ISO 23494: Biotechnology – Provenance Information Model for Biological Specimen and Data*

Rudolf Wittner[1,2][0000−0002−0003−2024], Petr Holub[1,2][0000−0002−5358−616X], Heimo Müller[3][0000−0002−9691−4872], Joerg Geiger[4][0000−0002−7689−531X], Carole Goble[5][0000−0003−1219−2137], Stian Soiland-Reyes[5,11][0000−0001−9842−9718], Luca Pireddu[6][0000−0002−4663−5613], Francesca Frexia[6][0000−0003−1007−1286], Cecilia Mascia[6][0000−0002−8952−725X], Elliot Fairweather[7][0000−0003−0880−0785], Jason R. Swedlow[8][0000−0002−2198−1958], Josh Moore[8][0000−0003−4028−811X], Caterina Strambio[9][0000−0002−1069−1816], David Grunwald[9][0000−0001−9067−804X], and Hiroki Nakae[10][0000−0002−5064−8468]

[1] BBMRI-ERIC, AUT
rudolf.wittner@bbmri-eric.eu
[2] Institute of Computer Science & Faculty of Informatics, Masaryk University, CZ
[3] Medical University Graz, AUT
[4] Interdisciplinary Bank of Biomaterials and Data Würzburg (ibdw), Würzburg, DE
[5] Department of Computer Science, The University of Manchester, UK
[6] CRS4 – Center for Advanced Studies, Research and Development in Sardinia, IT
[7] King's College London, UK
[8] School of Life Sciences, University of Dundee, Dundee, UK
[9] University of Massachusetts, US
[10] Japan bio- Measurement and Analysis Consortium, JPN
[11] Informatics Institute, University of Amsterdam, NL

Abstract. Exchange of research data and samples in biomedical research has become a common phenomenon, demanding for their effective quality assessment. At the same time, several reports address reproducibility of research, where history of biological samples (acquisition, processing, transportation, storage, and retrieval) and data history (data generation and processing) define their fitness for purpose, and hence their quality. This project aims to develop a comprehensive W3C PROV based provenance information standard intended for the biomedical research domain. The standard is being developed by the working group 5 ("data processing and integration") of the ISO (International Standardisation Organisation) technical committee 276 "biotechnology". The outcome of the project will be published in parts as international standards or technical specifications. The poster informs about the goals of the standardisation activity, presents the proposed structure of the standards, briefly describes its current state and outlines its future development and open issues.

Keywords: provenance · biotechnology · standardization

* Supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 654248, project CORBEL; grant agreement No. 824087, project EOSC-Life; and grant agreement No. 823830, project BioExcel-2.

1 Introduction

Research in life sciences has undergone significant changes during recent years, evolving away from individual projects confined to small research groups to transnational consortia covering a wide range of techniques and expertise. At the same time, several reports addressing the quality of research papers in life sciences have uncovered an alarming number of ill-founded claims. The reasons for the deficiencies are diverse, with insufficient quality and documentation of the biological material used being the major issue [1–3]. Hence there is urgent need for standardized and comprehensive documentation of the whole workflow from the collection, generation, processing and analysis of the biological material to data analysis and integration.

The PROV[4] family of documents serves as a current standard for provenance information used to describe the history of an object. On the other hand, as discussed in the results from EHR4CR and TRANSFoRm projects [5, 6], its implementation for the biotechnology domain and the field of biomedical research in particular is still a pending issue. To address this, the International Standardisation Organisation (ISO) initiated the development of a *Provenance Information Model for Biological Specimen and Data* standard defining the requirements for interoperable, machine-actionable documentation intended to describe the complete process chain from the source of biological material through its processing, analysis, and all steps of data generation and data processing to final data analysis.

The standard is intended for implementers and suppliers of HW/SW tools used in biomedical research (e.g. lab automation devices or analytical devices used for research purposes) and also for organisations adopting generated provenance (e.g. to require or use standardised tools).

2 Goals of the Standard and Its Structure

The main goals of the standard are to (a) enable effective assessment of quality and fitness for purpose of the objects provided, such as biological material and data; (b) support reproducible research by exacting the capture of all relevant information; (c) track error propagation within scientific results; (d) track the source of biological material in order to prevent fabrication of data and enabling notification of subjects in case of relevant incidental findings; (e) propagate withdrawal of or changes to an informed consent along the process chain.

The proposed structure of the standard reflects the intention to interconnect and integrate distributed provenance information furnished by all kinds of organisations involved in biotechnology research. Examples of such organisations are hospitals, biobanks, research centers, universities, data centers or pharmaceutical companies, where each of them is participating in research, thus generating provenance information describing particular activities or contributions.

In its current the standard is composed of the following 6 parts:

- **Part 1** stipulates common requirements for provenance information management in biotechnology to effectuate compatibility of provenance management at all stages of research and defines the design concept of this standard;
- **Part 2** defines a common provenance model which will serve as an overarching principle interconnecting provenance parts generated by all kinds of contributing organisations and enable access to provenance information in a distributed environment;
- **Parts 3, 4 and 5** are meant to complement the *horizontal* standards (1) and (2) as *vertical* standards defining domain specific provenance models describing diverse stages or areas of research in biotechnology (e.g. sample acquisition and handling, analytical techniques, data management, cleansing and processing; database validation);
- **Part 6** will contain optional data security extensions especially to address non-repudiation of provenance.

The proposed structure is also depicted in figure (1). Parts indicated by red boxes are considered as *horizontal* standards, i.e. providing a common basis for provenance information at all stages of research. The blue boxes indicate domain specific *vertical* standards build on top of the *horizontal* standards.

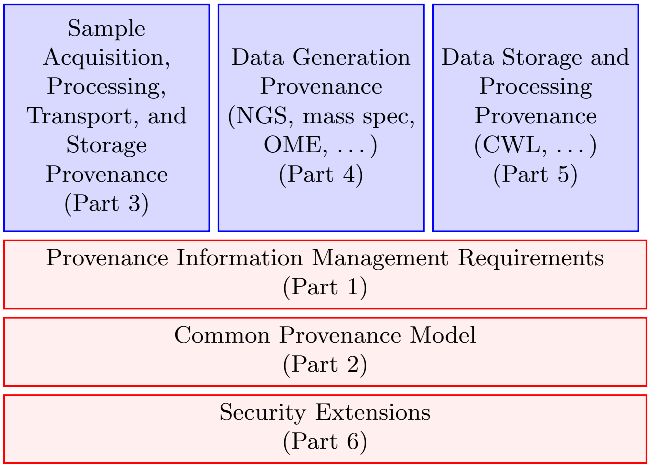

Sample Acquisition, Processing, Transport, and Storage Provenance (Part 3)
Data Generation Provenance (NGS, mass spec, OME, . . .) (Part 4)
Data Storage and Processing Provenance (CWL, . . .) (Part 5)
Provenance Information Management Requirements (Part 1)
Common Provenance Model (Part 2)
Security Extensions (Part 6)

Fig. 1. Overall structure of the standard

3 Current Status and Future Development

The standard is currently at an early stage of development. The PROV model has been already used to define new types of provenance structures, called *connectors*, that are used to interconnect provenance generated by different organizations. The concept of the connectors and a common mechanism for bundles versioning has been published as an EOSC-Life project provenance deliverable [7]. A publication describing use of the connectors at a specific use case is under development at the moment and its pre-print will be published in summer 2021. Continuously, the model will be enriched by new types of structures (e.g. relations, entities, etc.) to capture common objects. These structures will be subsequently used to design provenance templates[1] to define a common representation of usual scenarios in life sciences. Further aspects will be also targeted. The major focus areas are: opaque provenance components; privacy preservation and non-repudiation of provenance information; full syntactic and semantic interoperability of provenance information captured; rigorous formal verification process of provenance instance validity (provable compliance with the proposed model).

Another publication describing the standardization process in a more detailed way is under development. The publication will contain more detailed explanation of our motivation and the standardization activity itself, more detailed description of the standard structure, and finally, an important discussion on openness of the standard and related issues.

References

1. Freedman LP, Cockburn IM, and Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol 2015;13:e1002165. DOI: `10.1371/journal.pbio.1002165`.
2. Begley GC and Ioannidis JP. Reproducibility in Science. Circulation Research 2015;116:116–26. DOI: `10.1161/CIRCRESAHA.114.303819`.
3. 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`.
4. Groth P and Moreau L. PROV-Overview. An Overview of the PROV Family of Documents. W3C Working Group Note 30 April 2013. W3C, 2013. URL: `https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/`.
5. Curcin V, Miles S, Danger R, et al. Implementing interoperable provenance in biomedical research. Future Generation Computer Systems 2014;34:1–16. DOI: `10.1016/j.future.2013.12.001`.

[1] The templates can be considered as synonyms for named graphs or graph patterns. These concepts are used to abstract from actual instances of provenance and to describe repeating occurrences of components of provenance

6. Cuccuru G, Leo S, Lianas L, et al. An automated infrastructure to support high-throughput bioinformatics. In: *High Performance Computing & Simulation (HPCS), 2014 International Conference on*. IEEE. 2014:600–7. DOI: `10.1109/HPCSim.2014.6903742`.
7. Wittner R, Mascia C, Frexia F, et al. EOSC-Life Common Provenance Model. EOSC-Life deliverable D6.2. 2021. DOI: `10.5281/zenodo.4705074`.

ISO 23494

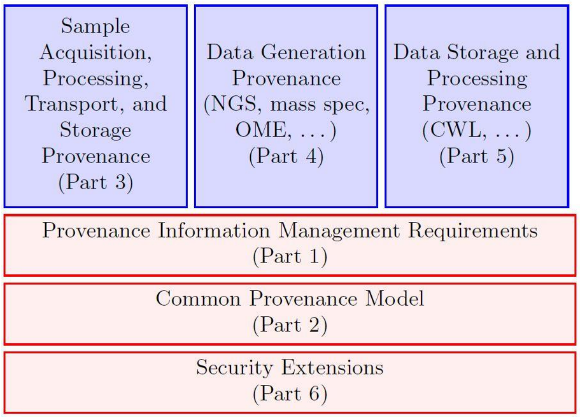
BIOTECHNOLOGY

PROVENANCE INFORMATION MODEL FOR BIOLOGICAL SPECIMEN AND DATA

The purpose of the standard is the standardization of provenance information for the biotechnology domain covering the whole process chain, from the source of biological material, through its processing, analysis, and all steps of data generation and processing.

GOALS OF THE STANDARDIZATION:

1. Enabling **effective assessment** of **quality** and **fitness for purpose** of the objects provided, such as biological material and data;
2. Supporting **reproducible research** by exacting the capture of all relevant information;
3. Tracking **error propagation** within scientific results;
4. Tracking the **source of biological material** in order to prevent fabrication of data and enabling the notification of subjects in case of relevant incidental findings;
5. Propagating **withdrawal** of or **changes** to an informed consent along the process chain;


Sample Acquisition, Processing, Transport, and Storage Provenance (Part 3)

Data Generation Provenance (NGS, mass spec, OME, . . .) (Part 4)

Data Storage and Processing Provenance (CWL, . . .) (Part 5)

Provenance Information Management Requirements (Part 1)

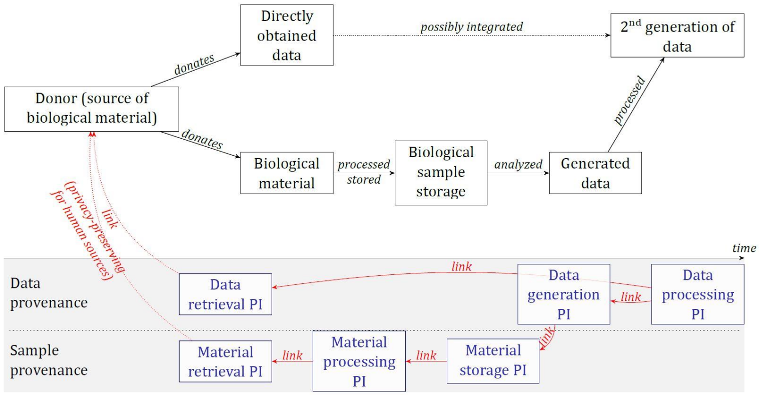
Common Provenance Model (Part 2)

Security Extensions (Part 6)

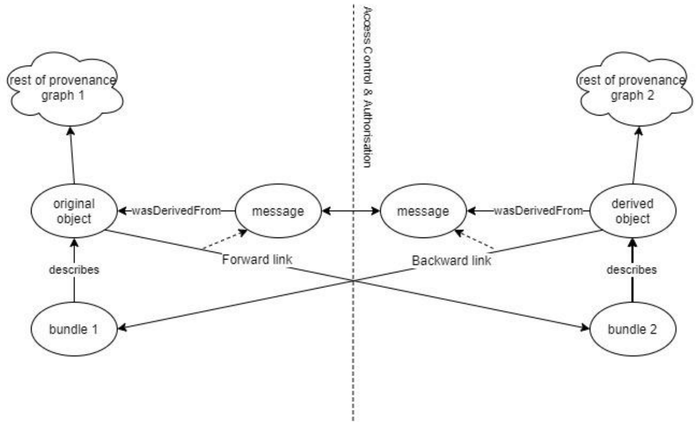
Proposed structure of the standard.

WG LEADERS AND MAIN CONTRIBUTORS

Petr Holub, Jörg Geiger, Rudolf Wittner, Carole Goble, Heimo Müller, Stian Soiland-Reyes, Elliot Fairweather, Luca Pireddu, Francesca Frexia, Cecilia Mascia, Gianluigi Zanetti, Hiroki Nakae, Caterina Strambio, Josh Moore, David Grunwald, Jason Swedlow



Standard coverage.



General schema of a distributed provenance model

FOCUS AREAS:

1. Applying **W3C PROV to describe** all phases of biomedical research and its enrichment by **new types of structures** (e.g. relations, entities, …) to capture common objects.
2. Definition of provenance templates as **common representation** of typical scenarios.
3. Interconnecting **distributed provenance** to enable processing of provenance stored within multiple organisations with support for **opaque provenance components** .
4. Full **syntactic and semantic interoperabili**ty of captured provenance.
5. Rigorous **formal verification process** of provenance instance validity (provable compliance with the model).
6. Access control, integrity and non-repudiation, protection of privacy.