2021-07-12

Towards a Common Standard for Data and Specimen Provenance in Life Sciences [preprint]

Petr Holub
BBMRI-ERIC

Et al.

Towards a Common Standard for Data and Specimen Provenance in Life Sciences

Petr Holub*[1], Rudolf Wittner[1,2], Cecilia Mascia[3], Francesca Frexia[3], Heimo Müller[4], Markus Plass[4], Clare Allocca[5], Fay Betsou[6], Tony Burdett[7], Ibon Cancio[8], Adriane Chapman[9], Martin Chapman[10], Mélanie Courtot[7], Vasa Curcin[10], Johann Eder[11], Mark Elliot[12], Katrina Exter[13], Elliot Fairweather[10], Carole Goble[14], Martin Golebiewski[15], Bron Kisler[16], Andreas Kremer[17], Sheng Lin-Gibson[18], Anna Marsano[19], Marco Mattavelli[20], Josh Moore[21], Hiroki Nakae[22], Isabelle Perseil[23], Ayat Salman[24,25], James Sluka[26], Stian Soiland-Reyes[14,27], Caterina Strambio-De-Castillia[28], Michael Sussman[29], Jason R. Swedlow[21], Kurt Zatloukal[4], and Jörg Geiger[30]

[1] *BBMRI-ERIC, Graz, AT*
[2] *Institute of Computer Science & Faculty of Informatics, Masaryk University, Brno, CZ*
[3] *CRS4 – Center for Advanced Studies, Research and Development in Sardinia, IT*
[4] *Medical University Graz, AT*
[5] *National Institute of Standards and Technology, Gaithersburg, MD, USA*
[6] *LNS - Laboratoire national de santé, LU*
[7] *EMBL's European Bioinformatics Institute (EMBL-EBI), UK*
[8] *Plentzia Marine Station (PiE-UPV/EHU), University of the Basque Country, EMBRC-Spain*
[9] *University of Southampton, UK*
[10] *King's College London, UK*
[11] *University of Klagenfurt, AT*
[12] *Department of Social Statistics, School of Social Sciences, The University of Manchester, UK*
[13] *Flanders Marine Institute (VLIZ), EMBRC-Belgium*
[14] *Department of Computer Science, The University of Manchester, UK*
[15] *Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, DE*
[16] *Independent consultant*
[17] *ITTM S.A., LU*
[18] *Biosystems and Biomaterials Division, NIST, USA*
[19] *Department of Biomedicine, The University of Basel, CH*
[20] *SCI-STI-MM, École Politechnique Fédérale de Lausanne, Lausanne, CH*
[21] *Centre for Gene Regulation and Expression and Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK*
[22] *Japan bio-Measurement and Analysis Consortium, JPN*
[23] *INSERM - Institut National de la Sante et de la Recherche Medicale, FR*
[24] *Standards Council of Canada*
[25] *Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Department of Family Medicine, Queen's University, CA*
[26] *Biocomplexity Institute, Indiana University, USA*

[27] *Informatics Institute, University of Amsterdam, NL*
[28] *Program in Molecular Medicine, UMass Medical School, USA*
[29] *US Department of Agriculture, USA*
[30] *Interdisciplinary Bank of Biomaterials and Data Würzburg (ibdw),Würzburg, DE*

The profound crisis of scientific reproducibility has its roots in the enhanced availability of large volumes of data that are produced at ever increasing velocity, which in turn often leads to the dissolution of the control mechanisms that traditionally ensured the quality of data and processes [1–7]. At the same time the origin and history of specimens used to generate research data often remains inexplicit. While considerable effort has been put in the development of standards for specimen quality, the actual documentation has been left to the discretion of the provider of the specimen. As a result the situation is exacerbated by the lack of consistent and comprehensive documentation of specimens, which could support the identification of suspected, or proven use of, fabricated data or specimen of unclear origin. Hence, the urgent need for the trustworthy documentation of the data lineage and specimens is evident, especially when considering the serious impact of irreproducible or even flawed scientific results on health, economics, and political decisions [8–12].

It is generally accepted that the properties and quality attributes of specimens used in the life sciences have significant impact on the reliability of data generated in downstream analytical procedures [13–15]. Experts from multiple life sciences domains have called for the improvement and standardization of the documentation of research and scientific service processes [16–22]. This has led in turn to the progressive development and implementation of data management and other functional tools, such as discovery services, access pipelines, and standardized data models, enabling the sharing of data and specimens [23–28]. In practice, however, there remains a gap between the needs and the reality of the requirements specified in accepted standards, including technical, operational and legal specifications needed to ensure the trustworthiness and traceability of data and specimens. Electronic lab notebooks (ELN) and laboratory information management systems (LIMS) adopted by research organizations might be considered attempts to electronically manage research workflows and data to promote reproducibility and traceability. However, these systems can not provide the degree of standardization an international standard would offer, as they are often proprietary and not subject to certification. In an effort to remedy these deficiencies in the provenance captured and reported, we are endeavoring to develop an *international standard on provenance information system for the life sciences* accepted by both academia and industry. Provenance information can be used to assess the quality and reliability, and hence the reusability of the object, i.e. the data, the metadata, the biological materials, or the specimens.

The need for an effort to address the issues in provenance was proposed to the International Standards Organization (ISO) Technical Committee 276 "Biotechnology" (ISO/TC 276) in 2017 and approved as a preliminary work item. In 2020, ISO/TC 276

*Corresponding author: Assoc. Prof. RNDr. Petr Holub, Ph.D., Neue Stiftingtalstr. 2/B/6, 8010 Graz, AT. Email: <petr.holub@bbmri-eric.eu>

approved a new work item proposal to develop an international standard for biological material and data provenance and registered it as a working draft (WD), ISO/WD 23494-1 *Provenance information model for biological specimen and data — Part 1: General requirements*. This standardization effort is in accordance with the FAIR principles, which provide high-level methodological recommendations, including guidance on provenance.[1] As the FAIR principles themselves do not provide detailed instructions for the implementation of provenance standards and documentation, ISO/WD 23494 is intended for data provenance of biological samples and will be built on the World Wide Web Consortium's (W3C) PROV [29], a generic provenance information standard that defines a general model, corresponding serializations[2] and other supporting specifications to enable the interoperable exchange of provenance information between data environments. W3C PROV serves as a framework that is adaptable and extensible to fit the needs of diverse domains. The W3C PROV standard has already been adopted in life science research areas [30], e.g., for computational workflows [31], pharmacologic pipelines [32], neuroscience [33, 34], microscopy experiments [35], medical sciences [36] and health implementation care[3] in HL7 FHIR [37]. Unfortunately, these implementations occurred without coordination and the resulting solutions are often incompatible, incomplete, expressed at different levels of granularity, and do not use a consistent approach for creating a continuous chain of provenance from the "source" to the resulting data. Instead of redefining the W3C PROV concepts, we have identified gaps that need to be filled in order to develop a distributed, fully technically and semantically interoperable provenance information standard that covers a given specimen and its associated metadata, and describes its uninterrupted history from its "source". The "source" can include a complex, multi-institutional environment and can be both the source specimen and data, but also link to a specific biological entity, or environmental specimen collected at a given time and location (*connectivity* requirement [38]). The main goals of the provenance information standard are

(i) to support improved reproducibility of life-sciences research, to provide a voluntary provenance framework enabling concordance of governments, businesses, academia and the international community

(ii) to achieve harmonization of documentation of specimens that is compliant with international conventions, recognized ethical practices and legal requirements such as the Nagoya Protocol [39] and the Declaration of Taipei [40].

(iii) to enable decision-making about the fitness-for-purpose of particular specimens and data, by collecting and linking provenance information from the whole life-cycle of the object (from specimen collection and processing, through data generation and analysis) as depicted in Figure 1.

The standard will enhance the trustworthiness of provenance information by including requirements and guidelines on its integrity, authenticity, and non-repudiation [41], to prevent the production and/or use of unreliable, flawed or fabricated data

[1] Principle R1.2: (Meta)data are associated with detailed provenance.

[2] As defined in ISO 21597-1:2020: encoding of an ontology or dataset into a format that can be stored, typically in a file.

[3] `https://www.hl7.org/fhir/provenance.html`

(the potential harms of which have become evident during the COVID-19 pandemic [2, 10]), as well as accidental or malicious modification of data. Since provenance information may also include sensitive or personal data (related, e.g., to the health condition of an individual), the standard aims to enable sensitive information to be concealed and disclosed only under strictly controlled conditions, while preserving its core properties of integrity, authenticity and non-repudiation. Additional advanced application scenarios include tracking of provenance information to: (i) track research error propagation, (ii) identify people affected by incidental research findings, (iii) check compliance with applicable regulations, or (iv) support production of reference material by maintaining full documentation of provenance (complementing work of ISO/TC 334 [42]). For research concerned with highly regulated fields in life sciences, such as development of medical products or drugs, the standardized provenance model will also contribute to a level of accountability and auditability of research organisations.

The proposed standard is designed to cover the majority of the organizations involved in life-sciences research, both academic and industrial, government labs and research centers. Included organizations are university and industrial research laboratories, biobanks and biorepositories, culture collections, hospitals, research centres, and private companies (e.g., pharmaceutical companies or lab reagent suppliers). The broader audience includes not only research data producers, but also those publishing, cataloguing, archiving or reusing research data [43]. The standard can also be adopted by manufacturers and vendors of laboratory instruments – e.g., automation devices, microscopes, sequencers, spectrometers – to enable automated standard-compliant generation of provenance information. Automated generation of provenance information will minimize human errors and the burden put on workers, both in terms of effort and training. Provenance information generated automatically by devices should be interoperable to enable automated integration and quality control as well as validity checks demonstrating standard-compliant provenance. The standard is intended to cover a wide range of research and applications in life sciences and for that reason a modular structure has been used to enable extensibility to evolving requirements, processes, or technologies.

The current draft proposal ISO/WD TS 23494 1 is the first part of a planned series of six parts, with the intent that each will become a distinct ISO standard:

1. *Provenance Information Management* defines the overall structure of the standard and provides general requirements on provenance information management, thus enabling interconnections between the various components of provenance information in distributed environments. It also specifies requirements applicable to entities responsible for generating the provenance information.

2. The *Common Provenance Model* builds on the W3C PROV model, defining representations of elements common to all stages of research, such as interlinking of distributed components of provenance information by sender and receiver objects, the identification of physical and digital objects, and provisions for non-repudiation. Provenance information patterns for common scenarios, such as the compound processes, versioning of provenance information or documen-

tation of accountabilities. The model will also define mechanisms to embed or reference entire records of provenance information.

3. *Provenance of Biological Materials* defines requirements and scope of the provenance information documenting biological material or specimen acquisition, handling and processing and builds on the Common Provenance Model. This includes, but is not limited to, data on collection and collection procedure, transport conditions, and documentation of legal and ethical basis (e.g. consent, terms of access and benefit sharing). It will also provide mechanisms to reference Standard Operating Procedures (SOPs) and compliance with or deviations from them. Referencing the widely accepted de-facto reporting standard for biological specimen quality SPREC [44] will also be enabled. Actual techniques or practices for handling biological material are not specified in the standard, in favor of technical specifications enabling consistent interoperable and machine-actionable documentation of handling biological material. With the provenance information provided, however, the standard facilitates the verification of compliance with other pre-analytical ISO standards covering biobanking, analytical and processing methods, generation of reference material and related fields (ISO 20387:2018, ISO 20184 series, ISO 20166 series, and ISO 20186 series).

4. *Provenance of Data Generation* defines the provenance of data generated from the analysis or observation of biological material, e.g., sequencing, microscopy, spectrometry, etc. Provenance information specific for diverse analytical or observational methods will be embedded in a way meeting the requirements of the particular domain, but as well compliant with the provenance model standard allowing seamless integration in a complete provenance chain.

5. *Provenance of Data Processing* defines provenance of computational aspects of life sciences research (such as computational workflows based on CWLProv [31] and RO Crate [45]).

6. *Security Extensions* define optional extensions supporting authenticity, integrity and non-repudiation of provenance information, and hence its trustworthiness and reliability. Demonstration of these properties will also be supported for sensitive elements of provenance information.

The ISO standards development process responds to a market need and is based on globally-relevant expertise. The product is a voluntary consensus standard developed through a multi-stakeholder process. ISO/WD 23494-1 has a proven market need and has passed through the preliminary stages of the ISO voting process – as a result, it is part of the ISO Work Programme. The document is under development and will evolve along the multi-stage ISO standard development process. ISO/WD 23494-1 *Provenance information model for biological specimen and data – Part 1: General requirements* is currently at the working draft stage, and is anticipated to move next to the committee draft (CD) stage. The document will be revised and reviewed throughout the ISO project stages until final approval and publication. Part 2 of this series, *Biotechnology — Provenance information model for biological material and data*

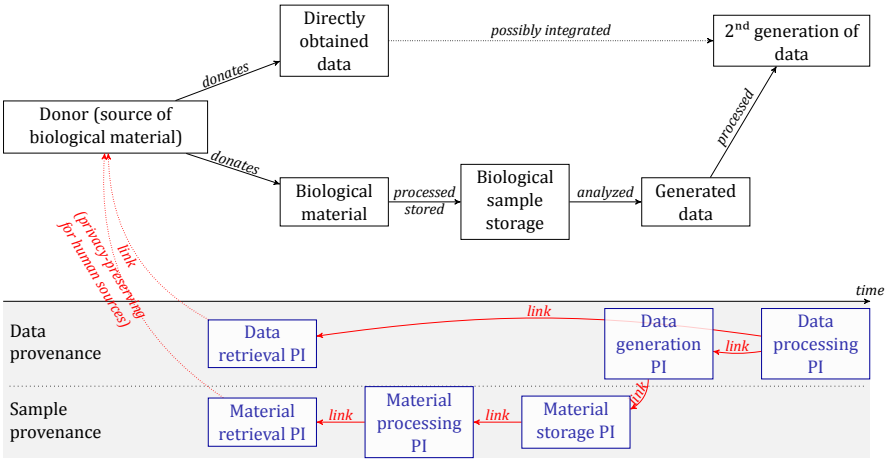

Figure 1: An overview of provenance chain. A sample obtained from a donor (or other source) is created and an initial set of provenance information (PI) is generated. As that sample moves through time and space, is processed and/or analyzed, additional provenance data is appended to the provenance chain for each new item. The chain can be extended as a complete unit of later stages of provenance or use unique identifiers to refer to early stages of provenance data.

— *Part 2: Common provenance model*, has been accepted by ISO/TC 276/WG 5 as preliminary work item ISO/PWI TS 23494-2 . The future documents in this series are in planning stages, but not yet submitted to ISO/TC 276/WG 5 . The standards development process builds on existing standards for collection and processing of specimens, analytical techniques and data generation and analysis, as well as use-cases from the biomedical domain. BBMRI-ERIC, which is also active in developing international standards for biobanking, has drafted use-cases for biological material provenance. Collaborations and ISO liasions with professional societies like the European, Middle Eastern and African Society for Biobanking (ESBB) and the International Society for Biological and Environmental Repositories (ISBER) have also contributed to the development of specimen provenance use cases. In addition, use cases on data generation and processing can come from subject matter experts and the scientific community including the European EOSC-Life project,[4] Open Microscopy Environment, OME,[5] genetic data compression (ISO/IEC JTC1/SC 29/WG 08 MPEG-G) [46], clinical decision support systems (Kings College London) and other life sciences domains such as biodiversity, marine biology and systems biology.

However, alternatives to ISO standards process[6] exist—some community-based efforts have developed widely adopted specifications that have become *de facto* global

[4] `https://www.eosc-life.eu/`

[5] `https://www.openmicroscopy.org/`

[6] `https://www.iso.org/developing-standards.html`

standards .[7] The success of these examples lies, at least in part, in the pairing of a specification with an accessible implementation that validates the utility of the specification and allows a broad community to explore integration into applications that extend far beyond the initial target [50]. We believe that community-led and ISO-based approaches for developing and delivering standards can complement each other and that a combination of parallel efforts for developing a provenance chain standard might ultimately be the most productive approach. As the provenance information model development is grounded in the EOSC-Life project, collaboration with these communities is already established. The ISO standard development is thus considered as a standardized instance of a publicly available model developed in parallel under auspices of EOSC-Life [51].

Another challenge is the continuous dissemination and periodic revision of the standard once published. Though ISO standards are not "open access", they can be purchased for a moderate fee[8] or accessed through institutional libraries, and, barring any patent restrictions, can be freely implemented, for instance, in Open Source software. ISO standards can also include Open Source reference implementations as specific normative or informative parts of the standards. ISO standards can be implemented independently or based on such source code, in compliance with the reasonable and non-discriminatory (RAND) licensing terms imposed by the ISO requirements. Such licensing terms, like for instance the one applied to all ISO/IEC/SC29 (MPEG) standards that are free from any charge for scientific and non-profit research purposes, may or may not include licensing fees.

We would like to invite experts from biotechnology and biomedical fields to further contribute to the standard, in particular to the provenance of biological specimens, the data-generation and data-processing modules. Help is needed to develop applications of the general modules and the development of specific use cases, as well as direct contributions to the text of the standard itself. Contributions are possible through a liaison organization, a national ISO body or by engaging with EOSC-Life project events and calls.

Acknowledgments This work has been co/funded by EOSC-Life supported by EU Horizon 2020, grant agreement no. 824087; EJP-RD supported by EU Horizon 2020, grant agreement no. 825575; BioExcel-2 supported by EU Horizon 2020, grant agreement no. 823830; the DIFRA Project, funded by the Sardinian Regional Authority. VC, EF, and MCh supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' National Health Service NHS) Foundation Trust and King's College London. TB, MCo acknowledge funding from EMBL-EBI Core Funds and the FAIRplus project (H2020 No 802750) . MCo supported by Wellcome Trust GA4GH award number 201535/Z/16/Z and the CINECA project (H2020 No 825775). AC was supported by EPSRC (EP/S028366/1). JS was supported by the US National institute of Health (U24 EB028887, R01 GM122424, and OT2OD026671). ME was supported by the Alan Turing Institute (ProvAnon). The

[7] E.g., for on-line cryptography (RSA public keys [47]), scientific workflows (Common Workflow Language [48]) and bioimaging data formats (OME-TIFF [49]).

[8] In some cases ISO standards can be obtained without any fee, e.g. `https://www.iso.org/covid19`

opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders.

Representation of communities The co-authors team represents wide coverage of life-sciences communities. PH, RW, CM, FF, HM, MP, JG come from human biobanking and biomolecular resources communities, BBMRI-ERIC Research Infrastructure, and are directly involved as experts in the ISO standardization process. KZ and JE come from cancer research, biobanking and medical informatics and are long-term contributors to data quality standardization efforts. TB, MCo lead development of the BioSamples database at EMBL-European Bioinformatics Institute. IC and KE come from marine biology and EMBRC Research Infrastructure. CG and SS-R have worked with bioinformatics, CWL, RO-Crate and the original W3C PROV standards developments. JRS and JM come from bio-imaging communities and EUBioImaging Research Infrastructure. VC, EF, and MCh come from health informatics. HN participates in provenance standardization process as an expert from Japan, MS and JS as experts from the U.S.A, and AK as an expert from Luxembourg. ME contributes to privacy protection and provenance aspects. FB is a biobanking expert and chairing the ISBER Biospecimen Science Working Group. AS is a biobanking expert and ESBB councillor. SL-G and CA are from NIST and convenor and secretary of ISO/TC 276/WG 3 "Analytical Methods". AM belongs to the tissue engineering and biomedical research community. MM is a standard expert in the digital media, genomic sequencing and annotation data fields, and convenor of ISO/IEC SC29/WG 8 "MPEG Genomic Coding". AC contributes to capture and handling of provenance within large organizations.

References

1. Begley CG and Ioannidis JP. Reproducibility in Science. Circulation Research 2015;116:116–26. DOI: 10.1161/CIRCRESAHA.114.303819.
2. Servick K and Enserink M. The pandemic's first major research scandal erupts. Science 2020;368:1041–2. DOI: 10.1126/science.368.6495.1041.
3. Lagoze C. Big Data, data integrity, and the fracturing of the control zone. Big Data & Society 2014;1:2053951714558281. DOI: 10.1177/2053951714558281.
4. Mobley A, Linder SK, Braeuer R, et al. A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. PLOS ONE 2013;8:1–4. DOI: 10.1371/journal.pone.0063221.
5. Morrison SJ. Time to do something about reproducibility. eLife 2014;3:1–4. DOI: 10.7554/eLife.03981.
6. Byrne JA, Grima N, Capes-Davis A, et al. The Possibility of Systematic Research Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Potential Solutions. Biomarker Insights 2019;14. DOI: 10.1177/1177271919829162.
7. Prinz F, Schlange T, and Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 2011;10. Number: 9 Publisher: Nature Publishing Group:712–2. DOI: 10.1038/nrd3439-c1.
8. Freedman LP, Cockburn IM, and Simcoe TS. The Economics of Reproducibility in Preclinical Research. PLOS Biology 2015;13:1–9. DOI: 10.1371/journal.pbio.1002165.
9. Nickerson D, Atalag K, Bono B de, et al. The Human Physiome: how standards, software and innovative service infrastructures are providing the building blocks to make it achievable. Interface Focus 2016;6. 00001:20150103. DOI: 10.1098/rsfs.2015.0103. URL: http://rsfs.royalsocietypublishing.org/lookup/doi/10.1098/rsfs.2015.0103 (visited on 03/10/2016).
10. Mahase E. Covid-19: 146 researchers raise concerns over chloroquine study that halted WHO trial. BMJ 2020;369. DOI: 10.1136/bmj.m2197.
11. Chaplin S. Research misconduct: how bad is it and what can be done? Future Prescriber 2012;13:5–76. DOI: 10.1002/fps.88.
12. Committee on Responsible Science, Committee on Science, Engineering, Medicine, and Public Policy, Policy and Global Affairs, et al. Fostering Integrity in Research. Pages: 21896. Washington, D.C.: National Academies Press, 2017. DOI: 10.17226/21896.
13. Simeon-Dubach D and Perren A. Better provenance for biobank samples. Nature 2011;475:454–5. DOI: 10.1038/475454d.
14. Holub P, Kohlmayer F, Prasser F, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. Biopreservation and Biobanking 2018;16:97–105. DOI: 10.1089/bio.2017.0110.

15. Müller H, Reihs R, Zatloukal K, et al. State-of-the-Art and Future Challenges in the Integration of Biobank Catalogues:13. DOI: 10.1007/978-3-319-16226-3_11.

16. Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. The Lancet 2014;383:166–75. DOI: 10.1016/S0140-6736(13)62227-8.

17. Freedman LP and Inglese J. The Increasing Urgency for Standards in Basic Biologic Research. Cancer Research 2014;74:4024–9. DOI: 10.1158/0008-5472.CAN-14-0925.

18. Begley CG and Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature 2012;483:531–3. DOI: 10.1038/483531a. arXiv: 9907372v1.

19. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 2012;490. nature11556[PII]:187–91. DOI: 10.1038/nature11556.

20. Consortium of European Taxonomic Facilities (CETAF) Code of Conduct and Best Practice for Access and Benefit-Sharing. URL: https://ec.europa.eu/environment/nature/biodiversity/international/abs/pdf/CETAF%20Best%20Practice%20-%20Annex%20to%20Commission%20Decision%20C(2019)%203380%20final.pdf (visited on 02/15/2021).

21. Benson EE, Harding K, and Mackenzie-dodds J. A new quality management perspective for biodiversity conservation and research: Investigating Biospecimen Reporting for Improved Study Quality (BRISQ) and the Standard PRE-analytical Code (SPREC) using Natural History Museum and culture collections as case studies. Systematics and Biodiversity 2016;14:525–47. DOI: 10.1080/14772000.2016.1201167.

22. A-E. K and Tillin H. The EMBRC guide to ABS compliance. Recommendations to marine biological resources collections' and users' institutions. A handbook produced by the European Marine Biological Resource Centre. European Marine Biological Resource Centre. 2020. URL: https://bluebiobank.eu/docs/EMBRCGuideABS.pdf.

23. Villanueva AG, Cook-Deegan R, Koenig BA, et al. Characterizing the Biomedical Data-Sharing Landscape. The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics 2019;47:21–30. DOI: 10.1177/1073110519840481.

24. Hulsen T. Sharing Is Caring-Data Sharing Initiatives in Healthcare. International Journal of Environmental Research and Public Health 2020;17:E3046. DOI: 10.3390/ijerph17093046.

25. Banzi R, Canham S, Kuchinke W, et al. Evaluation of repositories for sharing individual-participant data from clinical studies. Trials 2019;20:169. DOI: 10.1186/s13063-019-3253-3.

26. Toh S. Analytic and Data Sharing Options in Real-World Multidatabase Studies of Comparative Effectiveness and Safety of Medical Products. Clinical Pharmacology and Therapeutics 2020;107:834–42. DOI: 10.1002/cpt.1754.

27. Grossman RL. Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data. Trends in genetics: TIG 2019;35:223–34. DOI: 10.1016/j.tig.2018.12.006.

28. Wilson SL, Way GP, Bittremieux W, et al. Sharing biological data: why, when, and how. FEBS Letters 2021;595. _eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1002/1873-3468.14067:847–63. DOI: 10.1002/1873-3468.14067. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/1873-3468.14067 (visited on 06/22/2021).

29. Groth P and Moreau L. PROV-Overview: An Overview of the PROV Family of Documents. 2013. URL: https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

30. Huynh TD, Groth P, and Zednik S. PROV Implementation Report. 2013. URL: http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/.

31. Khan FZ, Soiland-Reyes S, Sinnott RO, et al. Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. GigaScience 2019;8. giz095. DOI: 10.1093/gigascience/giz095.

32. Mammoliti A, Smirnov P, Safikhani Z, et al. Creating reproducible pharmacogenomic analysis pipelines. Scientific Data 2019;6:166. DOI: 10.1038/s41597-019-0174-7.

33. McClatchey R, Shamdasani J, Branson A, et al. Traceability and Provenance in Big Data Medical Systems. In: *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*. 2015:226–31. DOI: 10.1109/CBMS.2015.10.

34. Giesler A, Czekala M, Hagemeier B, et al. UniProv: A Flexible Provenance Tracking System for UNICORE. In: *High-Performance Scientific Computing*. Ed. by Di Napoli E, Hermanns MA, Iliev H, et al. Cham: Springer International Publishing, 2017:233–42. DOI: 10.1007/978-3-319-53862-4_20.

35. Samuel S. Integrative Data Management for Reproducibility of Microscopy Experiments. In: *The Semantic Web*. Ed. by Blomqvist E, Maynard D, Gangemi A, et al. Cham: Springer International Publishing, 2017:246–55. DOI: 10.1007/978-3-319-58451-5_19.

36. Curcin V, Fairweather E, Danger R, et al. Templates as a method for implementing data provenance in decision support systems. Journal of Biomedical Informatics 2017;65:1–21. DOI: 10.1016/j.jbi.2016.10.022.

37. HL7 and its participants. FHIR Release #4 [Standard], version 4.0.1. 2019. URL: http://hl7.org/fhir/R4/.

38. Curcin V, Miles S, Danger R, et al. Implementing interoperable provenance in biomedical research. Future Generation Computer Systems 2014;34. Special Section: Distributed Solutions for Ubiquitous Computing and Ambient Intelligence:1–16. DOI: 10.1016/j.future.2013.12.001.

39. Unit B. The Nagoya Protocol on Access and Benefit-sharing. Publisher: Secretariat of the Convention on Biological Diversity. 2021. URL: https://www.cbd.int/abs/ (visited on 06/22/2021).

40. WMA - The World Medical Association-WMA Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks. URL: https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/ (visited on 07/02/2020).

41. Fairweather E, Wittner R, Chapman M, et al. Non-repudiable provenance for clinical decision support systems. preprint. 2020. arXiv: 2006.11233 [cs.CR].

42. 14:00-17:00. ISO/WD Guide 85. URL: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/55/75538.html (visited on 06/09/2021).

43. Cheney J, Chapman AP, Davidson J, et al. Data provenance, curation and quality in metrology. 2021. arXiv: arXiv:2102.08228v1.

44. Betsou F, Bilbao R, Case J, et al. Standard PREanalytical Code Version 3.0. Biopreservation and Biobanking 2018;16:9–12. DOI: 10.1089/bio.2017.0109. (Visited on 02/18/2020).

45. Carragáin EÓ, Goble C, Sefton P, et al. A lightweight approach to research object data packaging. In: *Bioinformatics Open Source Conference (BOSC) 2019*. 2019. DOI: 10.5281/zenodo.3250687.

46. Voges J, Hernaez M, Mattavelli M, et al. An Introduction to MPEG-G: The First Open ISO/IEC Standard for the Compression and Exchange of Genomic Sequencing Data. Proceedings of the IEEE 2021. DOI: 10.1109/JPROC.2021.3082027.

47. Rivest RL, Shamir A, and Adleman L. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Commun. ACM 1978;21:120–6. DOI: 10.1145/359340.359342. URL: https://doi.org/10.1145/359340.359342.

48. Crusoe MR, Abeln S, Iosup A, et al. Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language. 2021. arXiv: 2105.07028v1 [cs.DC].

49. Linkert M, Rueden CT, Allan C, et al. Metadata matters: access to image data in the real world. Journal of Cell Biology 2010;189:777–82. DOI: 10.1083/jcb.201004104. URL: https://doi.org/10.1083/jcb.201004104.

50. Swedlow JR, Kankaanpää P, Sarkans U, et al. A global view of standards for open image data formats and repositories. Nature Methods 2021. DOI: 10.1038/s41592-021-01113-7. URL: https://doi.org/10.1038/s41592-021-01113-7.

51. Wittner R, Mascia C, Frexia F, et al. EOSC-Life Common Provenance Model. EOSC-Life deliverable D6.2. 2021. DOI: 10.5281/zenodo.4705074.