Standardisation and Quality Assurance in life-science research - crucially needed or unnecessary and annoying regulation?

Susanne Hollmann[1, 2], Teresa K. Attwood[4], Erik Bongcam-Rudloff[9], Deborah Duca[8], Domenica D'Elia[5], Christoph Endrullat[3], Marcus Frohme[7], Katrin Messerschmidt[6], Babette Regierer[2]

[1] Potsdam University, Potsdam 14476, Germany
[2] SB Science Management UG (haftungsbeschränkt), 12163 Berlin, Germany
[3] formerly at Technical University of Applied Sciences Wildau, Germany
[4] University of Manchester, UK
[5] CNR - Institute for Biomedical Technologies, 70126 Bari, Italy
[6] Potsdam University, Department of Synthetic Biosystems, Potsdam 14476, Germany
[7] Technical University of Applied Sciences Wildau, Molecular Biotechnology and Functional Genome Analysis, Germany
[8] Research Support Services; University of Malta
[9] SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Open Science describes the ongoing transitions in the way research is performed, i.e. researchers collaborate, knowledge is shared, and science is organised. It is driven by digital technologies and by the enormous growth of data, globalisation, enlargement of the scientific community and the need to address societal challenges [2]. It has now widely been recognised that making research results more accessible to all societal actors contributes to better and more efficient science, as well as to innovation in the public and private sectors [3,4]. However, the reuse of research results can only be achieved reliably and efficiently, if these data are valorized in a specific manner. Data are to be generated, formatted and stored according to Standard Operating Procedures (SOPs) and according to sophisticated Data Management Plans [5]. Hence, to generate accurate and reproducible data sets, to allow interlaboratory comparisons as well as further and future use of research data it is mandatory to work in line with good laboratory practices and well-defined and validated methodologies. Within this article, members of the Cost Action CHARME [6] will discuss aspects of quality management and standardization in context with Open Access (OA) efforts. We will address the question: Are Standardisation and Quality Management measures in life-science research crucially needed or introduce further unwanted means of regulation?

Keywords: Open Access (OA), Standardisation, Validity and reliability of data, Standard Operating Procedures (SOPs), Quality Management (QM), Quality Control (QC), Quality Assurance (QA), Seal of Science, Education.

Introduction into the problem

The root word for “science” is the Latin word “scientia” which means 'knowledge”, with research being the tool towards obtaining this knowledge [15]. Posing research questions and designing experiments to answer those questions have enabled the scientific community and the society in general to gain a deep understanding of the world around us [1]. Access to scientific knowledge is essential for any research activity. Over hundreds of years researchers have been focused on generating data and gaining knowledge to answer particular research questions and to expand the insight to their field of interest. Results and conclusions were published but not open to everybody, and thereby shared with only a limited number of the peer researchers. This resulted in recognition, intellectual merit, and building a common knowledge base but the potential of research results obtained was not fully exploited In contrast, unimpeded flow of knowledge is important for the implementation of research results in innovations and as a source of inspiration for new ideas [2, 22]. The easier research results are findable and accessible to anyone, the better they can be the basis of further research and innovation. Open Access allows for quick and easy access to relevant scientific content by making scientific information openly available. Besides availability, reusability of research results is crucial for the success of Open Access. This demands a high quality of the data and information to be shared. This also benefits society as a whole: anyone interested can find research results and scientific publications on the Internet, download, read and share them [7,8].

With the advent of the new Framework Program Horizon 2020 the Open Access policy is broadly implemented throughout the funding program, and many national funders across Europe follow this example. The sharing of research results in Open Access format publications is no longer an option, but a mandatory task for all publicly funded research to ensure accessibility and reusability of research results [2,6]. With this new policy of openness we face a new challenge of ensuring quality of research across all scientific disciplines and actors resulting in a need to share and implement standards, SOPs and Good Scientific Practice among all these groups involved.

General repositories of data are not sufficient; rather descriptions and detailed annotations are required to provide open science in order to accelerate the innovation processes. This includes the background behind the generation of certain data formats as well as the potency for interoperability and transferability between different data formats. Definition of quality benchmarks for data are also important in order to define metrics which are applicable and reasonable for building a framework around good data quality [7]. The data quality is directly linked to the quality of the biological samples and procedure/protocols used. Hence, high-quality data can only be obtained by the respective use of high-quality samples[10].

Those aspects face questions about who is responsible for reviewing the data quality and how the handling of low quality data should be performed. Proficiency testing methods for data generation using well-known and high quality reference data sets are an additional point.

The reproducibility debate in the life sciences revealed that scientific results are not only suffering from lack of reusability but have been demonstrated by some examples

that the data seem to a large extent to be irreproducible [17]. The impact is enormous, not only affecting the scientific progress but also limiting the translation of research results into application and increasing the costs of research. Furthermore the perception of the ”truth of science” in the public may be deteriorated. In knowledge-based societies this is unacceptable and the recent discussions about fake and true news (also from science) started by political leaders for economic interests opens a Pandoras Box of misinformation with global impact.

The reasons for this “reproducibility crisis” could be manifold, among them lack of a good study design, controls or insufficient documentation, but also non-scientific reasons might contribute to this like pressure to publish, lack of funding to replicate experiments thoroughly or simply the exclusion of negative results. Therefore, a proper implementation of both, standards and standard operating procedures (SOPs) is crucial if we want to overcome the problem. Many initiatives have emerged in recent years to provide standards and tools to their scientific communities ensuring reproducibility through a common framework. The basis to achieve such a common framework demands the agreement for a common language that could allow a successful interaction and cooperation. Besides the commonly used scientific jargon, it needs an agreement on a harmonised terminology and ontology for a successful implementation of a standards framework. Well known examples of existing standardisation in today’s life-science research are the usage of SI units in all publications as well as defined formats for data of nucleic acids (DNA and RNA) and protein sequences (FASTA format), and for three-dimensional (3D) structure of biological molecules such as the Protein Data Bank (PDB) format which provides a standard representation for macromolecular structure data derived from X-ray diffraction and NMR studies [23].

Hence, standards are the key for addressing and neutralizing the majority of problems related to the management and reuse of big data. There is a strong need for consensus agreements on measurements and stringent performance criteria when dealing with process, data and differences in definitions/terminology.

Data are the most valuable resource for investigating biological systems but their value is null if their formats do not allow sharing and integration from different sources. Interoperability is a must for omics disciplines. To answer to this pressing need, in 2005 the Research Data Management (RDM) initiative introduced a new set of principles for data management services. Following their principles, data should be FAIR – Findable, Accessible, Interoperable and Reusable [2,18]. Applying the FAIR data principles mainly addresses the metadata levels in research and does not necessarily take into account the quality of the source datasets. Hence, even if the datasets are published following the FAIR data principles, the quality of the data might be unsatisfactory. As a result, downstream calculations, analyses and proceedings based on such data might be questionable [7,17].

High-throughput technologies, such as Next Generation Sequencing (NGS), have turned the Life Sciences into a data-intensive discipline and require that the analysis of data is performed using high-performance computing resources. Researchers are now using informatics tools and computational models to decipher the biological information and predict the functioning of cells, organs and whole organisms. These approaches require the integration of data from different types of sources and of

different levels of biological information. To achieve interoperability is therefore mandatory and the only way to make data and resources available for their easy exchange and integration is the use of standards. Working across scales and (biological) systems demands now also the harmonisation of existing standards - including a common language - between particular fields or analytical technologies. But "...the lack of full standardisation hinders effective integration of results from cross-disciplinary collaboration studies" [17, 19]. This implies also to carry out data management and analysis tasks on large scale. One way to standardise such data analysis is by the use of bioinformatics Workflow systems that simplify and automatise the construction of analysis pipelines. Well documented and deposited on suitable databases these workflows can then support reproducibility and provide measures for fault-tolerance. Workflow systems for data analysis need therefore to be part of the documentation connected to any deposited data.

Harmonization and interfacing on the level of data formats and structures, descriptors and metadata represent just one side of the coin. The quality of the data provided is an issue of fundamental importance which, as we have described above, has not yet been resolved satisfactorily. The diversity of data sources precludes any straightforward and coherent strategy for maintaining and documenting the quality of the data. Data quality implies not only the fit for use of the data but also metrological traceability, repeatability, reproducibility, consistency and comparability. At best some but not all of the requirements are met by the prevailing data standards. More work is required to ensure that the experimental (including clinical) data underlying the computational models are not only appropriate for the context in which it is being used but is also of sufficient quality for that purpose. [8]

For raw data to be of value and of use, they must be both reliable and valid. Reliability refers to the repeatability of findings. Reliability also applies to individual measures starting with the experiments in every wet lab. To test the validity of instruments, procedures, or experiments, research may replicate elements of prior projects or the project as a whole.

A good starting point in standardisation measures would be the introduction of quality documentation of experiments which is frequently an obvious lack. Thus, it is crucial to develop and establish procedure-, operating- and inspection instructions as well as quality records. Furthermore, verification documents, particularly for providing a string of documents for the verifiable origin of data is an essential point. Especially the quality records could act as a certificate for potential users (customers) and the general documentation would improve the traceability and transparency with the aim to prove the reliability of results. Another important parameter in quality management (QM) considerations is the quality assurance (QA). A QA program should contain predetermined quality control (QC) checkpoints for monitoring QA and an extensive documentation including, among others, used devices, reagent lot numbers and any deviation from standard procedures (9;11). Moreover, for sequencing data, the QA program should contain QC methods for contamination identification at several stages within the sequencing workflow. These stages comprise the initial sample evaluation, the fragmentation step, the final library assessment, the monitoring of error rates during the sequencing process and the raw data analysis with a focus on reads quality (9;10).

Synopsis

To maximise the impact of research it is not sufficient to ask for OA alone. To this end, the availability of data via portal, platforms and repositories is indeed not enough. The principle of OA can only enter in force successfully, if the actors know and understand clearly what needs to be done at all stages towards them. Here, a huge backlog exists for measures in context with education and training at all connected levels. Validated research allows reproducible performing of experiments based on different techniques and technologies and this would serve in parallel as an indicator and a quality seal for reliable generated data. The idea behind education is that it is dynamic and continuous with questioning itself, its means and content. Education in life science gets stuck without providing validated and shareable research. Without, there is danger of misinterpretation, misuse and a drift towards fundamentalist dullness. Validation could become realised by trainers and "train the trainer" workshops which should be offered to all research data generators. Those trainings could be set as mandatory for certification issues, where the EU and EC should be responsible for encouraging and even enforcing standardisation and quality issues in research data generated within EU Member Country Institutions.

A huge obstacle nowadays to OA is antiquated mindsets regarding the claim of ownership of results. Although is necessary to deal with the three O (Open Access, Open Science and Open Innovation) in research projects, the idea of making results and data available for further use is still looked upon with distrust by many researchers and university leaders. Here several factors are contributing. There is a delicate balance to be struck between the freedom to preserve the autonomy of scientists on the one hand and the requirements of economic use on the other [2, 6, 20].

Opinions differ on this question by different stakeholders. The researcher or research group often claims the ownership of generated results because they were generated during their projects and are therefore intellectual property. On the other hand, the public claims free accessibility due to the nature of public funding. Furthermore, open data are subject to conflict of interest issues as far as the generated data are confidential and therefore linked to financial aspects.

In addition to the Open Access policy the European Commission also promotes Open Science as a new strategy. "Open Science represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools. The idea captures a systemic change to the way science and research have been carried out for the last fifty years: shifting from the standard practices of publishing research results in scientific publications towards sharing and using all available knowledge at an earlier stage in the research process." [26] This new policy of the European Commission now acknowledges that science and innovation are not restricted to the academic world and the industry, but also takes place in other societal groups. The European Commission intends to embrace these groups typically excluded from the research process and promotes their inclusion. Prerequisite is the openness of research data, results and scientific communication.

With this new policy of openness, we face a new challenge of ensuring quality of research across all scientific disciplines and actors. There will be a need to share and

implement standards, SOPs and Good Scientific Practice among all these groups involved.

Outlook/Perspective

The **initiative CHARME aims to harmonise** standardisation strategies to increase efficiency and competitiveness of European life-science research. The members of CHARME welcome the implementation of OA and the resulting possibility of sharing data and results. However, a general uploading of data is not sufficient; rather descriptions and detailed annotations are required to provide open science in order to accelerate the desired innovation processes. This includes the background behind the generation of data as well as the potency for interoperability and transferability between different data formats. At the metadata level huge efforts have been made in the past (Data Management Plan, FAIR Principle, etc.), whereas in context with the raw data hardly any efforts have been made.

We think, that there is a strong need for mechanisms of control for the quality of data which are openly accessible. This data check must be upstream of the open access.

A "seal of quality" similar to a DMP with clear definition of quality benchmarks for data is needed in order to define metrics which are applicable and reasonable for building a framework around good data quality [7] which than are unthinkingly usable for further proceeding by everyone.
This seal of quality should be supported by incentives by funder and publishers.
Incentives should also consider another important aspect, that is education to acceptance of QA plans. There are many examples of resistance of researchers to accept rules that QA plans impose and of how these resistances can be easily overcome by education [25]. To enlarge as much as possible the possibility for courses and implementation of these courses as part of university curricula is a crucial step forward the universalization of a safe and reliable way to make research.

We encourage interested stakeholder to join our discussion and to contribute to enabling the credibility of data and publications which are available from OA.

Acknowledgement

This publication is based upon work from COST Action CHARME (CA15110) supported by COST (European Cooperation in Science and Technology).
www.cost.eu

References

1. Understanding Science. 2018. University of California Museum of Paleontology. 3 January 2018 http://www.understandingscience.org
2. The transition towards an Open Science system-Council conclusions (adopted on 27/05/2016) 8791/16 RECH 133 TELECOM 74 http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf
3. Open Access in Deutschland https://www.bmbf.de/pub/Open_Access_in_Deutschland.pdf

4. Benefits of sharing. Nature. 530 (7589): 129–129. 2016-02-11. doi:10.1038/530129a.
5. Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research (CHARME) Cost Action CA15110 https://www.cost-charme.eu/the-action/about-charme
6. 'The transition towards an Open Science system' – Council conclusions adopted on 27 May 2016, http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf
7. Nickerson, D.; Atalag, K; de Bono, B.; Geiger, J.; Goble, C.; Hollmann, S.; Lonien, J.; Müller,W.; Regierer, B.; Stanford, N.J.; Golebiewski, M.; Hunter, P. The Human Physiome: how standards, software and innovative service infrastructures are providing the building blocks to make it achievable. The Royal Society Publishing, Interface Focus 2016 6 20150103; DOI: 10.1098/rsfs.2015.0103. Published 19 February 2016
8. Stuart D. et al.: Whitepaper: Practical challenges for researchers in data sharing, (2018). DOI: https://doi.org/10.6084/m9.figshare.5975011.v1
9. Rehm, H.L., Bale, S.J., Bayrak-Toydemir, P., Berg, J.S., Brown, K.K., Deignan, J.L., Friez, M.J., Funke, B.H., Hegde, M.R., Lyon, E.,Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Commtitee, 2013. ACMGclinical laboratory standards for next-generation sequencing. Genet. Med. 15, 733–747. http://dx.doi.org/10.1038/gim.2013.92.
10. Endrullat, C.; Glöckler, J.; Franke, P.; Frohme, M. Standardization and quality management in next-generation sequencing, Applied & Translational Genomics, Vol 10, September 2016, Pages 2-9 https://doi.org/10.1016/j.atg.2016.06.001
11. Gargis, A.S., Kalman, L., Berry,M.W., Bick, D.P., Dimmock, D.P., Hambuch, T., Lu, F., Lyon, E.,Voelkerding, K.V., Zehnbauer, B.A., Agarwala, R., Bennett, S.F., Chen, B., Chin, E.L.H.,Compton, J.G., Das, S., Farkas, D.H., Ferber, M.J., Funke, B.H., Furtado, M.R., Ganova-Raeva, L.M., Geigenmüller, U., Gunselman, S.J., Hegde, M.R., Johnson, P.L.F.,Kasarskis, A., Kulkarni, S., Lenk, T., Liu, C.S.J., Manion, M., Manolio, T.A., Mardis, E.R.,Merker, J.D., Rajeevan, M.S., Reese, M.G., Rehm, H.L., Simen, B.B., Yeakley, J.M., Zook,J.M., Lubin, I.M., 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat. Biotechnol. 30, 1033–1036. http://dx.doi.org/10.1038/nbt.2403
12. OECD (2015), Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris. Retrieved on 27 October 2017
13. Nancy Baron, Escape from the Ivory Tower ISBN: 9781597266635 Pub Date: 8/2010
14. Docherty AB, Lone NI. Exploiting big data for critical care research. Curr Opin Crit Care. 2015 Oct;21(5):467-72. doi: 10.1097/MCC.0000000000000228.
15. Source:Definition of 'science' https://www.collinsdictionary.com/dictionary/english/science
16. Smalter Hall, A.; Shan, Y.; Lushington,G.; Visvanathan, M. An overview of computational life science databases & exchange formats of relevance to

chemical biology research. Comb Chem High Throughput Screen. 2013 March ; 16(3): 189–198.
17. Freedman, L.P.; Venugopalan, G.; Wisman, R. Reproducibility 2020: Progress and priorities [version 1; referees: 2 approved]. *F1000Research* 2017, **6**:604 (doi: 10.12688/f1000research.11334.1)
18. Sansone, S.; Mcquilton, P.; Rocca-serra, P.; Gonzalez-beltran, A.; Izzo, M.; Lister, A.; Thurston, M. FAIRsharing: Working with and for the Community to Describe and Link Data Standards , Repositories and Policies. *bioRxiv* , No. 1, 0–9. (2018)
19. Ellingsen,G.; Monteiro, E. Seamless Integration: Standardisation across Multiple Local Settings, Computer Supported Cooperative Work (2006) 15:443–466 DOI 10.1007/s10606-006-9033-0
20. OECD Principles and guidelines for access to research data from public funding – OECD 2007 http://www.oecd.org/sti/sci-tech/38500813.pdf
21. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci Data, 3, 160018.(2016) Winkler-Nees, http://www.ariadne.ac.uk/issue64/datacite-2010-rpt
22. European Commission EUR 22836 - Improving knowledge transfer between research institutions and industry across Europe: embracing open innovation. 2007 ISBN 978-92-79-05521 http://ec.europa.eu/invest-in-research/pdf/download_en/knowledge_transfe_07.pdf
23. Nature Structural Biology **10**, 980 (2003) doi: 10.1038/nsb1203-980
24. Impact of Emerging Technologies on the Biological Sciences, Report of a Workshop,June 26-27, 1995, https://nsf.gov/bio/pubs/reports/stctechn/stcmain.htm
25. Monya Baker (2016) How quality control could save your science. Nature News 529, 456–458
26. Directorate-General for Research and Innovation ISBN 978-92-79-65567-8 doi:10.2777/380389