The Biomedical Data Journal in the New Era of Open Access to Scientific Information

Iordanis Mourouzis, Constantinos Pantos (✉)

Department of Pharmacology, University of Athens, 75 Mikras Asias Ave.,11527 Goudi, Athens, Greece
http://school.med.uoa.gr/en/

ABSTRACT

The new scientific landscape requires fuller and wider access to scientific information and research data as an important means to accelerate innovation and improve transparency of the scientific process. In this emerging environment, the new publication model of data journals will play a key role by establishing high standards for dataset publication and promoting the deserved credit and recognition to authors. Biomedical Data Journal (BMDJ) is an open access journal aiming to facilitate the presentation, validation, use, and re-use of datasets with focus on publishing biomedical datasets that can serve as a source for simulation and computational modelling of diseases and biological processes. The rise of computational science could lead to exciting developments in the field of biomedical research. BMDJ has the ambition to pave the way in this direction through meeting the existing need for high-quality, well structured, open-access, biomedical datasets. BMDJ has established a new approach that includes a number of initiatives to potentiate the use of data papers as a source of new scientific creation, to maximize reach and visibility of data publications and to introduce formal citation of data papers. Sufficient motivation for researchers to release high quality experimental biomedical data goes through providing the respective credit and recognition to authors of data papers.

Introduction

Research progress is based on former achievements and the success largely depends on accessing and sharing scientific information. The World Wide Web has fundamentally changed the scientific world by creating the potential to make knowledge available immediately on a large scale. This potential generates new opportunities and challenges in biomedical research. In particular, fuller and wider access to scientific information and research data is now considered an important way to accelerate innovation and economic growth, foster collaboration and avoid duplication of efforts, improve transparency of the scientific process and quality of research results.[1]

The new scientific landscape of open-access to biomedical data

A new scientific landscape is gradually formed where research data are no longer regarded as single use but as intermediate products that can be re-used to create innovation and economic growth. Research funding agencies, licensing authorities and academic institutions adopt policies

✉ Corresponding Author: Fax: (+30210) 7462560; E-mail: cpantos@med.uoa.gr

and promote initiatives to make research data open and readily accessible. Horizon 2020, the biggest EU Research and Innovation programme ever, has established the Open Research Data Pilot which aims to improve and maximise access to and re-use of research data generated by EU-funded projects.[2] In the UK, the Research Information Network (RIN), in association with the Joint Information Systems Committee (JISC) and the Natural Environment Research Council (NERC), commissioned a study on how the process of data publishing and quality assurance could be enhanced for the future.[3] Furthermore, the NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. The policy expects final research data, especially unique data, from NIH-supported research efforts to be made available to other investigators. Data sharing applies to human subjects and laboratory research. Thus, researchers submitting grant applications to NIH are asked to include a data sharing plan or an explanation of why data sharing is not possible.[4]

Several traditional journals have also adopted new policies requiring authors to make their datasets publicly available in order to promote reliability of research findings. For example, the new PLOS data policy states that *authors must make all data publicly available, without restriction, immediately upon publication of the article*. Since March 2014, all authors who submit to a PLOS journal have been asked to provide a statement, describing where and how others can access each dataset that underlies the findings.[5] An increasing number of traditional journals are now adopting similar obligatory or optional requirements.

In this emerging environment, the new publication model of data journals will play a key role by establishing high standards for dataset publication and by promoting the deserved credit and recognition to authors.

Data papers and journals as a new publication model to promote open-access to research information

Data journals can play an important role as part of a joint effort by the biomedical research community towards open-access to research information. The primary purpose of data journals is to provide the infrastructure and scholarly reward opportunities to encourage the publication of data papers. Data papers are peer reviewed publications created to describe high quality datasets in a detailed

and comprehensible manner and provide a statement of how the data was collected.[6] Data papers differ from regular publications because they lack any analyses or conclusions and should be considered as a companion to a website-repository that will host the dataset.

BMDJ's focused scope

Biomedical Data Journal (BMDJ), ISSN 2367-5322, is an open access journal aiming to facilitate the presentation, validation, use, and re-use of datasets with focus on publishing *biomedical datasets* that can serve as a source for simulation and computational modelling of diseases and biological processes. Computational modelling can bridge the gap between experiments and patients by integrating data obtained from experimental cell and animal based models to patients. Datasets availability is critical for training, optimization and validation of 'integrative' mathematical models based on experimental (cell and animal datasets) and clinical observations (human datasets). However, acquiring high-quality biomedical datasets requires a substantial technical endeavour. Indeed, the time and financial cost of experimental data production can become a barrier to entry and slow down the development of systems medicine approaches. It is increasingly recognized that the integration of a variety of biological and medical research data to produce or refine disease models using advanced statistical, computational and mathematical approaches could tackle the complexity of biological systems. In fact, one of the main focuses of the EU Research and Innovation programme is the issue of "Understanding diseases via systems medicine." The European Research Council recognizes that computational approaches could address the current knowledge gaps in disease pathophysiology in order to support innovation in the development of novel evidence-based treatments.[7]

The rise of computational science is a reality (figure 1) and has led to exciting developments in many scientific areas.[8] The availability of high-quality, open-access datasets has allowed researchers to make substantial scientific contributions without the use of traditional tools in other disciplines beyond biomedical research. As an example of the potential of computational science, we could refer to computational chemistry which was awarded with the 2013 Nobel Prize in Chemistry. The prize was awarded for the development and application of methods to simulate the behaviour of molecules at various scales, from single

molecules to proteins, shedding light on phenomena as diverse as protein folding, catalysis, electron transfer and drug design. The challenges in the field of biomedical research are even greater and BMDJ has the ambition to pave the way in this direction through meeting the existing need for high-quality, well structured, open-access, biomedical datasets.

We can refer to representative examples from cardiology and cancer research to give some idea about the magnitude of challenges we have to face. In fact, computational modelling is becoming essential for bringing cancer drug discovery into a new era by predicting disease progress and outcomes, simulating anti-cancer drugs tissue delivery and predicting resistance to chemotherapy, developing non-invasive diagnosis to monitor treatment response and designing new chemotherapy targets.[9,10] In the case of cardiology research, mathematical modelling creates new tools to investigate complex hemodynamics of the circulatory system, understand cardiac development, construct patient-specific models of reconstructive surgeries for the treatment of congenital heart diseases, understand the pathogenesis of complex arrhythmias and make easier screening for new drug discovery.[11,12,13,14]

The entity of data publications in BMDJ

BMDJ will accept submissions of biomedical data papers based on datasets that have been already analysed and their results published in traditional journals. These data papers are regarded as complementary to traditional research publications and among others promote the scientific gold standard of reproducibility. Reproducibility calls for data and is considered the minimum criterion to assess the value of scientific claims.[15] However, BMDJ will also accept submissions of biomedical data papers not related to traditional publications that describe their analysis. This policy will allow authors to publish valuable datasets whose analysis may be considered by traditional journals to produce insignificant results.

The Biomedical Data Journal will be empowered by the functionalities established in the *OpenScienceLink (OSL)* platform. *OSL* is an EU-funded project which is designed to introduce and pilot a holistic approach to the publication, sharing, linking, review and evaluation of research results, based on the open access to scientific information. *OSL* creates a range of novel pilot services that could alleviate the lack of structured data journals and associated data models, the weaknesses of the review process, the poor link-

ing of scientific information, as well as the limitations of current research evaluation metrics and indicators. Authors will be able to submit their data papers via OSL platform. In this regard, submissions should include a document and the dataset(s). The document is expected to outline the broader goals that have motivated the creation of this dataset and the potential re-use value, explain in detail the methods used to create the data, describe each data record associated with the submission and give an overview of the data files and their structure. BMDJ encourages authors to submit dataset files in any format via OSL platform. Alternatively, authors could deposit their datasets in an appropriate open-access repository and provide a link.

Validation of data papers based on peer-review

Validation of data papers is necessary in order to apply the esteem of traditional papers to data publication. One of the best known methods to assess the validity of publications is peer-review. However, the ideal model of peer review for data papers is an open issue. Both technical and scientific criteria have been proposed for the evaluation of data papers.[16] Technical criteria refer to the completeness of the dataset itself and the adequate descript-tion, while scientific criteria refer to evaluation of the methods of data collection, the plausibility of the data and the re-use value.

BMDJ has established a set of specific technical and scientific criteria that will be used to validate and select high quality datasets for publication.[17] According to these criteria, the *dataset* will be evaluated in terms of internally consistent data values (falling within an appropriate range for the phenomenon being measured/observed), sufficient accuracy and precision, avoidance of missing values, plausibility and compatibility in comparison with similar datasets, appropriate extent and coverage of the data to match the expectations of re-use. In addition, *the data description document* will be evaluated in terms of adequacy of the data collection methods including methodological expertise of the authors, complete and comprehensible description of the dataset and clarity of presentation and adequacy of citations. An important part of the evaluation will be the appropriate scientific argumentation about the re-use value of the dataset with emphasis on computational modelling. The peer-review process will be conducted with the guidance of the editorial board
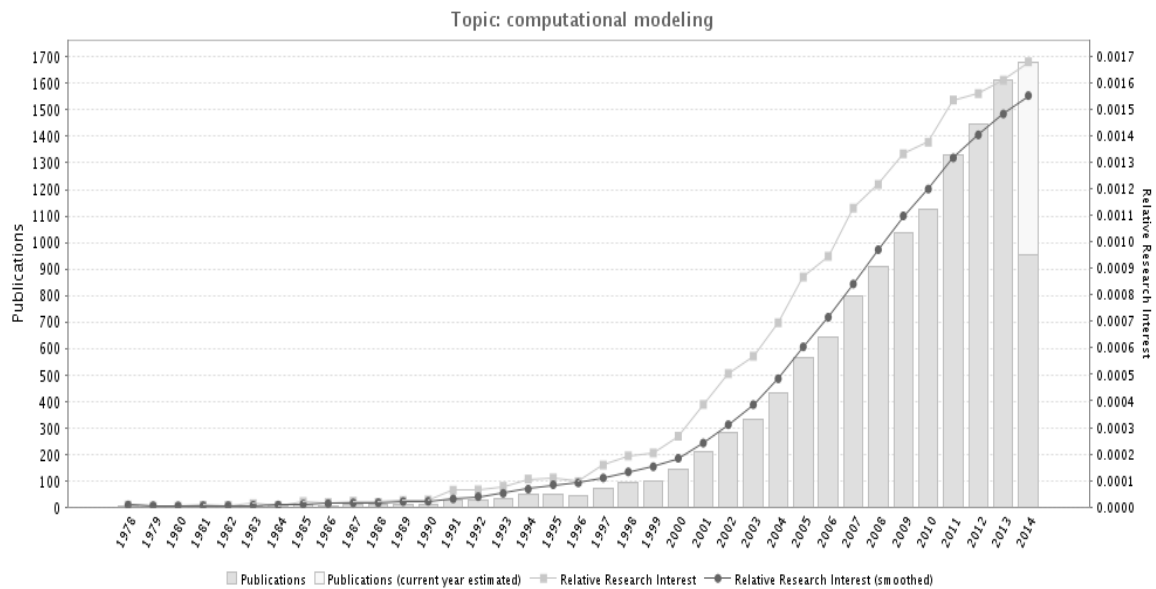
Figure 1: Continuous rise of research interest for the field of computational modelling (based on data from the GoPubmed Search Engine).
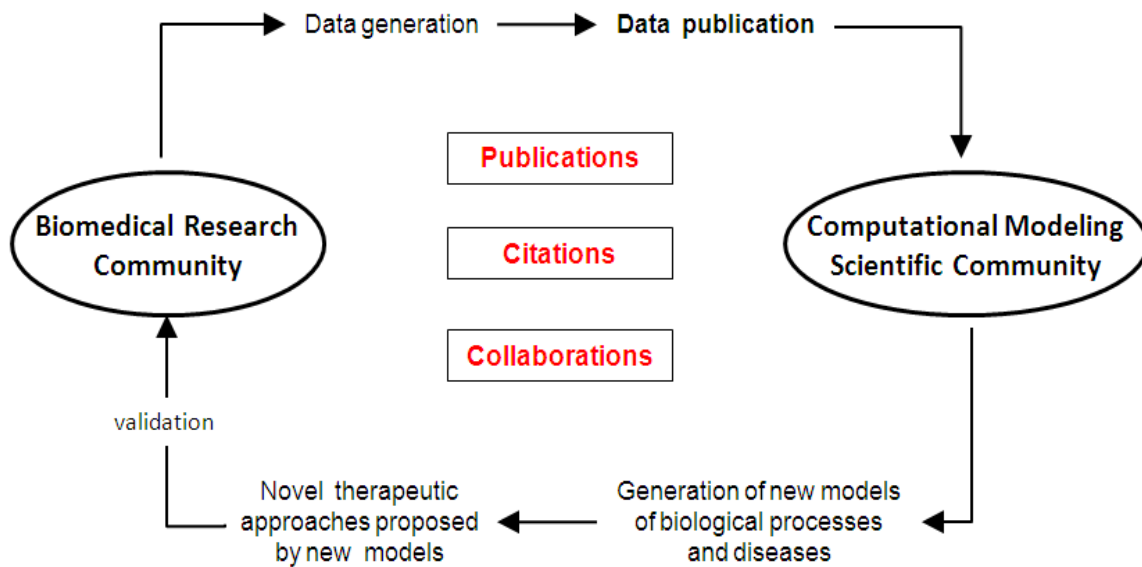
Figure 2: Open-access, high quality biomedical data papers as the catalyst to create new models of disease and therapy and create an impact to society.

which includes numerous experts covering nearly all fields of biomedical research and also includes experts in the field of computational modelling.

The development of peer-review for data papers is a major challenge considering that traditional peer-review is now facing several criticisms including reduced reliability, objectivity and inefficiency to detect error or fraud.[18] The process of peer review in BMDJ will be automatically handled by the OSL platform based on a number of new tools designed to improve the speed and effi-

ciency of the process. These tools include the online discussion forum for the reviewers group, the reviewer's evaluation tool and the interactive open post-review discussion.

Steps towards providing higher visibility and impact for data papers

A key point for the success of a data journal is to provide sufficient motivation for authors to release high quality, experimental biomedical data

that need great human and financial resources to collect. Motivation goes through providing the respective credit and recognition as well as career-related rewards for authors of data papers. BMDJ has established a new approach that includes a number of initiatives to potentiate the use of data papers as a source of new scientific creation, to maximize reach and visibility of data publications and to introduce formal citation as well as alternative measures of research impact and influence of data papers.

The direct link between BMDJ and the scientific community of computational modelling will permit to actively promote the use of data papers and create a prototype to show how data publication may initiate a cascade of events that will foster scientific collaboration and result in innovation. Computational modelling of biological processes and diseases is restrained by the limited available experimental data that are appropriate for this purpose. Thus, data papers in BMDJ could function as the catalyst to create new models of disease and therapy and create an impact on society. This process is expected to bring significant credit for authors through the production of citations and a generation of new experimental data that will be necessary to validate and refine the respective models (figure 2).

The value and impact of data papers greatly depend on how easily they can be discovered and accessed by other scientists. To facilitate this process, BMDJ is indexed by CrossRef and Google Scholar and is in the process of arranging indexing by PubMed. At a second stage, it will be indexed by Scopus and other relevant services.

BMDJ will be also empowered by a new functionality created by OSL platform: a search engine especially designed for data papers providing links to both text files and datasets. This search engine will be based on the construction of detailed metadata file during the submission process. Metadata is a detailed machine-readable file that provides a structured description of the dataset. BMDJ has adopted specific standards for metadata that will serve the needs of biomedical datasets based on experts' consultations.

Data citations will facilitate giving scholarly credit and normative and legal attribution to all contributors to the data. The Joint Declaration of Data Citation Principles states that "data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications." BMDJ will enable formal citation of published data papers by assigning unique permanent identifiers: Digital Object Identifiers (DOIs). At present, there is consensus that a citation of a data paper should use a minimum of five elements similar to traditional article citations: creator(s), title, year, publisher and identifier. This format includes all the information necessary to obtain a DataCite DOI or be listed in the Thomson-Reuters Data Citation Index, and is also in agreement with CODATA's recommendation.[19,20]

Besides citations, the online publishing medium and the advantages inherent in Open Access content provide tremendous potential for maximizing reach and visibility of publications and creating alternative measures of research impact and influence. BMDJ works with OSL to help authors maximize the impact of their published work through communication to interested groups of experts according to scientific profile, email newsletters, Social Media coverage and targeted press releases. OSL will also provide new tools of measuring scientific impact related to online usage, downloads, blog and media coverage, commenting activity and social bookmarking.

In conclusion, the new scientific environment requires that there is open-access to biomedical research data as an important means to accelerate innovation and improve transparency of the scientific process. Biomedical Data Journal is based on a new publication model that will play a key role in establishing high standards for dataset publication and promoting the deserved credit and recognition to authors. BMDJ aims to facilitate the presentation, validation, use, and re-use of datasets, with focus on publishing biomedical datasets that can serve as a source for simulation and computational modelling of diseases and biological processes. BMDJ has the ambition to pave the way towards high-quality, well structured, open-access, biomedical datasets creating the basis for exciting developments in the field of systems medicine and biology.

Acknowledgment

References

1. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. European Commission. 20 March 2014. Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

2. Guidelines on Data Management in Horizon 2020, version 1.0, 11 December 2013. Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

3. Jubb M. To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Report commissioned by the Research Information Network. June 2008.

4. Final NIH Statement on Sharing Research Data, NOT-OD-03-032 (National Institutes of Health, 2003). Available at http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

5. Silva, L. PLOS' New Data Policy: Public Access to Data. 24 February 2014. Available at http://www.plos.org/plos-data-policy-faq.

6. Newman P, Corke P. Data papers – peer reviewed publication of high quality datasets. Int J Rob Res. 2009; 28(5): 587–587. DOI: 10.1177/0278364909104283.

7. HORIZON 2020 Work programme 2014-2015, Health, demographic change and wellbeing. PHC2-Understanding diseases: systems medicine. Available at http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-health_en.pdf.

8.Hanson B, Sugden A, Alberts B. Making data maximally available. Science. 2011 Feb 11; 331(6018): 649. DOI: 10.1126/science.1203354.

9.Kim M, Gillies RJ, Rejniak KA. Current advances in mathematical modeling of anti-cancer drug penetration into tumor tissues. Front Oncol. 2013; 3:278. DOI: 10.3389/fonc.2013.00278.

10. Zhang P, Brusic V. Mathematical modeling for novel cancer drug discovery and development. Expert Opin Drug Discov. 2014 Jul 25: 1-18. DOI: 10.1517/17460441.2014.941351.

11. Lindert S, Li MX, Sykes BD, McCammon JA. Computer-Aided Drug Discovery Approach Finds Calcium Sensitizer of Cardiac Troponin. Chem Biol Drug Des. 2014 Jun 20. DOI: 10.1111/cbdd.12381.

12. Pennati G, Corsini C, Hsia TY, Migliavacca F. Computational fluid dynamics models and congenital heart diseases. Front Pediatr. 2013 Feb 26;1: Article 4. DOI: 10.3389/fped.2013.00004.

13. Gould RA, Aboulmouna LM, Varner JD, Butcher JT. Hierarchical approaches for systems modeling in cardiac development. Wiley Interdiscip Rev Syst Biol Med. 2013 May-Jun; 5(3): 289-305. DOI: 10.1002/wsbm.1217.

14. Roberts BN, Yang PC, Behrens SB, Moreno JD, Clancy CE. Computational approaches to understand cardiac electrophysiology and arrhythmias. Am J Physiol Heart Circ Physiol. 2012 October 1; 303(7): H766-83. DOI: 10.1152/ajpheart.01081.2011.

15. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007 March 20; 146(6): 450-3. DOI: 10.7326/0003-4819-146-6-200703200-00154.

16. Callaghan S, Donegan S, Pepler S, et al. Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. Int J Digital Curation. 2012; 7(1): 107–113. DOI: 10.2218/ijdc.v7i1.218.

17. Lawrence B, Jones C, Matthews B, et al. Citation and peer review of data: Moving towards formal data publication. Int J Digital Curation. 2011;6(2):4–37. DOI: 10.2218/ijdc.v6i2.205.

18. Herron DM. Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review. Surg Endosc. 2012; 26(8): 2275–2280. DOI: 10.1007/s00464-012-2171-1.

19. CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. Data Sci J. 2013;12:1–75. DOI: 10.2481/dsj.OSOM13-043.

20. Starr J, Gastl A. isCitedBy: a metadata scheme for DataCite. D-Lib Magazine. 2011; 17 (1/2). DOI: 10.1045/january2011-starr.