secure virtual workspaces to integrate and manipulate data, with shared software programs (e.g., bioinformatics tools), can facilitate the FAIR (Findable, Accessible, Interoperable and Reusable) use of data for near- and long-term research needs.





Keywords

Open, Archival, Information, System, Biomedical, Data, Preservation, Access

Open Peer Review

Referee Status:

	Invited Referees				
	1	2	3	4	5
version 1					
published 29 Aug 2018	report	report	report	report	report

- 1 **Jane Greenberg** , Drexel University, USA
- 2 **George Alter** , University of Michigan, USA
- 3 **Chaitan Baru**, University of California San Diego, USA
- 4 **David Giarretta** , Primary Trustworthy Digital Repository Authorisation Body Ltd (PTAB Ltd), UK
- 5 **Ravi Madduri** , University of Chicago, USA

Discuss this article

Comments (0)

Corresponding author: Vivek Navale (Vivek.Navale@nih.gov)

Author roles: Navale V: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; McAuliffe M: Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2018 Navale V and McAuliffe M. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

How to cite this article: Navale V and McAuliffe M. **Long-term preservation of biomedical research data [version 1; referees: 4 approved, 1 approved with reservations]** *F1000Research* 2018, 7:1353 (doi: [10.12688/f1000research.16015.1](https://doi.org/10.12688/f1000research.16015.1))

First published: 29 Aug 2018, 7:1353 (doi: [10.12688/f1000research.16015.1](https://doi.org/10.12688/f1000research.16015.1))

Introduction

Over the past decade, major advancements in the speed and resolution of acquiring data has resulted in a new paradigm, ‘Big Data.’ The impact of Big Data can be seen in the biomedical field. Billions of DNA sequences and large amounts of data generated from electronic health records (EHRs) are produced each day. Continued improvements in technology will further lower the cost of acquiring data, and by 2025, the amount of genomics data alone will be astronomical in scale¹. In addition to large data sets and the large number of data sources, challenges arise from the diversity, complexity and multimodal nature of data generated by researchers, hospitals, and mobile devices around the world. Research programs like the All of Us Research Program envision using Big Data to transform healthcare from case-based studies to large-scale data-driven precision medicine endeavors².

Harnessing the power of digital data for science and society, requires developing management strategies that enable data to be accessible and reusable for immediate and future research needs. With the preponderance of bigger datasets, the volume, variety and magnitude of biomedical data generation is significantly higher than existing analytical capabilities. The time lag between data accumulation and thorough analysis will result in more data being passive or inactive for extended time intervals. Meaningful associations for data reuse for applications beyond the

purpose for which it was collected will also be a time-intensive endeavor. Therefore, our opinion is that attention should be focused on developing a data preservation strategy that can ensure biomedical data availability for longer term access and reuse.

Model for long-term data preservation

Challenges to manage vast amount of data from space missions led to the development of the Open Archival Information System (OAIS) model³. The OAIS model defined “as an archive that consists of an organization of people and systems with responsibility to preserve information and make it available for a designated community” provides the framework for long term preservation of data⁴.

The functional model (Figure 1) illustrates that during the Ingest process, Submission Information Packages (SIP) are produced. Metadata and descriptive information are important for developing Archival Information Packages (AIP) for data storage. Metadata can include attributes that establish data provenance, authenticity, accuracy, and access rights. The Dissemination Information Packages (DIP) are produced in response to queries from consumers. The OAIS model includes six functions (shown in Figure 1) - Ingest, Access, Data Management, Archival Storage, Administration and Preservation Planning. Figure 1 contains a pictorial representation of the OAIS model.

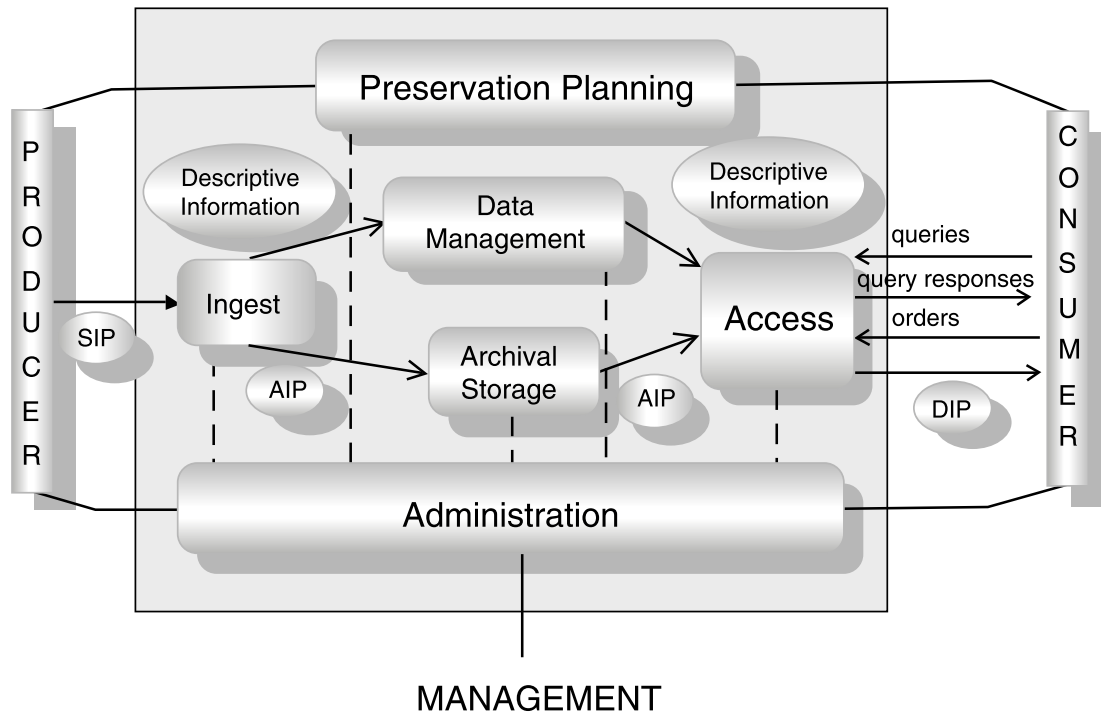


Figure 1. Open Archival Information System (OAIS) functional model. Information flow within the OAIS model is by means of “packages”, SIP, AIP and DIP with the related interfaces (both solid and dotted lines) that show the interaction between the various functions⁵. Various OAIS implementations have led to development of digital repository systems (e.g. Dspace, Fedora) and customized repositories (e.g. the US National Oceanic and Atmospheric Association). Reproduced with permission from The Consultative committee for Space Data Systems (<https://public.ccsds.org/pubs/650x0m2.pdf>). The source for this OAIS implementation was originally provided by Ball (2006) (<http://www.ukoln.ac.uk/projects/grand-challenge/papers/oaisBriefing.pdf>)⁶.

The wide variety of examples illustrate that the OAIS model is content and technology agnostic. Therefore, we posit that the model can be used for developing biomedical digital data preservation strategy. In the following sections, we contextualize the functional aspects of the model needed for successful implementation of biomedical data repository ecosystems.

Preservation planning

As shown in [Figure 1](#), preservation planning is an important bridge between the data producers and consumers. During the planning stage several questions (some of which are listed below) must be addressed:

- How will data be collected and managed?
- What data (and metadata) is required for establishing provenance?
- What type of common data elements and bio-ontology are needed?
- How will data curation be carried out for the data sets?
- Which data types will be stored and preserved?
- How will data access be provided?
- What methods are needed to maintain data quality?

In the past these questions have been the responsibility of biomedical data custodians and curators working in libraries, archives and repositories, who are usually engaged during the latter part of data lifecycle management (during data preservation and access services). We think that importance should be placed on data preservation during the planning of initial activities (e.g. collection, processing, storage), along with the ensuing activities (data analysis, sharing and reuse). In our opinion, developing a community of data stewards for biomedical research programs within institutions is an important step towards long-term preservation of biomedical data.

Considering the interdisciplinary skill set needed for data stewards, we propose that institutions leverage the expertise of staff (e.g. biologists, physicians, informaticians, technologists, library science specialists, etc.) for their respective biomedical research programs. We envision data steward teams to be engaged early in research data lifecycle management, developing digital data stewardship plan(s) for biomedical data sets. These activities can promote a culture of semantic scientists for biomedical programs, which can help reduce the time and cost of data interpretation by biocurators⁷.

We think that establishing data stewards' responsibility within the biomedical research program can improve data quality, provide traceability, and support reproducibility.

Typically, a designated community for biomedical data are researchers of a sub-discipline for a disease (e.g. cancer). Reviewing the research sponsors' requirements, understanding the volume and types of data to be collected, and defining how the data will be organized and managed can all promote the reuse of data⁸.

Goodman *et al.* provide a short guide to consider when caring for scientific data. The guide highlights the use of permanent identifiers, depositing data in established repository archives and publishing code and workflows that can facilitate data use/reuse⁹.

It is also essential that research group leaders and institutions emphasize data management best practice principles¹⁰. An important practice for ensuring good research management in laboratories includes selecting the right medium (paper-based and/or electronic) for laboratory notebooks¹¹.

Administration

Both producers and consumers of data will be best served by implementing established procedures for digital preservation. Producers of biomedical data should develop a comprehensive data management plan (DMP) that addresses policies and practices needed to acquire, control, protect and deliver data, and the steps needed for the preservation and reuse of data¹².

As a first step, data stewards should establish a DMP to identify the types of data that will be collected, provide information on the organization of data, assign roles and responsibilities for description of the data and document processes and procedures for Ingest and methods for data preservation and dissemination.

Data collection strategies need to be established in context of institutional policies for biomedical archives. We recommend that DMP be used as a planning tool to communicate all operations performed on data, and details of software used to manage data. Williams *et al.* provide a comprehensive review of data management plans, their use in various fields of biomedical research, and reference material for data managers¹³.

As part of administration, data stewards should engage with a designated community (data creators, funding agencies, stakeholders, records managers, archivists, information technology specialists) to appraise the data to determine whether all the data produced during the research program should be preserved, or whether different data types (raw, processed, etc.) require different degrees of preservation (e.g. temporary with a time stamp for review or permanent indefinitely).

Establishing data provenance should be part of data collection and management strategy. This may not be always easy, because contextual information (metadata) about experimental data (wet/dry lab) and workflows is often captured informally in multiple locations, and details of the experimental process are not extensively discussed in publications. Contacting the original source for additional information may or may not yield fruitful results, and the reproducibility of experiments becomes challenging in many cases.

Security controls should be part of data collection and management strategy. For initial security controls assessment, guidance documents, FISMA, NIST-800-53 and FIPS, can provide tools for an organizational risk assessment and validation purposes^{14,15}. A wide range of issues involving ethical, legal and technical boundaries influence biomedical data privacy, which

can be specialized for the type of data being processed and supported¹⁶. Important points to consider are confidentiality, disclosure specifications, data rights ownership, and eligibility criteria to deposit data to an established repository.

Ingest

Capturing relevant data from the experiment in real time can be one of the better practices for establishing biomedical data provenance. Automated metadata capture when possible (using a laboratory information management system), and digitization where automation is not possible, can reduce errors, minimize additional work and ensure data and metadata integrity¹⁷. We believe that establishing data provenance will result in successful preparation of SIP during ingest (Figure 1).

SIP for clinical research, patient registries and other human subject research can be developed by use of common data elements (CDEs). A CDE is defined as a fixed representation of a variable to be collected within a clinical domain. It consists of a precisely defined question with a specified format or a set of permissible values for responses that can be interpreted unambiguously in human and machine-computable terms. There are many examples of CDE usage and information on CDE collections, repositories, tools and resources available from the National Institutes of Health (NIH) CDE Resource Portal¹⁸. The advantage of using CDEs was highlighted by the Global Rare Disease Repository (GRDR), where researchers integrated de-identified patient clinical data from rare disease registries and other data sources to conduct biomedical studies and clinical trials within and across diseases¹⁹.

Ontologies are useful for annotating and standardizing the data description so that the querying and interpretation of data can be facilitated. Selecting a biontology requires knowledge about the specific domain, including current understanding of biological systems. Several ontologies have been reported for various biological data and can be selected for research data²⁰. An online collaborative tool (e.g. [OntoBrowser](#)) can be used to map reported terms to preferred ontology (code list), which can be useful for data integration purposes²¹. We believe that use of CDEs and Ontologies can result in developing AIP for long-term preservation of biomedical data.

Data management

Data authenticity, accuracy and reliability influence data quality. For that purpose, controls from the very beginning of research (as part of DMP) need to be established. For experimental work, instrument calibration and validation of data analysis methods contributes significantly to the quality of data produced in a lab. Currently, many approaches for data quality assessment exist and their strengths and weakness have been discussed²². The most common approach for obtaining a first look at the quality of new data is by reviewing supporting data provided with research articles that contextualizes data to support the research goals and conclusions. Additional quality assessment is obtained by the evaluation provided by data producers and data curators and, when appropriate, with automated processes and algorithms.

In the context of Electronic Health Records (EHR), the five dimensions of data quality are: completeness, correctness, concordance, plausibility and currency. The data quality assessment of these dimensions has been carried out by one or more of the seven categories: comparison with gold standards, data element agreement, data source agreement, distribution comparison, validity checks, log review, and element presence²³. Validated and systematic methods for EHR data assessment are important, and with shared best practices, the reuse of EHR data for clinical research can be promoted.

We think data curation needs should be assessed during DMP. One of the ways to assess data curation is by using a Data Curation Maturity model²⁴. The model assumes that new areas of research (evolving areas) may not have best practice(s) from the very beginning, but having an indicator to show maturity levels at different stages of an organization or group in performing tasks enables in improving curation practices. A staging approach is proposed to aid in developing good practices (and even best practices) and identify ineffective practices for various tasks so that the quality of data can be improved upon from the beginning. The maturity model can be useful for determining steps that are needed to improve data quality.

We opine that data stewards should engage with established repositories and develop data sustainability plans. Many well-known biomedical repositories are known to host wide ranges of biomedical data (for example, [GenBank](#) for nucleotide sequence, [Gene Expression Omnibus](#) for microarray and high throughput gene expression data, [miRbase](#) for annotated sequences, [dbSNP](#) for single nucleotide polymorphism (SNPs), [Protein Data Bank](#) for 3D structure data for macromolecules (proteins and nucleic acids), [RefSeq](#) for non-redundant DNA, RNA and protein sequences. Additionally, disease-specific repositories for traumatic brain injury and Parkinson's disease are also available²⁵.

A tabulated listing of 21 established life science repositories with various types of user support services (e.g. for visualization, data search, analysis, deposition, downloads, and online help) is also available²⁶. Additional helpful resources, [re3data.org](#): the Registry of Research Data Repositories, can be used to identify appropriate repositories for storage and search of research data²⁷.

Archival storage

Both raw and processed data are produced during biomedical research. Therefore, developing a storage roadmap is important and should consider data types, volume, data format and the applications required for current and future processing. Broadly, file, object and block are three types of data storage options available to biomedical researchers²⁸.

File Storage has been used for storing large and smaller scale biomedical datasets providing direct and rapid access to local computing resources, including High Performance Computing clusters. Object Storage is ideal for systems that need to grow and scale for capacity. Block Storage is useful when the software application needs to tightly control the structure of the data, usually the case with databases.

Depending on access needs a tiered data storage strategy can be used for migrating data from high input/output (I/O) disks to lower I/O media, like magnetic tapes. Data Storage strategy should consider at least two types of media (disks and tapes) to mitigate the probability of data loss due to media failure. In addition, primary and backup copy of data should be stored at two different geographically separated locations (at least several hundred miles apart).

Considering the diversity, complexity and increasing volume of biomedical research data, we posit that cloud based platforms can be leveraged to support varieties of ingest modes (e.g. machine, software or human entry modes) to make data findable, accessible, interoperable and reusable (FAIR)²⁹. In our opinion, a cloud-based data archive platform (shown in Figure 2) can provide a dynamic environment for managing research data life cycle along with capabilities for long-term preservation of biomedical data³⁰.

Access

Needs in biomedical research can vary from simple queries (as shown in Figure 1) to a wide range of capabilities (work-flow and software tools) usually employed for analysis of large scale data sets (genomics data)³¹. Dissemination information package (DIP) can result from discovery search engines, e.g. DataMed³², and machine readable methods (e.g. repositio.io) for extracting new knowledge from the datasets³³, with online available resources for digital sharing purposes³⁴.

Broadly speaking, access to data and metadata can be discussed in terms of the web and Application Programming Interface

(API). In the web mode (Figure 2), user utilizes an interactive “browser” that presents overviews, summaries, and familiar search capabilities. In the API mode, the same underlying data and metadata can be consumed by a computer. The API mode is composed of a set of protocols and instructions that can serve the needs of both software developers and users. APIs commonly use Representational State Transfer (REST)³⁵. REST utilizes the standard ‘http’ protocol access to manipulate data or metadata, and standards and toolsets for developing, documenting, and maintaining REST-based APIs are available³⁶. In our opinion, type of API adoption will be driven by research questions and user community needs, evident from the comparison of three Genomics API’s (Google Genomics, SMART Genomics, and 23andMe)^{37,38}.

For longer term access needs using file formats that have a good chance of being understood in the future is one of the ways of overcoming technology obsolescence. File formats that have characteristics of “openness”, “portability”, and maintain “quality” are better choices for long term preservation needs.

Information on the data types (e.g. text, image, video, audio, numerical), structure and format are essential for ensuring that it can be used and reused over time. Access to data will be greatly enhanced if data are archived in “open formats” not restricted by proprietary software, hardware, and/or by purchase of a commercial license. Some examples of open data formats in use are: comma or tab-separated values (csv or tsv) for tabular data, hierarchical data format and NetCDF for structured scientific data, portable network graphics for images, Open Geospatial Consortium format for geospatial data, and

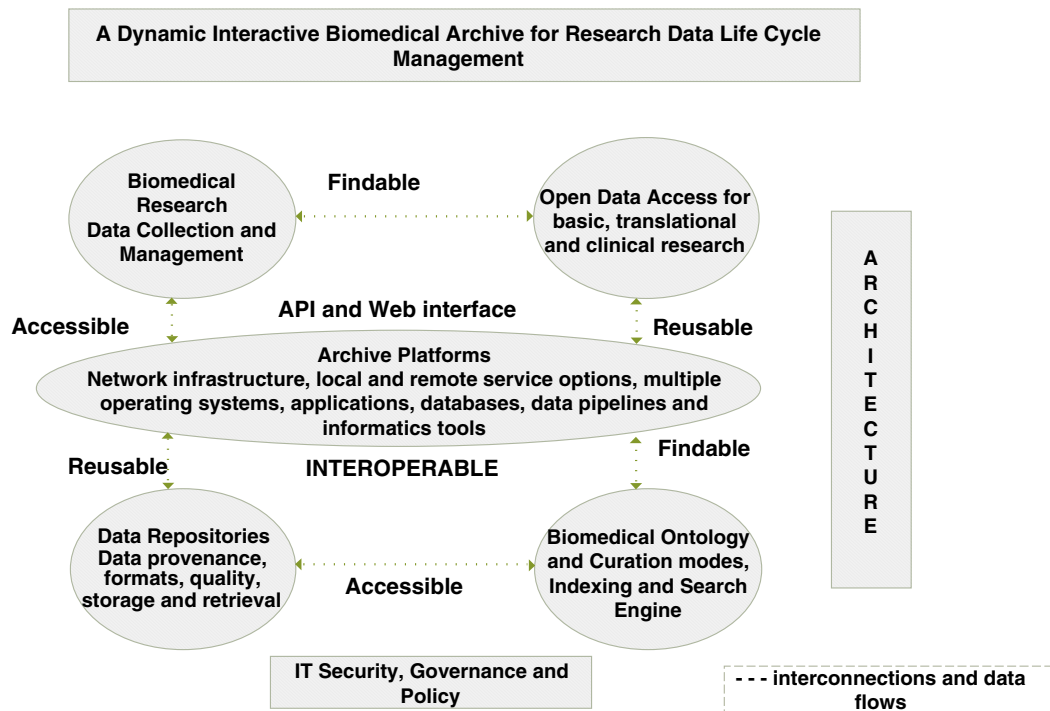


Figure 2. Conceptual model of bio-archive platform powered by cloud resources for long-term preservation of biomedical needs. Figure adapted from Navale and Bourne (2018)³⁰.

extensible markup language for documents³⁹. If proprietary formats are used for initial data collection and analysis work, it should be exported to an open format for archival purposes. In some cases, proprietary formats have become standard formats when popularity and utility have driven tools and algorithms purpose-built to ingest and modify those formats (e.g., Affymetrix .CEL and .CDF formats).

We also think that the reuse of preserved data can be enhanced by the open availability of client software to user communities. One example is Bioconductor for genomic data⁴⁰. In addition, developing and applying ontology-driven transformation and integration processes can result in open biomedical repositories in semantic web formats⁴¹.

Conclusion

Valuing, protecting, enabling access, and preserving data resources for current and future needs of researchers, laboratories, institutes and citizens is a critical step in maturing the biomedical research process of any organization or community.

With advent of Big Data, biomedical researchers need to become more proficient in understanding and managing research data throughout its lifecycle. Establishing the responsibilities of data stewards within the biomedical research program can improve data quality, provide traceability and support reproducibility. Determining specifically what to preserve and for how long

are policy decisions that require data steward teams to engage with funding agencies, designated communities and established repositories.

We opine that the likelihood of maintaining the authenticity, accuracy and reliability of biomedical data for longer-term access will be enhanced by application of the OAIS model. Implementation of the model for biomedical data sets will provide renewed opportunities for data integration, analysis and discovery for basic, translational and clinical research domains.

Data availability

No data are associated with this article.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgments

The authors thank Mr. Denis von Kaeppler and Mr. William Gandler, Center for Information Technology, and Dr. Sean Davis, Center for Cancer Research at the National Cancer Institute, National Institutes of Health for discussions and suggestions during the preparation of the manuscript. The opinions expressed in the paper are those of the authors and do not necessarily reflect the opinions of the National Institutes of Health.

References

- Stephens ZD, Lee SY, Faghri F, *et al.*: **Big Data: Astronomical or Genomical?** *PLoS Biol.* 2015; **13**(7): e1002195.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Collins FS, Varmus H: **A new initiative on precision medicine.** *N Engl J Med.* 2015; **372**(9): 793–795.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- ISO 14721:2012 - Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model.** 2018; [cited 1 Aug 2018].
[Reference Source](#)
- Wikipedia contributors: **Open Archival Information System.** In: *Wikipedia, The Free Encyclopedia.* 2018; [cited 1 Aug 2018].
[Reference Source](#)
- Standard ISO: 14721: 2003: Space Data and Information Transfer Systems - Open Archival Information System Reference Model.** International Organization for Standardization. 2003.
[Reference Source](#)
- Ball A: **Briefing Paper: The OAIS Reference Model.** UKOLN: University of Bath. 2006.
[Reference Source](#)
- Haendel MA, Vasilevsky NA, Wirz JA: **Dealing with data: a case study on information and data management literacy.** *PLoS Biol.* 2012; **10**(5): e1001339.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Michener WK: **Ten Simple Rules for Creating a Good Data Management Plan.** *PLoS Comput Biol.* 2015; **11**(10): e1004525.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goodman A, Pepe A, Blocker AW, *et al.*: **Ten simple rules for the care and feeding of scientific data.** *PLoS Comput Biol.* 2014; **10**(4): e1003542.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Schreier AA, Wilson K, Resnik D: **Academic research record-keeping: best practices for individuals, group leaders, and institutions.** *Acad Med.* 2006; **81**(1): 42–47.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Schnell S: **Ten Simple Rules for a Computational Biologist's Laboratory Notebook.** *PLoS Comput Biol.* 2015; **11**(9): e1004385.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- National Library of Medicine, National Institutes of Health: **Key Elements to Consider in Preparing a Data Sharing Plan Under NIH Extramural Support.** U.S. National Library of Medicine; 26, Jun 1012, updated 3 Jan 2013 [cited 19 Jun 2017].
[Reference Source](#)
- Williams M, Bagwell J, Nahm Zozus M: **Data management plans: the missing perspective.** *J Biomed Inform.* 2017; **71**: 130–142.
[PubMed Abstract](#) | [Publisher Full Text](#)
- National Institute of Standards, Technology: **FIPS 200, Minimum Security Requirements for Federal Information and Information Systems.** CSRC, 2006; [cited 7 Feb 2018].
[Reference Source](#)
- O'Reilly PD: **Federal Information Security Management Act (FISMA) Implementation Project.** Created June 12, 2009; updated March 19, 2018.
[Reference Source](#)
- Malin BA, Emam KE, O'Keefe CM: **Biomedical data privacy: problems, perspectives, and recent advances.** *J Am Med Inform Assoc.* 2013; **20**(1): 2–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kazic T: **Ten Simple Rules for Experiments' Provenance.** *PLoS Comput Biol.* 2015; **11**(10): e1004384.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- U.S. National Library of Medicine: first published 18 June, 2012; updated 29 March 2016.
[Reference Source](#)
- Rubinstein YR, McInnes P: **NIH/NCATS/GRDR® Common Data Elements: A leading force for standardized data collection.** *Contemp Clin Trials.* 2015; **42**: 78–80.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Malone J, Stevens R, Jupp S, *et al.*: **Ten Simple Rules for Selecting a Bio-ontology.** *PLoS Comput Biol.* 2016; **12**(2): e1004743.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ravagli C, Pognan F, Marc P: **OntoBrowser: a collaborative tool for curation**

- of ontologies by subject matter experts. *Bioinformatics*. 2017; **33**(1): 148–149.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Leonelli S: **Global Data Quality Assessment and the Situated Nature of “Best” Research Practices in Biology**. *Data Science Journal*. 2017; **16**: 32.
[Publisher Full Text](#)
23. Weiskopf NG, Weng C: **Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research**. *J Am Med Inform Assoc*. 2013; **20**(1): 144–151.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Alqasab M, Embury SM, Sampaio S: **A Maturity Model for Biomedical Data Curation**.
[Reference Source](#)
25. Navale V, Ji M, McCreedy E, et al.: **Standardized Informatics Computing platform for Advancing Biomedical Discovery through data sharing**. *bioRxiv*. 2018.
[Publisher Full Text](#)
26. Kirlow PW: **Life Science Data Repositories in the Publications of Scientists and Librarians**. [cited 31 Oct 2017].
[Publisher Full Text](#)
27. Pampel H, Vierkant P, Scholze F, et al.: **Making research data repositories visible: the re3data.org Registry**. *PLoS One*. 2013; **8**(11): e78080.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
28. **Data Storage Best Practices**. In: *Fred Hutch Biomedical Data Science Wiki*. [cited 22 Jul 2018].
[Reference Source](#)
29. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al.: **The FAIR Guiding Principles for scientific data management and stewardship**. *Sci Data*. Nature Publishing Group; 2016; **3**: 160018.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Navale V, Bourne PE: **Cloud computing applications for biomedical science: A perspective**. *PLoS Comput Biol*. 2018; **14**(6): e1006144.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Liu B, Madduri RK, Sotomayor B, et al.: **Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses**. *J Biomed Inform*. 2014; **49**: 119–133.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Ohno-Machado L, Sansone SA, Alter G, et al.: **Finding useful data across multiple biomedical data repositories using DataMed**. *Nat Genet*. 2017; **49**(6): 816–819.
[PubMed Abstract](#) | [Publisher Full Text](#)
33. Corpas M, Kovalevskaya NV, McMurray A, et al.: **A FAIR guide for data providers to maximise sharing of human genomic data**. *PLoS Comput Biol*. 2018; **14**(3): e1005873.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Jagodnik KM, Koplev S, Jenkins SL, et al.: **Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop**. *J Biomed Inform*. 2017; **71**: 49–57.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Fielding RT: **Architectural Styles and the Design of Network-based Software Architectures**. Taylor RN, Dissertation Committee Chair. PhD, University of California, Irvine. 2000.
[Reference Source](#)
36. The Linux Foundation and Open API Initiative: **Open API Initiative**. In: *Open API Initiative*. [cited 8 Aug 2017].
[Reference Source](#)
37. Swaminathan R, Huang Y, Moosavinab S, et al.: **A Review on Genomics APIs**. *Comput Struct Biotechnol J*. 2015; **14**: 8–15.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
38. Cheemalapati S, Chang YA, Daya S, et al.: **Hybrid Cloud Data and API Integration: Integrate Your Enterprise and Cloud with Bluemix Integration Services**. IBM Redbooks; 2016.
[Reference Source](#)
39. Hart EM, Bamby P, LeBauer D, et al.: **Ten Simple Rules for Digital Data Storage**. *PLoS Comput Biol*. 2016; **12**(10): e1005097.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. Davis S, Meltzer PS: **GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor**. *Bioinformatics*. 2007; **23**(14): 1846–1847.
[PubMed Abstract](#) | [Publisher Full Text](#)
41. Carmen Legaz-García MD, Miñarro-Giménez JA, Menárguez-Tortosa M, et al.: **Generation of open biomedical datasets through ontology-driven transformation and integration processes**. *J Biomed Semantics*. 2016; **7**: 32.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Referee Status:



Version 1

Referee Report 09 October 2018

doi:10.5256/f1000research.17492.r38251



Ravi Madduri 

Computation Institute, University of Chicago, Chicago, IL, USA

The article titled "Long-term preservation of biomedical research data" by Navale et.al, is a timely article that highlights the need for a long term strategy for preservation of data products generated from research projects. Often times a lot of time is spent in the initial activities of a research project which involves data collection, processing, analysis, sharing and publishing but substantially less time is spent in curating the data, making data reusable and finally long term preservation of data products. The paper presents a strategy for long term data preservation which the authors have broken down into multiple stages. This reviewer agrees with the strategy and the overall presentation of the strategy in the manuscript. There is, however, one important challenge in long term data management that this reviewer felt has not been covered adequately which is the economics of long term data preservation. Determining which data products are preserved and for how long is an important part of the puzzle. Additionally, the economics of long term storage along with who the stakeholders are and what the incentives for this to happen is also important. The article would be made better with these additions.

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Are arguments sufficiently supported by evidence from the published literature?

Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 24 September 2018

doi:10.5256/f1000research.17492.r37766



David Giaretta 

Primary Trustworthy Digital Repository Authorisation Body Ltd (PTAB Ltd), Dorset, UK

The reason I ticked "Partly" in the above checkboxes is that the article has omitted clear discussion of the key concepts in OAIS for preservation, which are also key to the "R" in FAIR i.e. Reuse.

These may be addressed by including a discussion of the OAIS Information Model.

OAIS defined Long Term Preservation as *The act of maintaining information, Independently Understandable by a Designated Community, and with evidence to support its Authenticity, over the Long Term.*

To explain what I mean briefly, to ensure understandability the archive should collect the appropriate Representation Information, and ensure that as the Designated Communities Knowledge Bases, the Representation Information must be supplemented. For example if the ontology, which is used to understand the biomedical data, goes out of use over time, for example the URL for its location no longer works, the archive will need to ensure that the original ontology remains available.

Similarly, as evidence for Authenticity the archive should collect Provenance about the data, as is briefly mentioned in the article.

The Archival Information Package, which is the AIP shown in the Functional Model, is a way to ensure that the archive has captured all the information required for Long Term Preservation, including Representation Information, Provenance Information and several other items. In order to create the AIP for a dataset the archive should ensure that the required information is captured during Ingest, and is maintained over time.

The data management plan should help to ensure that the appropriate information is captured over the course of the project in order to provide to the archive.

One last point which is worth mentioning is that the evaluation of the ability of an archive is also possible through ISO 16363 audit and certification.

References

1. Standard ISO: 14721: International Organization for Standardization: 2012: Space Data and Information Transfer Systems - Open Archival Information System Reference Model. 2012.

Is the topic of the opinion article discussed accurately in the context of the current literature?

Partly

Are all factual statements correct and adequately supported by citations?

Partly

Are arguments sufficiently supported by evidence from the published literature?

Partly

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: I am an expert in digital preservation - see www.iso16363.org and www.giaretta.org

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 19 September 2018

doi:[10.5256/f1000research.17492.r37763](https://doi.org/10.5256/f1000research.17492.r37763)



Chaitan Baru

San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA

In sum, this paper is highlighting the importance and need for "data stewardship", viz., well-considered data management plans for every project that produces/generates data. Data stewardship plans should be developed early in a project cycle--indeed, along with the development of the science/research goals of the project itself.

The authors caution that this problem becomes even more urgent in the era of "big data". They recommend use of an existing approach, viz., the Open Archival Information System (OAIS).

The issues mentioned and the approaches suggested are very reasonable, and very much in step with similar concerns and approaches in several other domains, which are all facing the data deluge.

In fact, I have heard so much discussion and read so many articles on this topic--supporting the approaches described in this paper as well--that I am now concerned that, as a community, we are probably not taking the right approach to this problem.

First, the only way that the community may pay attention and spend resources on this problem is if they see value. These type of articles should probably begin with the value of doing this work, rather than the cost. Almost all articles on this topic talk in detail about the costs, and simply presume that the value exists. Value could be demonstrated by showing real science examples that benefited from archival data; examples of studies that went into archival data and found something new and interesting. Or, conversely, studies that duplicated effort, or failed in other ways, for not digging into archival data.

Second, curation is not a static process. The costs of curation, done properly, may actually be way more than what these papers suggest. Since science is not static, the relevance or "meaning" of a particular dataset is also not static. Data may become less or more valuable as the field progresses. All of that speaks to what I would call "continuous curation" of scientific data, and not just one-time curation at the time of creation.

Finally, what to do when everyone is extolling the need and virtue of curation but no one is spending nearly enough resources to do the job right. One would think that this is the classic use case for AI and "smart" techniques. Why not let the computer do the job that no human is willing or able to do for the amount of money we are willing to spend. Rather than AI, this may be the classic use case for IA--intelligence augmentation, with a human in the loop.

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Are arguments sufficiently supported by evidence from the published literature?

Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Data science; database systems; informatics; scalable data systems.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 17 September 2018

doi:[10.5256/f1000research.17492.r38250](https://doi.org/10.5256/f1000research.17492.r38250)



George Alter 

Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan, Ann Arbor, MI, USA

This article explains the importance of engaging data stewards in planning for data preservation at the beginning of data collection. It makes an important contribution by offering useful guidance about each stage of the data life cycle, such as the use of CDEs and ontologies, the Data Curation Maturity model, and open formats. Data would be more plentiful, better documented, and easier to reuse if this planning took place.

The following suggestions are intended to strengthen and expand the current draft of the paper. Since this is explicitly an "opinion article," some of these comments reflect my own opinions, which the authors may not share.

1. Under "Mandatory Responsibilities" the OAIS standard says that an archive must: "Ensure that the information to be preserved is Independently Understandable to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information." I think a statement like this provides a focus for the activities involved in preparing data for preservation. OAIS is not simply about assuring that the data survive. Preservation also assures that the data will be reusable (the 'R' in FAIR) in the future. It might be worthwhile to point out the relevance of OAIS for FAIR earlier in the paper.

2. Although Figure 1 suggests that the SIP arrives at the archive fully formed, the text of the OAIS standard emphasizes that the relationship between the archive and the data producer may involve a lot of negotiation. Data repositories often need to contact the data producer several times to get the information that they need. This is a costly process, and the recommendations in this paper would reduce

those costs. This is worth mentioning, because data producers often only see the costs of preparing data for sharing.

3. The discussion of CDEs and ontologies could also mention the Center for Expanded Data Annotation and Retrieval (CEDAR), which has developed tools for creating metadata.

4. When choosing a data repository, data stewards should favor repositories that offer an assurance of permanence and trustworthiness. This is especially important in the biomedical community, because valuable data have been lost when repositories and databases closed. There are several bodies that certify data repositories as trustworthy. ISO has a "Standard for Trusted Digital Repositories" (ISO 16363), which is very comprehensive and usually involves an external auditor. The CoreTrustSeal has a smaller list of requirements and relies on a self-audit.

5. Confidentiality and disclosure are mentioned in a paragraph about security controls, but this deserves a little more space. Data producers set the terms of future reuse of data when they make informed consent agreements with subjects. If data sharing is not anticipated in the informed consent agreement, it is very difficult to share the data. Since the informed consent agreement is supervised by an IRB, the IRB should also approve plans for future data sharing. These plans could involve an array of legal (data use agreements) and technical (anonymization, "data enclaves") measures to protect confidential information.

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Are arguments sufficiently supported by evidence from the published literature?

Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 13 September 2018

doi:10.5256/f1000research.17492.r37761



Jane Greenberg 

Metadata Research Center, College of Computing & Informatics, Drexel University, Philadelphia, PA, USA

This is a timely scientific undertaking and stands as an important, original contribution to the body of research on data preservation in the context of big data, biomedical research, and data management/archiving. Furthermore, the research makes important links to key underpinnings and research covering data quality, modeling, along with metadata, and ontologies.

Much of the big data research to date has focused on algorithmic work, visualization, and related topics, whereas data preservation and archival research has been driven largely from the context of institutional repositories that are not necessarily storing big data.

This paper bridges these two research areas, using the OAIS model as a platform to weave together these topics, and provide salient discussion about keys pillars of covering preservation planning, administration, ingest, data management, archival storage, and access. The research is further contextualized by the FAIR principles that seek to ensure that data not only findable and accessible, but also interpretable and reusable. Further, the authors hone in on the role of data modeling, metadata, including the provenance model and ontologies.

The writing is excellent, the synthesis of the literature and integration of other research is solid. Additionally, the diagrams are illustrative. (Note, this my first review in this system, and I am super excited and pleased to have had the opportunity to serve as a reviewer for this original research, and eager to share with colleagues. In fact, have already shared this link with colleagues, so they can include this piece as a key reading across related courses as our academic quarter at Drexel is almost underway).

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Are arguments sufficiently supported by evidence from the published literature?

Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Metadata, Ontologies, Semantics, Linked data, Data management, Economics of metadata, Big data

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research