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

R. Wittner & P. Holub & H. Müller
BBMRI-ERIC, AUT

J. Geiger
IBDW, DE

C. Goble & S. Soiland-Reyes
University of Manchester, UK

L. Pirredu & F. Frexia & C. Mascia & G. Zanetti
CRS4, IT

E. Fairweather
KCL, UK

H. Nakae
JMAC, JPN

C. Strambio & D. Grunwald
University of Massachusetts, US

J. Swedlow & J. Moore
University of Dundee, UK

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) defines their fitness for purpose, and hence their quality. The project aims at developing 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.

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 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 [1,4,5]. 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 [6] 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 [2,3], 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 an organisations are hospitals, biobanks, research centers, universities, data centers or pharma companies, where each of them is participating in research, thus generating provenance information describing particular activities or contributions.

In current planning the standard is assembled of 6 parts as follows:

- **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.

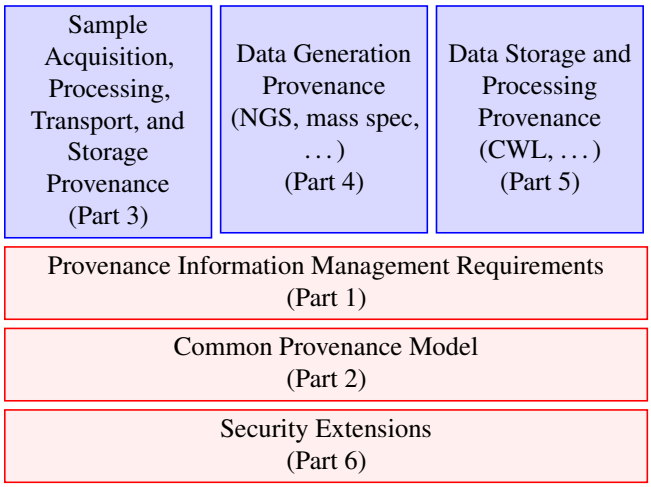


Figure 1: Overall structure of the standard

3 Current Status and Future Development

The standard is currently at a preliminary stage of development. The PROV model is being analysed in order to verify its usability in the context of biotechnology and identify necessary adaptions or amendments. Additionally, the model will be enriched by new types of structures (e.g. relations, entities, ...) to capture common objects. These structures will be subsequently used to design provenance templates[1] to define a common representation of usual scenarios. Further aspects will be targeted in future. The major focus areas are: full syntactic and semantic interoperability of provenance information captured; rigorous formal verification process of provenance instance validity (provable compliance with the proposed model); privacy preservation and non-repudiation of provenance information.

Acknowledgments

This work has been supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 654248, project CORBEL, and under grant agreement No 824087, project EOSC-Life.

References

[1] Glenn C. Begley and John P.A. Ioannidis. Reproducibility in science. *Circulation Research*, 116(1):116–126, 2015. https://www.ahajournals.org/doi/pdf/10.1161/CIRCRESAHA.114.303819.

[2] Gianmauro Cuccuru, Simone Leo, Luca Lianas, Michele Muggiri, Andrea Pinna, Luca Pireddu, Paolo Uva, Alessio Angius, Giorgio Fotia, and Gianluigi Zanetti. An automated infrastructure to support high-throughput bioinformatics. In *High Performance Computing & Simulation (HPCS), 2014 International Conference on*, pages 600–607. IEEE, 2014.

[3] Vasa Curcin, Simon Miles, R Danger, Y Chen, Richard Bache, and Adel Taweel. Implementing interoperable provenance in biomedical research. *Future Generation Computer Systems*, 34:1–16, 2014.

[4] Leonard P Freedman, Iain M Cockburn, and Timothy S Simcoe. The economics of reproducibility in preclinical research. *PLoS Biol*, 13(6):e1002165, 2015. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165.

[5] Leonard P. Freedman and James Inglese. The increasing urgency for standards in basic biologic research. *Cancer Research*, 74(15):4024–4029, 2014. https://cancerres.aacrjournals.org/content/74/15/4024.

[6] Paul Groth and Luc Moreau. Prov-overview. an overview of the prov family of documents, 2013. https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

[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

ISO 23494

BIOTECHNOLOGY

PROVENANCE INFORMATION MODEL FOR BIOLOGICAL SPECIMEN AND DATA

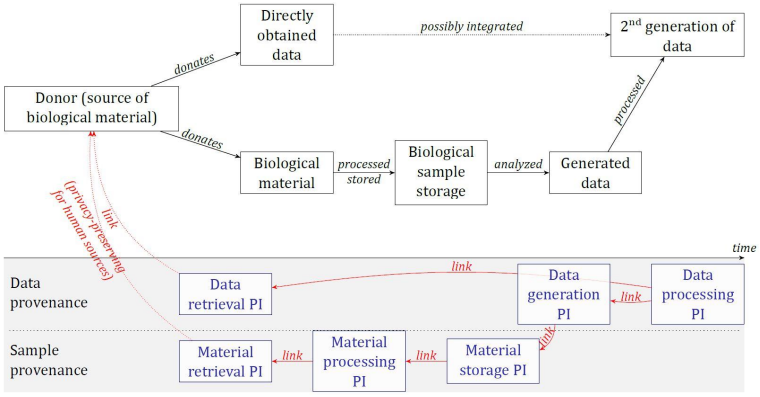
The purpose of the standard is the standardization of provenance information for the biotechnology domain covering the whole process chain, from the source of biological material, through its processing, analysis, and all steps of data generation and processing.

GOALS OF THE STANDARDIZATION:

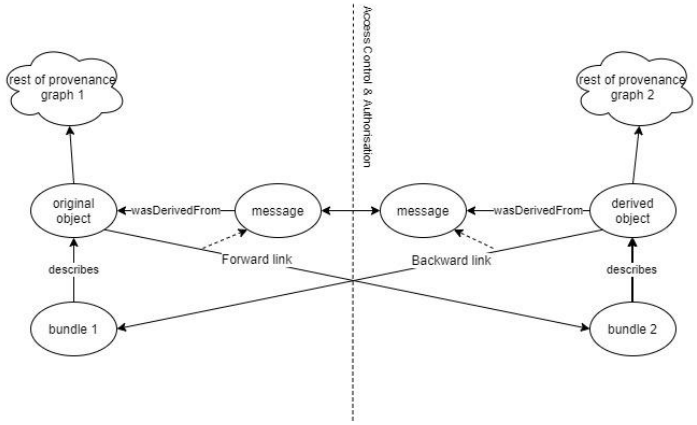
1. Enabling **effective assessment** of **quality** and **fitness for purpose** of the objects provided, such as biological material and data;
2. Supporting **reproducible research** by exacting the capture of all relevant information;
3. Tracking **error propagation** within scientific results;
4. Tracking 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 (IS) (Part 3)	Data Generation Provenance (NGS, mass spec, ...) (IS) (Part 4)	Data Storage and Processing Provenance (CWL, ...) (IS) (Part 5)
Provenance Information Management Requirements (TS) (Part 1)		
Common Provenance Model (TS) (Part 2)		
Security Extensions (TS) (Part 6)		

Proposed structure of the standard.



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 interoperabilit**y 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.

WG LEADERS AND MAIN CONTRIBUTORS

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

rudolf.wittner@bbmri-eric-eu | www.bbmri-eric.eu