TECHNICAL SPECIFICATION

ISO/TS 23494-1

First edition
2023-04

Biotechnology — Provenance information model for biological material and data —

Part 1: Design concepts and general requirements

Reference number
ISO/TS 23494-1:2023(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 276, *Biotechnology*.

A list of all parts in the ISO 23494 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

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. The exchange of research data and biological materials has become essential for the research in life sciences and biotechnology, and consequently interoperability and quality measures of data have become imperative.

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.

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 statistics. This provenance information serves as a quality indicator and provides information on the reliability thus enabling transparency and comparability of research results.

The purpose of these documents is the standardization of provenance information management for the biotechnology domain in a way that allows for meaningful data integration. To this end, provenance information needs to be prepared in a way that enables interoperability between prevailing tools for data generation, processing and analysis. While in information technology well-established approaches to provenance information management in general are available (e.g. OPM[1] or W3C PROV[2]), the implementation for the biotechnology domain and related fields in particular is still a pending issue (as discussed in the results from the Electronic Health Records Systems for Clinical Research (EHR4CR) and TRANSFoRm projects in several papers[3][4]).

Since data in biotechnology mostly originate from analysis of biological material, it is essential that the provenance information covers the entire process chain, from the source of biological material, throughout processing and analysis of the material, generation, and processing of the data to final analysis and interpretation. With the increasing adoption of data-intensive technologies, such as next-generation sequencing (NGS), high-throughput mass spectrometry as used for proteomics or metabolomics, or high-throughput microscopy in digital pathology, and their impact on data collection strategies, consistent and comprehensive documentation of data provenance has become a necessity.

In fact, experimental designs in life sciences have moved from individual experiments with a limited amount of data towards pipelines generating a vast volume of raw digital data using massively parallel acquisition systems demanding complex data processing workflows to extract biologically relevant information. This trend is particularly evident in NGS, where the actual data acquisition device at the wet lab-digital interface (i.e. the sequencer) is completely oblivious to the details of the experiment being performed, with all the specialization pushed to the protocols used by the sample preparation procedure and to the software pipelines processing the data. Software pipelines are continuously changing due to the evolution of analytical algorithms and reference data sets, which is having a significant effect on result concordance.

In addition, particular issues, relevant to scientific domains utilizing biological material and data obtained from humans, must be considered. These include aspects of data privacy, ethics or management of identities. Notably, issues such as withdrawal of an informed consent or communication of incidental findings require the implementation of appropriate mechanisms.

The major objectives for collecting and storing provenance information are summarized as follows:

— retrospective evaluation of experimental results and data analysis with respect to the influence of standard operating procedures (SOPs) and workflow parameters;

— quality monitoring of biological materials and data entered in a workflow or analysis pipeline (e.g. against reference ranges and tolerances);

— automation of quality control procedures (e.g. comparisons between different pipelines);

— profiling of sample and data analysis to identify bottlenecks;

— assessment of fitness for purpose of biological materials and data for the intended use.

To achieve these objectives, a digitally processable description of provenance information is required.

This overarching document will be complemented by appropriate vertical standards for specific fields (e.g. collection of biological material, data generation, processing of biological material and data). The basic requirements contained in this document do not impose any limitations to future, domain-specific standards based on this document.

The standardization of provenance information requires the conceptualization and essential specifications for the generation, management, provisioning and maintenance as described in this document. Not covered in this document are additional fundamental components such as a generic model for provenance information and extensions common to all kinds of provenance information, ensuring security, privacy and non-repudiation. For particular domains in biotechnology, detailed specifications building on a common provenance model are required, covering provenance information describing:

— the life cycle of biological materials, including acquisition, processing, transport and storage.

— the data generation by analytical methods.

— the data processing and analysis in computational workflows.

This document provides definitions for relevant terms used and specifies fundamental requirements for provenance information generation, management and provisioning.

Biotechnology — Provenance information model for biological material and data —

Part 1: Design concepts and general requirements

1 Scope

This document specifies a general concept for a provenance information model for biological material and data and requirements for provenance data interoperability and serialization.

The provenance information model covers any information relevant to the quality and fitness for purpose of the biological material generated throughout the preanalytical phase of the materials life cycle from collection to analysis, data originating from analytical procedures applied to the biological material and results from further mathematical processing of the data.

This document is applicable to organizations, authorities and industries that are:

a) collecting, processing or distributing biological material for research;

b) generating, collecting, analysing or storing data on biological material.

This document does not apply to biological material and data used for other than research or in fields that are regulated by national, regional or international laws, such as medical diagnosis and therapy or food production.

NOTE International, national, or regional regulations or requirements can also apply to specific topics covered in this document.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

United Nations Treaty Collection. 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 [online]. Available from: https://treaties.un.org/pages/ViewDetails.aspx?src=IND&mtdsg_no=XXVII-8-b&chapter=27&clang=_en

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
common provenance model
CPM
extension of *PROV-DM* (3.12) for generating, maintaining and provisioning *provenance information* (3.13) on biological material and data

3.2
described activity
activity performed on a *described object* (3.3)

Note 1 to entry: Examples for activities performed on physical objects can be biobanking activities as specified in ISO 20387:2018, 3.6. Examples for activities performed on digital objects can be data analytics as specified in ISO/IEC 20546:2019, 3.1.6.

3.3
described object
physical or digital object which is being documented by *provenance information* (3.13)

3.4
finalization event
time instance, at which generated and assembled *provenance information* (3.13) is transformed into *finalized provenance information* (3.5)

3.5
finalized provenance information
provenance information (3.13) transformed into a representation specified by the *common provenance model* (3.1), and which is prepared to be conserved or archived and which is considered as being immutable

Note 1 to entry: Finalized provenance information is a subset of provenance information.

3.6
interoperability
capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units

[SOURCE: ISO/IEC 20944-1:2013, 3.6.1.24]

3.7
machine-readable
pertaining to data in a form that can be automatically input to a computer

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.2300, modified — Example deleted.]

3.8
opaque provenance component
part of the *finalized provenance information* (3.5) that is findable but not directly accessible and is stored as an opaque object at the respective *responsible subject* (3.15)

3.9
persistent identifier
unique identifier that ensures permanent access for a digital object by providing access to it independently of its physical location or current ownership

[SOURCE: ISO 24619:2011, 3.2.4, modified — Abbreviated term and note to entry deleted.]

3.10
preparation
activities, taking place in a laboratory after acquisition, to make biological material ready for further use in the life cycle, storage or distribution

Note 1 to entry: These activities can include, e.g. centrifuging, homogenizing, purifying, fixing, stabilizing, replicating, filtering, sorting, culturing, vacuum drying, freeze drying, freezing and thawing, tissue sectioning, fractionating, dispensing/aliquoting, cryopreserving.

[SOURCE: ISO 20387:2018, 3.37]

3.12
PROV-DM
conceptual data model forming a basis for the W3C provenance (PROV) family of specifications

[SOURCE: PROV-DM:2013[2]]

3.13
provenance information
information that documents the history of a *described object* (3.3) and related *described activities* (3.2), and that contains information about the origin or source of the described object, any changes that can have taken place since it was originated, and who has had custody of it since it was originated

3.14
provenance provider
institution or its organizational unit that is responsible for the transformation of *provenance information* (3.13) into *finalized provenance information* (3.5) and its storage, and which can be either a *responsible subject* (3.15) or acting on behalf of a responsible subject

3.15
responsible subject
institution, or the organizational unit it is part of, which is responsible for providing and assembling *provenance information* (3.13) documenting a *described activity* (3.2) on a *described object* (3.3) it is involved in

3.16
serialization
process of translating data structures or object states into a format that can be stored or transmitted and reconstructed later

4 Requirements and recommendations on provenance information management

4.1 Provenance information generation

4.1.1 Output of machine-readable finalized provenance information shall occur only upon completion of the respective activity or process.

NOTE 1 Creating finalized provenance information only upon completion of a process is reasonable to avoid inconsistencies resulting from constant changes during recording of the data. Otherwise, requirements in terms of non-repudiation, trust and verifiability can be affected. In general, finalized provenance information is not meant to record the process conditions and parameters continuously, thus substituting for logging functionality, nor to detail any minor variation in the process conditions (e.g. temperature fluctuations during storage).

Finalized provenance information should represent an aggregate of the data available, indicating adherence or deviation from predefined boundary conditions as necessary to assess fitness for purpose (e.g. exceeding the temperature range for safe storage for a certain period, frequency and duration of deviations). If, however, detailed information is regarded vital or necessary to comply with particular regulations, embedding of comprehensive data is also possible.

If required, additional, specific finalized provenance information can be generated on request. Any additional finalized provenance information shall only be provided by the responsible subject.

NOTE 2 Such a request can be addressed by providing more detailed provenance information from recorded information. If only a subset of the finalized provenance information was made generally available (e.g. for privacy reasons or because of confidentiality), additional finalized provenance information can be generated for a particular user.

To ascertain consistency and validity any additional finalized provenance information should be compared to the original provenance information for plausibility.

4.1.2 Machine-readable provenance information generated by the devices involved in processing, handling, transporting or storing biological material, or involved in data generation, shall be used, if available.

The compilation and processing of information relevant for the generation of finalized provenance information shall be documented and should be automated as much as possible, to minimize errors and optimize the amount and quality of information. This includes also any transformation applied to the information gathered. For any algorithmic transformation, the underlying algorithms should be either archived or referenced, so that any modification and processing of the information used to generate finalized provenance information is reproducible and documented.

Devices qualified for providing finalized provenance information as specified in this document shall be set to provide the provenance information accordingly and the information shall be used unmodified.

If a device is able to provide provenance information in a documented, machine-readable format which does not conform to this document, the device shall be set to produce machine-readable provenance information. The provenance information obtained shall be transformed subsequently to conform to the standard format. The implementation of the algorithm used for transformation shall be documented and referenced in the finalized provenance information.

If machine-readable provenance information cannot be obtained, the provenance information shall be created by the responsible subject from the device output conforming to the document. The provenance information generated shall be then transformed into finalized provenance information during a finalization event.

4.1.3 Finalized provenance information on biological material and data shall invariably be linked to the finalized provenance information on the biological material or data from which it originates.

NOTE Creation of backward links to the origin of the biological material and data is essential for an assessment of the fitness for purpose or responding to reproducibility concerns, see also 4.1.5.

4.1.4 Finalized provenance information on biological material and data should invariably be linked to the finalized provenance information on the biological material or data derived therefrom.

NOTE Forward links pointing to the derivatives of biological material, analytical results, data processing or data analysis become relevant when any questions or problems with the biological material or data become apparent which can affect the validity of results obtained in subsequent processes. By means of forward references, users of the biological material or data can be notified accordingly.

4.1.5 Relevant ethical or legal issues such as withdrawal of consent or changes in usage rights should be forwarded to any subsequent user of the biological material or data.

4.1.6 Finalized provenance information should provide an unbroken chain to the source of biological material or to initially collected data, unless it is intentionally broken for legal or ethical reasons, such as the protection of individual rights.

NOTE 1 A valid assessment of the quality of biological material or data and their fitness for purpose can be secured only if any step in the sequence of collecting, processing and analysing biological material or data is thoroughly documented, findable and accessible. For biological material subject to the Nagoya Protocol, an uninterrupted chain of finalized provenance is fundamental to meet the requirements on access to genetic resources and the fair and equitable sharing of benefits arising from their use.

Finalized provenance information on biological material subject to the Nagoya Protocol shall record the geographical coordinates of where the material has been retrieved.

Biological material and data are commonly pseudonymized or anonymized, or both, to protect individual rights of the donor or uninvolved third parties (e.g. uninvolved family members) or in order to facilitate the access to and use of biological material and data.[5][6] The link to the original source shall be maintained after pseudonymization, though obscured, thus enabling the verification of the source without revealing the identity.

NOTE 2 Anonymized biological material and data are completely detached from the source, so that no provenance information on the source is available, and hence, cannot be verified.

4.1.7 Finalized provenance information is the sole basis for quality assessment, verification of provenance, and traceability of biological material and data. Once generated, finalized provenance information shall be held immutable. Alterations and manipulations should be prohibited by appropriate measures.

If changes to finalized provenance information, such as fixes or corrections, are necessary, a new version of the finalized provenance information shall be created for this purpose, whereby the original version is retained and only referenced by the newer version.

4.1.8 Plausibility checks of finalized provenance information shall be specified and implemented, based on the characteristics and constraints of the particular process.

NOTE For any process, particular characteristics (e.g. parameters, conditions) or constraints apply. Documentation of these characteristics and compliance with the constraints forms the basis for quality assessment and demonstrates the fitness for purpose of the biological material or data. These characteristics can also be used for a plausibility check of the finalized provenance information. By implementing obligatory plausibility checks, based on appropriate value ranges, the finalized provenance information can be tested automatically for consistency and validity. However, as a matter of principle, the properties monitored are highly process and device specific and hence cannot be covered by a general specification.

The range of characteristics (e.g. values) to be tested shall be defined by the responsible subject.

Sanity checks should be implemented as algorithms and stored for future inspection. Sanity checks shall be performed automatically and on a regular basis. The test results of the plausibility check shall be documented and archived.

4.2 Provenance information availability and findability

4.2.1 Finalized provenance information shall be made available according to the intended or foreseeable use of the biological material or data. Terms for accessibility can range from strictly confined to an organization to publicly disclosed. To protect a legitimate interest of confidentiality, publicly accessible finalized provenance information can be made available only in parts or entirely opaque. The provenance provider shall provide or commission a service provider to enable access to finalized provenance information as appropriate.

4.2.2 Finalized provenance information shall be entered in a registry using a resolvable, stable reference. The registry can be maintained by the provenance provider, or a service provider commissioned to maintain the registry.

4.2.3 Any responsible subject shall be identified using a globally resolvable persistent identifier.