Received: 29 December 2022 | Revised: 17 March 2023 | Accepted: 24 March 2023

DOI: 10.1002/lrh2.10365

COMMENTARY

Learning Health Systems

Toward a common standard for data and specimen provenance in life sciences

Rudolf Wittner[1,2] | Petr Holub[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[11] | Vasa Curcin[10] | Johann Eder[12] | Mark Elliot[13] | Katrina Exter[14] | Carole Goble[15] | Martin Golebiewski[16] | Bron Kisler[17] | Andreas Kremer[18] | Simone Leo[3] | Sheng Lin-Gibson[19] | Anna Marsano[20] | Marco Mattavelli[21] | Josh Moore[22,23] | Hiroki Nakae[24] | Isabelle Perseil[25] | Ayat Salman[26,27] | James Sluka[28] | Stian Soiland-Reyes[15,29] | Caterina Strambio-De-Castillia[30] | Michael Sussman[31] | Jason R. Swedlow[22] | Kurt Zatloukal[4] | Jörg Geiger[32]

Correspondence
Rudolf Wittner, BBMRI-ERIC, Graz, Austria.
Email: wittner@ics.muni.cz

Funding information
Alan Turing Institute; Bundesministerium für Bildung, Wissenschaft und Forschung (Federal Ministry of Education, Science and Research of Austria), Grant/Award Number: BMBWF-10.470/0010-V/3c/2018; Chan Zuckerberg Initiative, Grant/Award Number: 2019-198155; Engineering and Physical Sciences Research Council, Grant/Award Number: EP/S028366/1; Horizon 2020 Framework Programme, Grant/Award Numbers: 802750, 823830, 824087, 825575, 825775; National Institutes of Health, Grant/Award Number: U01CA200059; National Science Foundation, Grant/Award Number: NSF 2054061; National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' National Health Service (NHS) Foundation Trust, Grant/Award Number: RJ112/N027; Sardinian Regional Authority; U.S. Environmental Protection Agency, Grant/Award Number: RD840027; US National institute of Health, Grant/Award Numbers: OT2OD026671, R01 GM122424, U24 EB028887; Wellcome Trust GA4GH, Grant/Award Number: 201535/Z/16/Z; NIHR Application Research Collaboration South London

Abstract

Open and practical exchange, dissemination, and reuse of specimens and data have become a fundamental requirement for life sciences research. The quality of the data obtained and thus the findings and knowledge derived is thus significantly influenced by the quality of the samples, the experimental methods, and the data analysis. Therefore, a comprehensive and precise documentation of the pre-analytical conditions, the analytical procedures, and the data processing are essential to be able to assess the validity of the research results. With the increasing importance of the exchange, reuse, and sharing of data and samples, procedures are required that enable cross-organizational documentation, traceability, and non-repudiation. At present, this information on the provenance of samples and data is mostly either sparse, incomplete, or incoherent. Since there is no uniform framework, this information is usually only provided within the organization and not interoperably. At the same time, the collection and sharing of biological and environmental specimens increasingly require definition and documentation of benefit sharing and compliance to regulatory requirements rather than consideration of pure scientific needs. In this publication, we present an ongoing standardization effort to provide trustworthy machine-actionable documentation of the data lineage and specimens. We would like to invite experts from the biotechnology and biomedical fields to further contribute to the standard.

For affiliations refer to page 6

KEYWORDS
biotechnology, International Organization for Standardization, provenance information, standardization

1 | INTRODUCTION

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.[1-11] At the same time, the origin and history of specimens used to generate research data often remain inexplicit. While considerable effort has been put into the development of standards for specimen quality, the actual documentation has been left to the discretion of the provider of the specimen and data. As a result, the situation is exacerbated by the lack of consistent and comprehensive documentation of specimens and data, which could support the identification of suspected, or proven use of, fabricated data or specimens of unclear origin. Hence, the urgent need for 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.[12-16]

It is generally accepted that the reliability of data generated in downstream analytical procedures[17-19] is significantly impacted by the properties and quality attributes of specimens, which are precursors of the data. Experts from multiple life sciences domains have called for the improvement and standardization of the documentation of research and scientific service processes.[20-26] 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.[27-32] 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. 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, that is, the data, the metadata, the biological materials, or the specimens.

1.1 | Objectives for a provenance standard

One of the main characteristics of present-day research in life sciences is that the research objects, such as datasets or specimens, are exchanged between organizations. Therefore, each of the organizations involved can only provide documentation for a part of the object's life cycle. Consequently, an uninterrupted chain of provenance information documenting the whole life cycle can only be formed from individual parts of provenance distributed across different sources. To enable meaningful integration and harmonized processing of the distributed provenance parts, semantic interoperability between standalone distributed provenance parts must be ensured. In addition, the processing of the resulting chain of distributed provenance must be designed to (a) deal with missing provenance components in the chain, so the chain is not interrupted or corrupted when an intermediary organization has not generated appropriate provenance information, or if the organization ceased to exist; (b) handle sensitive or confidential information contained in provenance information, keeping it opaque and disclosed only by authorization; (c) handle several versions of the same provenance information, for instance, when an error in provenance is found and is fixed; and (d) enable verification of the integrity and authenticity of provenance components, even for opaque provenance components, to ensure the trustworthiness of provenance.

The distributed provenance chain must be suited to answer essential queries independent of the research domain, such as *"What are the precursors of a given dataset?"* or *"Which processes precede a given dataset creation?"*. The underlying query resolution mechanism must be able to navigate through the chain, regardless of the actual site where the corresponding part of the distributed provenance is stored, which processes or objects are documented, or what the actual source of the provenance is.

The provenance standard must therefore include a general concept, providing a basis for common aspects shared between various domains which are part of the life cycle of a documented research object. In particular, these common aspects include (a) traversing distributed provenance chains; (b) implementing domain-independent properties for the provenance, such as confidentiality, authenticity, integrity, non-repudiation, and validity; and (c) locating a specific part of provenance in the distributed provenance. In addition, support for any domain-specific aspect, such as quality-related queries, must be provided and aligned with the common foundation without disrupting the general properties of the chain.

2 | RESULTS AND DISCUSSION

The novelty of the proposed standard is that it is the first provenance information standard for the biomedical domain that aims to address the aforementioned requirements. In addition, the standard covers both, physical and digital objects and links them to a common provenance chain, while ensuring the common properties of resulting provenance parts. It supports fully distributed provenance information management and aims to handle a wide range of complex real-world scenarios. As part of the standard development, we have proposed the Common Provenance Model (CPM),[33] which forms the conceptual

foundation of the standard. The CPM is the only provenance model that provides a baseline for distributed provenance chains, as they were described above.

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 approved a new work item proposal to develop an international standard for biological material and data provenance which is registered as a committee draft, ISO/DTS 23494-1 *Biotechnology—Provenance information model for biological material and data—Part 1: Design concepts and general requirements*. To the best of our knowledge, this standard is the first provenance information standard for the biotechnology domain, addressing the need for consistent documentation of the life-cycle of related research objects from the acquisition of a specimen to analytical procedures and downstream data processing and analysis. This standard is conceptualized according to the FAIR principles,[34] which provide high-level methodological recommendations, including guidance on provenance.* As the FAIR principles themselves do not provide detailed instructions for the implementation of provenance standards and documentation, the ISO 23494 series is intended for the provenance of data and biological samples and will be built on the World Wide Web Consortium's (W3C) PROV model,[35] a generic provenance information standard that defines a general model, corresponding serializations† 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,[36] for example, for computational workflows,[37] pharmacologic pipelines,[38] neuroscience,[39,40] microscopy experiments,[41] medical sciences,[42] and health implementation care‡ in HL7 FHIR.[43] 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 to develop a distributed, fully technically and semantically interoperable provenance information standard that covers uninterrupted documentation of the whole life cycle of a dataset back to its "source". The "source" can include a complex, multi-institutional environment and can be both the source specimen and data, but also a link to a specific biological entity, or environmental specimen collected at a given time and location (*connectivity* requirement[44]).

The main goals of the provenance information standard are as follows:

i. To support improved traceability and reproducibility of life-sciences research, to provide a voluntary provenance framework enabling concordance of governments, businesses, academia, and the international community.
ii. To enable decision-making about the fitness-for-purpose of particular data and specimens, 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.
iii. To achieve harmonization of documentation of specimens that is compliant with international conventions, recognized ethical practices, and legal requirements such as the Nagoya Protocol[45] and the Declaration of Taipei.[46]

The standard will enhance the trustworthiness of provenance information by including requirements and guidelines on its integrity, authenticity, and non-repudiation,[47] to prevent the production and/or use of unreliable, flawed, or fabricated data (the potential harms of which have become evident also during the COVID-19 pandemic),[2,14] as well as accidental or malicious modification of data. Since provenance information may also include sensitive or personal data (related, eg, 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 the production of reference material by maintaining full documentation of provenance (complementing work of ISO/TC 334[48]). For research concerned with highly regulated fields in life sciences, such as the development of medical products or drugs, the standardized provenance model will also contribute to a level of accountability and auditability of research organizations.

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, hospitals, biobanks and biorepositories, culture collections, research centers, and private companies (eg, pharmaceutical companies or lab reagent suppliers). The broader audience includes not only research data producers, but also those publishing, cataloging, archiving, or reusing research data.[49] The standard can also be adopted by manufacturers and vendors of laboratory instruments—for example, 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.

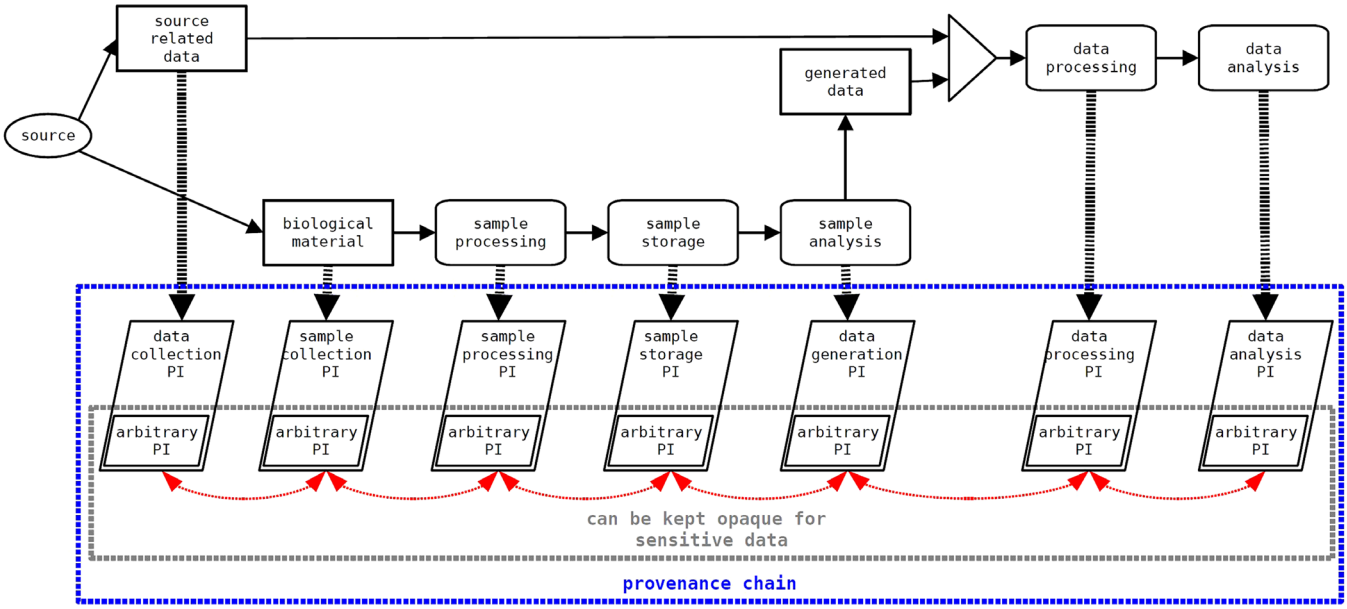

FIGURE 1 An overview of the provenance chain. A sample obtained from a donor (or other sources) is created and an initial set of provenance information (PI) is generated. As that sample moves through time and space, is processed and/or analyzed, additional provenance data are appended to the provenance 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. The figure cited from.[33]

The current draft proposal ISO/DTS 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. *Design concepts and general requirements* provide 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 the interlinking of distributed components of provenance information, the identification of physical and digital objects, provenance information patterns for common scenarios, such as missing provenance components in the chain, the compound processes, versioning of provenance information, or documentation of accountabilities. The model will also define mechanisms to embed or reference entire records of provenance information.
3. *Provenance of Biological Material* defines the 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 the legal and ethical basis (eg, consent, terms of access, and benefit sharing) of the collection. It will also provide mechanisms to reference Standard Operating Procedures and compliance with or deviations from them. Referencing the widely accepted de-facto reporting standard for biological specimen quality SPREC[50] 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, and 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, for example, sequencing, microscopy, spectrometry, and so on. Provenance information specific for diverse analytical or observational data generation methods will be embedded in a way meeting the requirements of a particular domain, but is well compliant with the provenance model standard allowing seamless integration in a complete provenance chain. This will be supported by the definition of standardized links from provenance to domain-specific information documenting the applied data generation method. As the syntax and semantics of the domain-specific information may be in the scope of another standard, the standardized links will provide information about the conformance of the domain-specific information to a particular standard.
5. *Provenance of Data Processing* defines the provenance of computational aspects of life sciences research such as the execution of computational workflows, for which we plan to leverage existing standards such as CWLProv[37] and RO-Crate,[51] which is being

complemented by a specialized profile to capture the provenance of workflow runs.§

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/DTS 23494-1 and ISO/PWI TS 23494-2 have a proven market need and have passed through the preliminary stages of the ISO voting process—as a result, they are part of the ISO Work Programme. ISO/DTS 23494-1 *Provenance information model for biological material and data—Part 1: Design concepts and general requirements* is published. Part 2 of this series, *Biotechnology—Provenance information model for biological material and data—Part 2: Common provenance model*, has been accepted by ISO/TC 276/WG 5 as preliminary work item ISO/PWI TS 23494-2. Part 3 of the series, *Biotechnology—Provenance information model for biological material and data—Part 3: Provenance of biological material*, will be proposed to become a Preliminary Work Item in 2023. The future documents in this series are in the planning stages but have not yet been submitted to ISO/TC 276/WG 5. The standards development process builds on existing standards for the 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 liaisons 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,¶ Open Microscopy Environment, OME,** genetic data compression (ISO/IEC JTC1/SC 29/WG 08 MPEG-G),[52] clinical trials and decision support systems and other life sciences domains such as biodiversity, marine biology, and systems biology.

2.1 | Industrial vs community-based standards

Alternatives to the ISO standards process†† exist—some community-based efforts have developed widely adopted specifications that have become de facto global standards.‡‡ 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.[56] 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. Industrial collaboration is established by grounding the standardization effort in the ISO, where industry experts drive all aspects of a standard development process through their involvement in the ISO Technical Committees. The presented ISO standard development is thus considered a standardized instance of a publicly available provenance model[33] developed in parallel under the auspices of the EOSC-Life project.[57]

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§§ 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.

2.2 | Open issues

The Common Provenance Model can be seen as a current state-of-the-art provenance model for distributed provenance, which is the most advanced provenance model that aims to provide a foundation for distributed provenance chains.[33] The development of the CPM was piloted using a distributed research pipeline covering biological material acquisition and storage, sample processing, data generation, and data processing. The prototype implementation of provenance generation was provided for the computational steps of the research pipeline.

However, the model should be rigorously validated in different domains, including multiple scientific communities and industries, to verify its applicability in diverse domains in life sciences. The model is currently being applied in the BY-COVID project,¶¶ which aims to develop a platform to integrate sources related to viral infections (clinical data, biological material, and research results). As part of this activity, the model will be integrated with RO-Crate[51] and applied to various use cases, including machine learning computational workflows and federated analysis.

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 BBMRI-ERIC.

AFFILIATIONS

[1]BBMRI-ERIC, Graz, Austria
[2]Institute of Computer Science & Faculty of Informatics, Masaryk University, Brno, Czechia
[3]CRS4—Center for Advanced Studies, Research and Development in Sardinia, Pula, Italy
[4]Medical University Graz, Graz, Austria
[5]National Institute of Standards and Technology, Gaithersburg, Maryland, USA
[6]Biological Resource Center of Institut Pasteur (CRBIP), Paris, France
[7]EMBL's European Bioinformatics Institute (EMBL-EBI), Cambridge, UK
[8]Plentzia Marine Station (PiE-UPV/EHU), University of the Basque Country, EMBRC-Spain, Bilbao, Spain
[9]University of Southampton, Southampton, UK
[10]King's College London, London, UK
[11]Ontario Institute for Cancer Research, Toronto, Ontario, Canada
[12]University of Klagenfurt, Klagenfurt, Austria
[13]Department of Social Statistics, School of Social Sciences, University of Manchester, Manchester, UK
[14]Flanders Marine Institute (VLIZ), EMBRC-Belgium, Ostend, Belgium
[15]Department of Computer Science, University of Manchester, Manchester, UK
[16]Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany
[17]Independent consultant
[18]ITTM S.A., Esch-sur-Alzette, Luxembourg
[19]Biosystems and Biomaterials Division, NIST, Gaithersburg, Maryland, USA
[20]Department of Biomedicine, University of Basel, Basel, Switzerland
[21]SCI-STI-MM, École Politechnique Fédérale de Lausanne, Lausanne, Switzerland
[22]Centre for Gene Regulation and Expression and Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK
[23]German BioImaging–Gesellschaft für Mikroskopie und Bildanalyse e.V., Konstanz, Germany
[24]Japan bio-Measurement and Analysis Consortium, Tokyo, Japan
[25]INSERM–Institut National de la Sante et de la Recherche Medicale, Paris, France
[26]Standards Council of Canada, Ottawa, Ontario, Canada
[27]Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Department of Family Medicine, Queen's University, Kingston, Ontario, Canada
[28]Biocomplexity Institute, Indiana University, Bloomington, Indiana, USA
[29]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[30]Program in Molecular Medicine, University of Massachusetts Chan Medical School, Worcester, Massachusetts, USA
[31]US Department of Agriculture, Washington, District of Columbia, USA
[32]Interdisciplinary Bank of Biomaterials and Data Würzburg (ibdw), Würzburg, Germany

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 PAM and the XDATA Projects, funded by the Sardinian Regional Authority. VC and MCh are 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 (RJ112/N027) and by NIHR Application Research Collaboration South London (ARC SL). TB, MCo acknowledges funding from EMBL-EBI Core Funds and the FAIRplus project (H2020 No 802750). MCo was 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), the US National Science Foundation (NSF 2054061), and the US EPA (RD840027). ME was supported by the Alan Turing Institute (ProvAnon). KZ was supported by the Bundesministerium für Bildung, Wissenschaft und Forschung (Federal Ministry of Education, Science and Research of Austria) (BMBWF-10.470/0010-V/3c/2018). CS was supported by NIH grant #U01CA200059 and by grant #2019-198155 (5022) awarded by the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation, as part of their Imaging Scientist Program. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders. Representation of communities: The co-author's team represents a wide coverage of life-sciences communities. PH, RW, CM, FF, HM, MP, and 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 is a director of Ontario Institute for Cancer Research. IC and KE come from marine biology and EMBRC Research Infrastructure. CG and SSR have worked with bioinformatics, CWL, and RO-Crate. JRS and JM come from bio-imaging communities and EUBioImaging Research Infrastructure. VC and MCh come from health informatics. HN participates in the provenance standardization process as an expert from Japan, MS and JS as experts from the United States, and AK as an expert from Luxembourg. ME contributes to privacy protection and provenance aspects. FB is a biobanking expert and director of the microbiological resource center CRBIP, Institut Pasteur. AS is a biobanking expert and ESBB councilor. 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 the capture and handling of provenance within large organizations. CS is a

Cell Biologists actively engaged in the development of quality control and reproducibility specifications and tools for light microscopy as a member of the Data Coordination and Integration Center of the NIH-funded 4D Nucleome initiative, Chair of the Quality Control and Data Management WG of BioImaging North America, and Co-Chair of the WG on Metadata (WG7) of the QUality Assessment and REProducibility for Instruments and Images in Light-Microscopy (QUAREP-LiMI) initiative. SLe is a member of the RO-Crate community and co-chair of a working group for the development of an RO-Crate profile for capturing the provenance of scientific workflow executions.

CONFLICT OF INTEREST STATEMENT

The authors report that they have no conflicts of interest.

ORCID

Rudolf Wittner https://orcid.org/0000-0002-0003-2024
Petr Holub https://orcid.org/0000-0002-5358-616X
Cecilia Mascia https://orcid.org/0000-0002-8952-725X
Francesca Frexia https://orcid.org/0000-0003-1007-1286
Heimo Müller https://orcid.org/0000-0002-9691-4872
Markus Plass https://orcid.org/0000-0003-2718-7648
Clare Allocca https://orcid.org/0000-0003-4132-0396
Fay Betsou https://orcid.org/0000-0002-0558-4653
Ibon Cancio https://orcid.org/0000-0003-4841-0079
Martin Chapman https://orcid.org/0000-0002-5242-9701
Mélanie Courtot https://orcid.org/0000-0002-9551-6370
Vasa Curcin https://orcid.org/0000-0002-8308-2886
Johann Eder https://orcid.org/0000-0001-6050-468X
Mark Elliot https://orcid.org/0000-0002-3142-4493
Katrina Exter https://orcid.org/0000-0002-5911-1536
Carole Goble https://orcid.org/0000-0003-1219-2137
Martin Golebiewski https://orcid.org/0000-0002-8683-7084
Andreas Kremer https://orcid.org/0000-0003-1466-0600
Simone Leo https://orcid.org/0000-0001-8271-5429
Sheng Lin-Gibson https://orcid.org/0000-0001-5092-1519
Anna Marsano https://orcid.org/0000-0002-3084-0823
Marco Mattavelli https://orcid.org/0000-0002-7742-0332
Josh Moore https://orcid.org/0000-0003-4028-811X
Hiroki Nakae https://orcid.org/0000-0002-5064-8468
Isabelle Perseil https://orcid.org/0000-0001-9058-9290
Ayat Salman https://orcid.org/0000-0002-7747-2757
James Sluka https://orcid.org/0000-0002-5901-1404
Stian Soiland-Reyes https://orcid.org/0000-0001-9842-9718
Caterina Strambio-De-Castillia https://orcid.org/0000-0002-1069-1816
Michael Sussman https://orcid.org/0000-0002-8432-0487
Jason R. Swedlow https://orcid.org/0000-0002-2198-1958
Kurt Zatloukal https://orcid.org/0000-0001-5299-7218
Jörg Geiger https://orcid.org/0000-0002-7689-531X

ENDNOTES

* Principle R1.2: (Meta)data are associated with detailed provenance.

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

‡ https://www.hl7.org/fhir/provenance.html.

§ https://www.researchobject.org/workflow-run-crate/.

¶ https://www.eosc-life.eu/.

** https://www.openmicroscopy.org/.

†† https://www.iso.org/developing-standards.html.

‡‡ E.g., for on-line cryptography (RSA public keys[53]), scientific workflows (Common Workflow Language[54]) and bioimaging data formats (OME-TIFF[55]).

§§ In some cases ISO standards can be obtained without any fee, for example, https://www.iso.org/covid19.

¶¶ https://by-covid.org/.

REFERENCES

1. Begley CG, Ioannidis JP. Reproducibility in science. *Circ Res*. 2015;116:116-126. doi:10.1161/CIRCRESAHA.114.303819
2. Servick K, Enserink M. The pandemic's first major research scandal erupts. *Science*. 2020;368:1041-1042. doi:10.1126/science.368.6495.1041
3. Lagoze C. Big data, data integrity, and the fracturing of the control zone. *Big Data Soc*. 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, Labbé C. 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, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*. 2011;10. Number: 9 Publisher: Nature Publishing Group:712–2. doi:10.1038/nrd3439-c1
8. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *NPJ Digit Med*. 2019;2:2. doi:10.1038/s41746-019-0079-z
9. Sheen MR, Fields JL, Northan B, et al. Replication study: biomechanical remod-eling of the microenvironment by stromal caveolin-1 favors tumor invasion and metastasis. *Elife*. 2019;8. Ed. by Sean J M and Joan M: e45120. doi:10.7554/eLife.45120
10. Errington TM, Denis A, Perfito N, et al. Reproducibility in Cancer Biology: Chal-lenges for assessing replicability in preclinical cancer biology. *Elife*. 2021;10. Ed. by Rodgers P and Franco E: e67995. doi:10.7554/eLife.67995
11. Tiwari K, Kananathan S, Roberts MG, et al. Reproducibility in systems biology modelling. *Mol Syst Biol*. 2021;17:e9982. doi:10.15252/msb.20209982
12. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol*. 2015;13:1-9. doi:10.1371/journal.pbio.1002165
13. Nickerson D, Atalag K, de Bono B, 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. doi:10.1098/rsfs.2015.0103
14. Mahase E. Covid-19: 146 researchers raise concerns over chloroquine study that halted WHO trial. *BMJ*. 2020;369:369. doi:10.1136/bmj.m2197
15. Chaplin S. Research misconduct: how bad is it and what can be done? *Future Prescriber*. 2012;13:5-76. doi:10.1002/fps.88
16. Committee on Responsible Science, Committee on Science, Engineering, Medicine, and Public Policy, Policy and Global Affairs, et al.

Fostering Integrity in Research. Washington, D.C.: National Academies Press; 2017:21896. doi:10.17226/21896
17. Simeon-Dubach D, Perren A. Better provenance for biobank samples. *Nature*. 2011;475:454-455. doi:10.1038/475454d
18. Holub P, Kohlmayer F, Prasser F, et al. Enhancing reuse of data and biological material in medical research: from FAIR to FAIR-health. *Biopreserv Biobank*. 2018;16:97-105. doi:10.1089/bio.2017.0110
19. 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
20. Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383:166-175. doi:10.1016/S0140-6736(13)62227-8
21. Freedman LP, Inglese J. The increasing urgency for standards in basic biologic research. *Cancer Res*. 2014;74:4024-4029. doi:10.1158/0008-5472.CAN-14-0925
22. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature*. 2012;483:531-533. doi:10.1038/483531a arXiv: 9907372v1.
23. 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]:187–91:187-191. doi:10.1038/nature11556
24. Consortium of European Taxonomic Facilities (CETAF). Code of Conduct and Best Practice for Access and Benefit-Sharing. 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 December 30, 2022)
25. Benson EE, Harding K, 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. *System Biodivers*. 2016;14:525-547. doi:10.1080/14772000.2016.1201167
26. A-E K, 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. *Eur Mar Biol Resource Centre*. 2020; https://bluebiobank.eu/docs/EMBRCGuideABS.pdf
27. Villanueva AG, Cook-Deegan R, Koenig BA, et al. Characterizing the biomedical data-sharing landscape. *J Law Med Ethics*. 2019;47:21-30. doi:10.1177/1073110519840481
28. Hulsen T. Sharing is caring-data sharing initiatives in healthcare. *Int J Environ Res Public Health*. 2020;17:E3046. doi:10.3390/ijerph17093046
29. Banzi R, Canham S, Kuchinke W, Krleza-Jeric K, Demotes-Mainard J, Ohmann C. Evaluation of repositories for sharing individual-participant data from clinical studies. *Trials*. 2019;20:169. doi:10.1186/s13063-019-3253-3
30. Toh S. Analytic and data sharing options in real-world multidatabase studies of comparative effectiveness and safety of medical products. *Clin Pharmacol Ther*. 2020;107:834-842. doi:10.1002/cpt.1754
31. Grossman RL. Data Lakes, clouds, and commons: A review of platforms for analyzing and sharing genomic data. *Trends Genet*. 2019;35:223-234. doi:10.1016/j.tig.2018.12.006
32. Wilson SL, Way GP, Bittremieux W, Armache JP, Haendel MA, Hoffman MM. Sharing biological data: why, when, and how. *FEBS Lett*. 2021;595:847-863. doi:10.1002/1873-3468.14067
33. Wittner R, Mascia C, Gallo M, et al. Lightweight distributed provenance model for complex real–world environments. *Sci Data*. 2022;9:503. doi:10.1038/s41597-022-01537-6
34. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. E FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. doi:10.1038/sdata.2016.18
35. Groth P, Moreau L. PROV-Overview: An Overview of the PROV Family of Documents. 2013 https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
36. Huynh TD, Groth P, Zednik S. PROV Implementation Report. 2013 http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/
37. Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR. Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. *GigaScience*. 2019;8:giz095. doi:10.1093/gigascience/giz095
38. Mammoliti A, Smirnov P, Safikhani Z, Ba-Alawi W, Haibe-Kains B. Creating reproducible pharmacogenomic analysis pipelines. *Sci Data*. 2019;6:166. doi:10.1038/s41597-019-0174-7
39. 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. 2015:226–231. doi:10.1109/CBMS.2015.10
40. Giesler A, Czekala M, Hagemeier B, Grunzke R. UniProv: A flexible provenance tracking system for UNICORE. In: Di Napoli E, Hermanns MA, Iliev H, et al., eds. *High-Performance Scientific Computing*. Cham: Springer International Publishing; 2017:233-242. doi:10.1007/978-3-319-53862-4_20
41. Samuel S. Integrative data management for reproducibility of microscopy experiments. In: Blomqvist E, Maynard D, Gangemi A, et al., eds. *The Semantic Web*. Cham: Springer International Publishing; 2017:246-255. doi:10.1007/978-3-319-58451-5_19
42. Curcin V, Fairweather E, Danger R, Corrigan D. Templates as a method for implementing data provenance in decision support systems. *J Biomed Inform*. 2017;65:1-21. doi:10.1016/j.jbi.2016.10.022
43. HL7 and its participants. FHIR Release #4B [Standard], version 4.3.0. 2022 http://hl7.org/fhir/R4B/
44. Curcin V, Miles S, Danger R, Chen Y, Bache R, Taweel A. Implementing interoperable provenance in biomedical research. *Future Gener Comput Syst*. 2014;34. Special Section: Distributed Solutions for Ubiquitous Computing and Ambient Intelligence: 1–16. doi:10.1016/j.future.2013.12.001
45. Secretariat of the Convention on Biological Diversity. Secretariat of the Convention on Biological Diversity. The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity. Convention on Biological Diversity, United Nations. 2021 https://www.cbd.int/abs/ (visited on December 30, 2022)
46. WMA—The World Medical Association—WMA Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks. https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/ (visited on December 30, 2022)
47. Fairweather E, Wittner R, Chapman M, et al. Non-repudiable provenance for clinical decision support systems. In: IPAW 2020, IPAW 2021: provenance and annotation of data and processes. In: Glavic B, Braganholo V, Koop D, eds. *Lecture Notes in Computer Science*. Vol 12839. Cham: Springer; 2021:162-182. doi:10.1007/978-3-030-80960-7_10 arXiv: 2006.11233 [cs.CR].
48. 14:00–17:00. ISO/WD Guide 85. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/55/75538.html (visited on December 30, 2022)
49. Cheney J, Chapman A, Davidson J, et al. Data provenance, curation and quality in metrology. In: advanced mathematical and computational tools in metrology and testing XII. Vol. 90. Series on advances in mathematics for applied sciences. *World Sci*. 2021;90:167-187. doi:10.1142/9789811242380_0009 arXiv: arXiv:2102.08228v1.
50. Betsou F, Bilbao R, Case J, et al. Standard PREanalytical code version 3.0. *Biopreserv Biobank*. 2018;16:9-12. doi:10.1089/bio.2017.0109

51. Soiland-Reyes S, Sefton P, Crosas M, et al. Packaging research artefacts with RO-Crate. *Data Sci*. 2022;5:97-138. doi:10.3233/ds-210053
52. Voges J, Hernaez M, Mattavelli M, Ostermann J. An introduction to MPEG-G: the first open ISO/IEC standard for the compression and exchange of genomic sequencing data. *Proc IEEE*. 2021;109:1607-1622. doi:10.1109/JPROC.2021.3082027
53. Rivest RL, Shamir A, Adleman L. A method for obtaining digital signatures and public-key cryptosystems. *Commun ACM*. 1978;21:120-126. doi:10.1145/359340.359342
54. Crusoe MR, Abeln S, Iosup A, et al. Methods included: standardizing computational reuse and portability with the common workflow language. *Commun ACM*. 2022;65. doi:10.1145/3486897
55. Linkert M, Rueden CT, Allan C, et al. Metadata matters: access to image data in the real world. *J Cell Biol*. 2010;189:777-782. doi:10.1083/jcb.201004104
56. Swedlow JR, Kankaanpää P, Sarkans U, et al. A global view of standards for open image data formats and repositories. *Nat Methods*. 2021;18:1440-1446. doi:10.1038/s41592-021-01113-7
57. Wittner R, Mascia C, Frexia F, et al. EOSC-Life Common Provenance Model. EOSC-Life deliverable D6.2. 2021. doi:10.5281/zenodo.4705074

How to cite this article: Wittner R, Holub P, Mascia C, et al. Toward a common standard for data and specimen provenance in life sciences. *Learn Health Sys*. 2023;e10365. doi:10.1002/lrh2.10365