ISO/PWI 23494-3: ####(E)

ISO/TC 276/WG 5

Secretariat: DIN

Biotechnology – Provenance information model for biological material and data – Part 3: Provenance of Biological Material

NWIP stage

Warning for WDs and CDs

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

To help you, this guide on writing standards was produced by the ISO/TMB and is available at https://www.iso.org/iso/how-to-write-standards.pdf

A model manuscript of a draft International Standard (known as "The Rice Model") is available at https://www.iso.org/iso/model_document-rice_model.pdf

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

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 on 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 the following URL: www.iso.org/iso/foreword.html.

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

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

1 Scope

This document specifies a general information model for a specimen and data provenance and requirements for provenance data interoperability and serialization. This document covers all steps of the specimen life-cycle from collection to analysis, any data generated or collected during the specimen life-cycle, data originating from analytical procedures applied to the specimen and results from further mathematical processing of the data.

The requirements defined here are applicable to organizations, authorities and industries collecting, processing or distributing specimen for research, generating, collecting, analyzing or storing data on specimen, or manufacturing devices or software for the aforementioned tasks or providing facilities for these tasks.

This standard defines a common provenance information model suitable for describing source material, from which the sample originates, its manipulation, generation of data from the material, as well as production of derived data. In order to achieve interoperability, serialization of the provenance information needs to be specified, too.

This document is meant as a horizontal standard stipulating a general guideline for provenance information management in biotechnology.

This document does not apply to specimen and data used for other than research or in fields that are highly regulated by national, regional or international laws, such as medical diagnosis and therapy or food and feed production. The specimen can be biological, environmental or other types relevant for biotechnology research.

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.

— ISO 8601-1:2019 Date and time – Representations for information interchange – Part 1: Basic rules *[recording dates/times in provenance information]*

— ISO 8601-2:2019 Date and time – Representations for information interchange – Part 2: Extensions *[uncertain or approximate dates]*

2.1 Suggested references

— ISO/TS 20658:2017 Medical laboratories – Requirements for collection, transport, receipt, and handling of samples *[defines minimum extent of the provenance coverage]*

— ISO 20387:2018 Biotechnology – Biobanking – General requirements for biobanking *[defines minimum extent of the provenance coverage]*

- ISO 20184-1:2018 Molecular in vitro diagnostic examinations – Specifications for pre-examination processes for frozen tissue – Part 1: Isolated RNA *[defines minimum extent of the provenance coverage]*
- ISO 20184-2:2018 Molecular in vitro diagnostic examinations – Specifications for pre-examination processes for frozen tissue – Part 2: Isolated proteins *[defines minimum extent of the provenance coverage]*
- ISO 20186-1:2019 Molecular in vitro diagnostic examinations – Specifications for pre-examination processes for venous whole blood – Part 1: Isolated cellular RNA *[defines minimum extent of the provenance coverage]*
- ISO 20186-2:2019 Molecular in vitro diagnostic examinations – Specifications for pre-examination processes for venous whole blood – Part 2: Isolated genomic DNA *[defines minimum extent of the provenance coverage]*
- ISO 20166-1:2018 Molecular in vitro diagnostic examinations – Specifications for pre-examination processes for formalin-fixed and paraffin-embedded (FFPE) tissue – Part 1: Isolated RNA *[defines minimum extent of the provenance coverage]*
- ISO 20166-2:2018 Molecular in vitro diagnostic examinations – Specifications for pre-examinations processes for formalin-fixed and paraffin-embedded (FFPE) tissue – Part 2: Isolated proteins *[defines minimum extent of the provenance coverage]*
- ISO 20166-3:2018 Molecular in vitro diagnostic examinations – Specifications for pre-examination processes for formalin-fixed and paraffin-embedded (FFPE) tissue – Part 3: Isolated DNA *[defines minimum extent of the provenance coverage]*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply. ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- IEC Electropedia: available at http://www.electropedia.org/
- ISO Online browsing platform: available at http://www.iso.org/obp

3.1
CPM

Common Provenance Model, an extension of PROV-DM for generating, maintaining, and provisioning provenance information on biological material and data

Note 1 to entry: The term already present in ISO 23494-1. DO NOT MODIFY THE DEFINITION!!!

3.2
described activity

an activity, performed on a described object

Note 1 to entry: Examples for activities performed on physical objects can be biobanking activities as specified in [ISO 20387:2018, 3.6], activities performed on digital objects can be data analytics as specified in [ISO/IEC 20546:2019(en), 3.1.6]. The term already present in ISO 23494-1. DO NOT MODIFY THE DEFINITION!!!

3.3
described object
a physical or digital object

Note 1 to entry: The term already present in ISO 23494-1. DO NOT MODIFY THE DEFINITION!!!

3.4
finalized provenance information
provenance information transformed into a representation specified by the Common Provenance Model (CPM) and which is prepared to be conserved or archived and which is considered immutable

Note 1 to entry: The term already present in ISO 23494-1. DO NOT MODIFY THE DEFINITION!!!

3.5
non-repudiation of origin
service intended to protect against the originator's false denial of having created the content of a message and of having sent a message

[SOURCE: ISO 13888:2009, 3.30[1]]

3.6
opaque provenance component
part of the finalized provenance information that is findable but not directly accessible, and is stored as an opaque object at the respective responsible subject

Note 1 to entry: The term already present in ISO 23494-1. DO NOT MODIFY THE DEFINITION!!!

3.7
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 etc. The term already present in ISO 23494-1. DO NOT MODIFY THE DEFINITION!!!

[SOURCE: ISO 20387:2018, 3.37[2]]

3.8
primary sample
discrete portion of a biological material taken for examination, study or analysis of one or more quantities or properties assumed to apply for the whole

3.9
PROV activity
a provenance structure, that represents something that occurs over a period of time and acts upon or with PROV entities and it may include consuming, processing, transforming, modifying, relocating, using, or generating PROV entities

Note 1 to entry: Modification of the original definition in PROV-DM in that the definition explicitly refers to a provenance structure. The term is prefixed by ”PROV” for disambiguation.

[SOURCE: W3C PROV-DM:2013[3]]

3.10
PROV agent
a provenance structure, that represents something or someone that bears some form of responsibility for an PROV activity taking place, for the existence of an PROV entity, or for another agent’s PROV activity

Note 1 to entry: Modification of the original definition in PROV-DM in that the definition explicitly refers to a provenance structure. The term is prefixed by ”PROV” for disambiguation.

[SOURCE: W3C PROV-DM:2013[3]]

3.11
PROV derivation
a provenance structure, that represents transformation of an PROV entity into another, an update of an PROV entity resulting in a new one, or the construction of a new PROV entity based on a pre-existing PROV entity

Note 1 to entry: Modification of the original definition in PROV-DM in that the definition in this document explicitly refers to a provenance structure. The term is prefixed by ”PROV” for disambiguation.

[SOURCE: W3C PROV-DM:2013[3]]

3.12
PROV entity
a provenance structure, that represents physical, digital, conceptual, or other kind of thing with some fixed aspects which may be real or imaginary

Note 1 to entry: Modification of the original definition in PROV-DM in that the definition in this document explicitly refers to a provenance structure. The term is prefixed by ”PROV” for disambiguation.

[SOURCE: W3C PROV-DM:2013[3]]

3.13
PROV usage
a provenance structure, that represents beginning of utilizing an PROV entity by an PROV activity

Note 1 to entry: Modification of the original definition in PROV-DM in that the definition in this document explicitly refers to a provenance structure. The term is prefixed by ”PROV” for disambiguation.

[SOURCE: W3C PROV-DM:2013[3]]

3.14
provenance information
information that documents the history of a described object and any described activities related to that object, including information on the origin or source of the described object and the chain of custody

Note 1 to entry: The term already present in ISO 23494-1. DO NOT MODIFY THE DEFINITION!!!

3.15
provenance structure
an object in finalized provenance information, such as a PROV entity, PROV activity, PROV agent or a relation

Note 1 to entry: In PROV-DM the term is used without an exact definition. PROV-DM distinguishes core structures which form the essence of provenance information, and extended structures, that enhance and refine core structures with more expressive capabilities.

Note 2 to entry: In PROV-DM the term is used in the context of provenance information in general, here it is used only in the context of finalized provenance information

3.16
responsible subject
an institution or the organizational unit it is part of, which is responsible for providing and assembling provenance information documenting a described activity on a described object it is involved in.

Note 1 to entry: The term already present in ISO 23494-1. DO NOT MODIFY THE DEFINITION!!!

3.17
sample
portion of a whole

[SOURCE: ISO 20387:2018, 3.45[2]]

3.18
SOP provenance structure
a provenance structure, that represents a standard operating procedure

3.19
standard operating procedure

4 Common Entities

4.1 Sample Entity

A sample is the central entity describing the provenance of biological material. A sample entity specialization of the `prov:entity`, written as `entity(id, [prov:type='cpm:sample', cpm:sampleType, cpm:samplePreservat cpm:sampleQuantity, cpm:sampleIdentifier, cpm:sampleContainerType, cpm:sampleCharacterization...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:sampleIdentifier*: a mandatory attribute describing the external ID of the sample

— *cpm:sampleType*: a mandatory value for the type of a sample

— *cpm:samplePreservation*: a mandatory value for the preservation of a sample

— *cpm:sampleQuantity*: an optional attribute for the quantity of the quantity of the primary sample

— *cpm:sampleContainerType*: an optional attribute describing the container type the sample is stored

Within the provenance documentation information about a container shall be documented and changing the sample container shall be a the processing activity. A sample shall either

— contain the information about the sample container type as an attribute, in this case the identifier of the sample, which is contained at the physical container, e.g. as a barcode or RF-ID, shall be an attribute of the sample.

— be related to a container by the attribute 'contained in'. In this case the identifier of the sample is an attribute of the container entity.

A container contains either samples and/or other containers, and can so organise samples in a hierarchical way. Entities in a container can be either unordered or ordered. For a ordered container the position of the entities shall be given in the the provenance documentation.

4.2 Container Entity

A container entity specialization of the `prov:entity`, written as `entity(id, [prov:type='cpm:container', cpm:containerType, cpm:constainerIdentifier, cpm:containerVolume, cpm:manufacturer,...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:containerType*: a mandatory value for the type of a container

— *cpm:constainerIdentifier*: a mandatory attribute describing the external ID of the container

— *cpm:containerVolume*:an optional attribute describing the storage capacity of a container

— *cpm:manufacturer*: an optional attribute describing the manufacturer of a container

4.3 SOP entity

SOP provenance structure is a provenance structure, that represents a standard operating procedure. Relation to the W3C PROV `cpm:sop` rdfs:subClassOf: `prov:entity`; rdfs:subClassOf: `prov:plan`

An SOP provenance structure, specialization of the `prov:plan`, written as `entity(id, [prov:type='cpm:sop', cpm:primaryId, cpm:sopPublicationDate, cpm:sopEffectivityDate, cpm:documentName, cpm:hashValue ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:sop*: a mandatory value for the *prov:type* attribute

— *cpm:primaryId*: a mandatory attribute

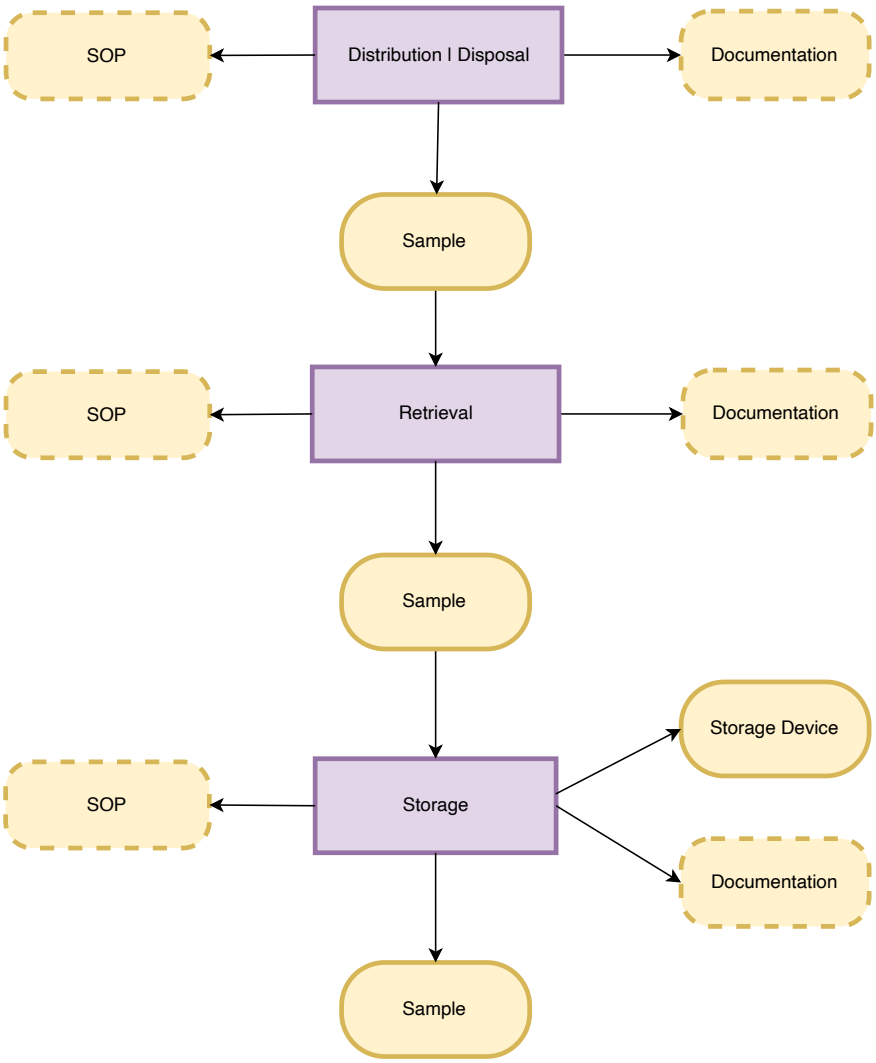
— *cpm:hashValue*: a mandatory attribute

— *cpm:sopPublicationDate*: an optional attribute

— *cpm:sopEffectivityDate*: an optional attribute

— *cpm:documentName*: an optional attribute

5 Provenance of Distribution, Disposal, Retrieval and Storage of Biological Material


SOP
Distribution I Disposal
Documentation
Sample
SOP
Retrieval
Documentation
Sample
Storage Device
SOP
Storage
Documentation
Sample

5.1 General

The provenance shall start either with an activity for distribution / disposal, linked to an activity of retrieval and storage.

5.2 Distribution/Disposal Activity

A distribution / disposal activity, specialization of the `prov:activity`, written as `activity(id, [prov:type='cpm:dist cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:timeStamp*: a mandatory value denoting retrieval according to ISO 8601

— *cpm:sop*: an optional attribute denoting the standard operating procedure the acquisition activity follows

— *cpm:documentation*: an optional attribute denoting the documentation of the activity and/or documentation of deviations towards the SOP

in case of a distribution the activity shall be linked to the receiving activity. If a sample is disposed, i.e. physical not more existent, also the corresponding PROV entity shall be invalidated.

5.3 Retrieval Activity

A retrieval activity, specialization of the `prov:activity`, written as `activity(id, [prov:type='cpm:retrieval', cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:timeStamp*: a mandatory value denoting retrieval according to ISO 8601

— *cpm:longTermStorageCurve*: an optional value denoting storage temperature profile

— *cpm:destinitionTemperature*: an optional value denoting temperature of the destination area

— *cpm:sop*: an optional attribute denoting the standard operating procedure the acquisition activity follows

— *cpm:documentation*: an optional attribute denoting the documentation of the activity and/or documentation of deviations towards the SOP

— *cpm:storageCharacterization*: an optional attribute providing relevant information about the receivement in a specific domain as ...

Every re-movement of a sample from the storage device shall be documented by a retrieval activity.

5.4 Storage Activity

A storage activity, specialization of the `prov:activity`, written as `activity(id, [prov:type='cpm:storage', cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:timeStamp*: a mandatory value denoting storage according to ISO 8601

— *cpm:longTermStorageTemperature*: a mandatory value denoting the storage temperature

— *cpm:storageDevice*: a mandatory value denoting the storage device

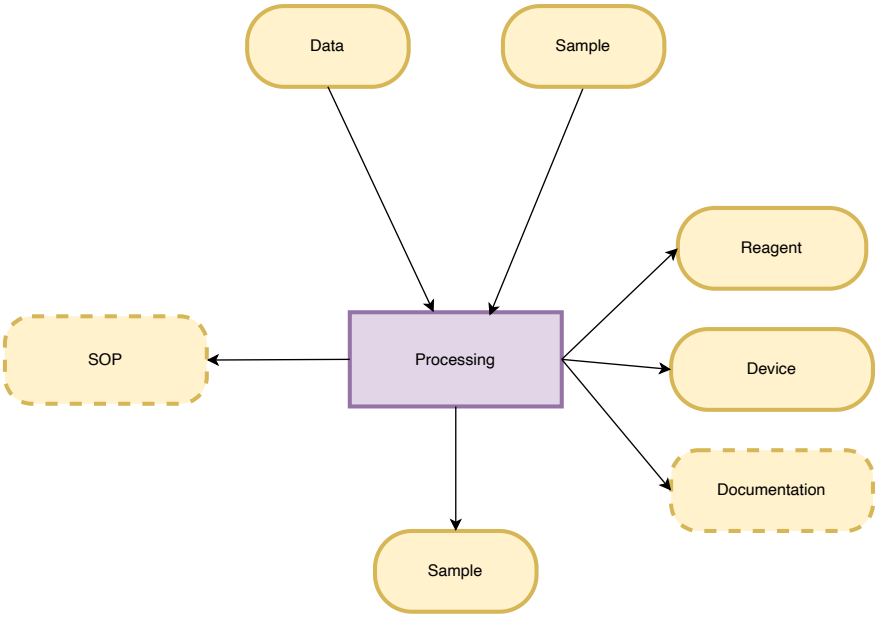
— *cpm:sop*: an optional attribute denoting the standard operating procedure the acquisition activity follows

— *cpm:documentation*: an optional attribute denoting the documentation of the activity and/or documentation of deviations towards the SOP

— *cpm:storageCharacterization*: an optional attribute providing relevant information about the storage in a specific domain as ...

The storage activity shall be linked to an entity representing the storage device.

6 Provenance of Processing of Biological Material


Data
Sample
Reagent
SOP
Processing
Device
Documentation
Sample

In the processing activity a sample together with optional data is transformed into a new samples and/or data. There are the following cases:

a) The processing of a sample and/or samples, where no new sample is generated, but the original sample is changed. In this case a new sample entity shall be derived with the same external ID

b) The processing of a sample and/or samples, where a new physical sample is derived. In this case the input sample is changed but will not get a new external ID. In a special cases the input sample is completely consumed or disposed. In this case the entity representing the input sample shall be invalidated.

c) The processing of a sample and/or samples, where the original sample is not changed. In this case no derived sample entity shall be generated. Example for such cases would be sample measurements, quality control, in this to cases only a data entity is generated.

6.1 Processing Activity

A processing activity, specialization of the `prov:activity`, written as `activity(id, [prov:type='cpm:processing', cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:timestamp*: acquisition time and/or data according to ISO 8601

— *cpm:sop*: an optional attribute denoting the standard operating procedure the acquisition activity follows

— *cpm:documentation*: an optional attribute denoting the documentation of the activity and/or documentation of deviations towards the SOP

Used reagents shall be linked with the processing activity with the PROV usage relation.

Used devices shall be linked with the processing activity with the PROV usage relation.

A reagent entity specialization of the `prov:entity`, written as `entity(id, [prov:type='cpm:reagent', cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

—

A device entity specialization of the `prov:agent`, written as `entity(id, [prov:type='cpm:device', cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

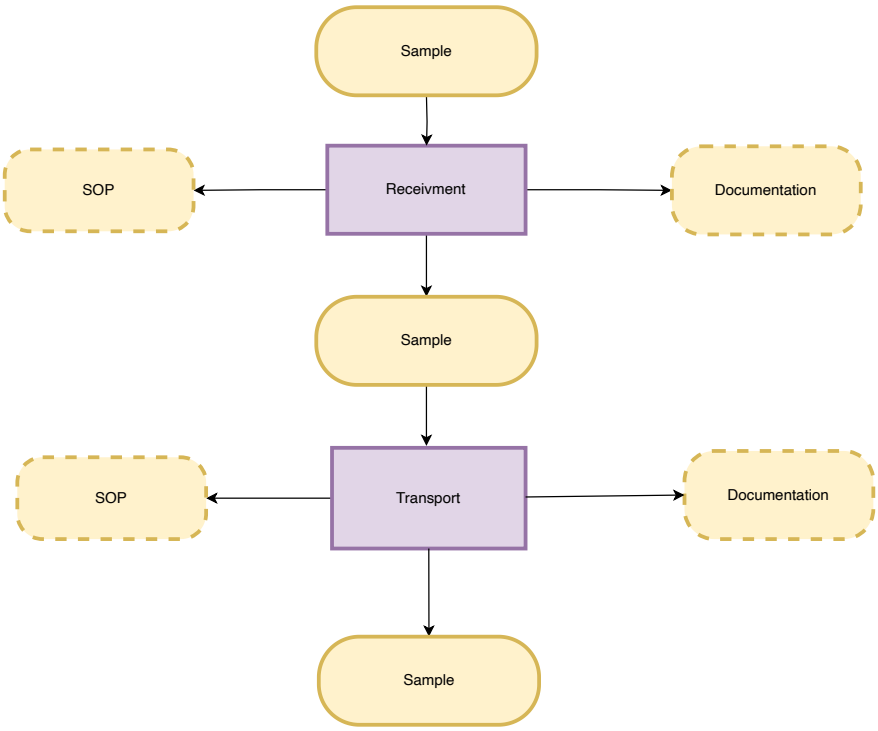
— *prov:type*: a mandatory attribute

—

7 Provenance of Receivement and Transportation of Biological Material

7.1 General

The provenance describing a sample and/or samples entering a biobank and/or laboratory shall include an activity for receivment and transportation. It should include a SOP entity, if the receivment/transport follows a specific procedure. It should include a documentation entity summarizing the storage process and/or deviation to the attached SOP.


Sample
SOP
Receivment
Documentation
Sample
SOP
Transport
Documentation
Sample

7.2 Receivement Activity

A receivement activity, specialization of the `prov:activity`, written as `activity(id, [prov:type='cpm:receivement' cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:timeStamp*: a mandatory value denoting receivement according to ISO 8601

— *cpm:transportEnd*: a mandatory value denoting end of the transport activity according to ISO 8601

— *cpm:sop*: an optional attribute denoting the standard operating procedure the acquisition activity follows

— *cpm:documentation*: an optional attribute denoting the documentation of the activity and/or documentation of deviations towards the SOP

— *cpm:receivementCharacterization*: an optional attribute providing relevant information about the receivement in a specific domain as ...

7.3 Transport Activity

A transport activity, specialization of the `prov:activity`, written as `activity(id, [prov:type='cpm:transport', cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:timeStampStart*: a mandatory value denoting the start of the transport activity according to ISO 8601

— *cpm:transportEnd*: a mandatory value denoting end of the transport activity according to ISO 8601

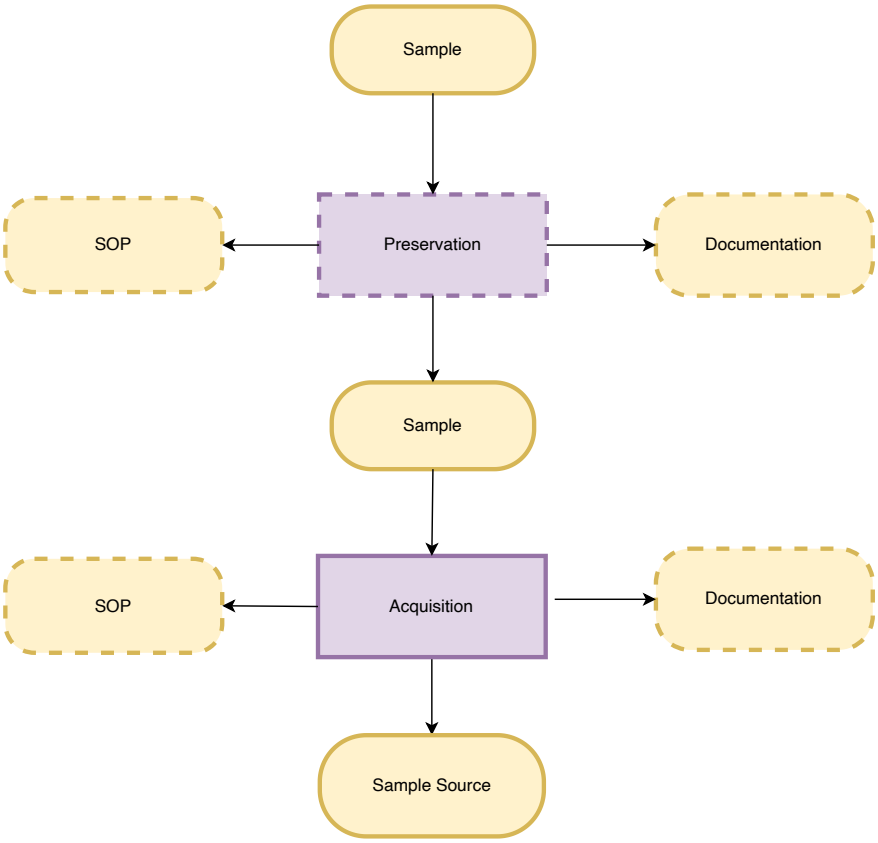
— *cpm:transportIdentifiers*: external identifiers used ti identify the transport process

— *cpm:sop*: an optional attribute denoting the standard operating procedure the acquisition activity follows

— *cpm:documentation*: an optional attribute denoting the documentation of the activity and/or documentation of deviations towards the SOP

— *cpm:transportCharacterization*: an optional attribute providing relevant information about the transport in a specific domain as ...

8 Provenance of Acquisition of Biological Material


Sample
SOP
Preservation
Documentation
Sample
SOP
Acquisition
Documentation
Sample Source

8.1 General

The provenance of the acquisition of biological material shall include an activity for preservation and acquisition. It shall contain entities for the sample source and an entity ready for transport. It shall contain actors for any activities. It should include a SOP entity, if the acquisitions follows a specific procedure. It should include a documentation entity summarizing the acquisition process and/or deviation to the attached SOP.

8.2 Acquisition Activity

An acquisition activity, specialization of the `prov:activity`, written as `activity(id, [prov:type='cpm:acquisition' cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:timestamp*: acquisition time and/or data according to ISO 8601

— *cpm:context*: a mandatory value for the acquisition context, e.g. medical treatment, research, quality control

— *cpm:method*: a mandatory attribute describing the acquisition method

— *cpm:acquisitionSource*: a mandatory attribute describing the the acquisition site, e.g. a farm, a hospital, etc

— *cpm:geographicLocation*: a mandatory attribute describing the geographical location of the acquisition site

— *cpm:sop*: an optional attribute denoting the standard operating procedure the acquisition activity follows

— *cpm:documentation*: an optional attribute denoting the documentation of the activity and/or documentation of deviations towards the SOP

The acquisition activity shall result in a sample entity, see section 4.1

8.3 Preservation, Processing Activity

The preservation activity shall be connected to entity and generate another entity with changed and/or additional attributes. The attributes shall follow the generic processing activity and shall include the following documentation:

a) any additions or modification of the sample after acquisition. e.g. labeling

b) any preservation procedure, e.g. fixation, stabilization, moisturising to prevent drying of the surface, etc. Both the preservation technique uses as well as additives/preservatives shall be documented with domain specific standards.

c) any specific packaging procedure, e.g. vacuum packaging, cooling, using container with pre-filled buffering solutions, etc.

d) for critical activities the temperature monitoring shall be included as an attribute

The acquisition activity shall be connected to a responsible actor and the source of the biological material entity.

8.4 Source Entity

The source of the biological material is a specialization of the `prov:entity`, written as `entity(id, [prov:type='cpm:so cpm:, cpm:, cpm:, cpm:, cpm: ...])` in PROV-N, has:

— *id*: an identifier property

— *prov:type*: a mandatory attribute

— *cpm:sourceTaxonomy*: a mandatory value for the taxonomy of the source

— *cpm:sourceIdentifier*: a mandatory unique identifier of the source of biological material

— *cpm:consentInformation*: an optional attribute describing consent information

9 Requirements

9.0.1 A described object represented by SOP provenance structure shall be digital (not physical). Hash value of the standard operating procedure being represented by a PROV entity shall be included in provenance information using the `cpm:hasValue` attribute of particular PROV entity.

NOTE: Integrity verification of standard operating procedures is easier to perform for digital standard operating procedures.

9.0.2 If a physical described object leaves controlled environment, a new derived PROV entity shall be created to represent the PROV entity in the new environment. The original and new entities shall be related using the PROV derivation relation.

NOTE: If a physical described object leaves controlled environment, it is presumed that the environmental conditions may change and those may affect state of the described object. The change may be not obvious from the first observation of the described object, thus leaving the controlled environment shall be always documented.

Appendix A
(informative)

CEN/TS Documentation Examples

The following examples have been implemented by BBMRI.at in collaboration with BBMRI-ERIC for assessment of compliance with CEN/TS specifications. They are not for public distribution beyond this WG and beyond BBMRI-ERIC nodes.

A.1 Encoding activities

Each activity needs to be encoding using the following attributes, in compliance with :

A.1.1 Semantic description of the process occuring.

A.1.2 Timestamp (date and time)s of activity beginning and end in ISO 8601 format.

A.1.3 0..n parameters of an activity, where each parameter must be encoded as a triplet:

— Semantic link of the parameters.

— Value

— Unit (empty if unit-less or irrelevant)

A.1.4 Actor performing the activity (or responsible for the activity in case of automated process).

A.2 Identified activities, their parameters, and mapping to ontologies

Activity: Start of warm ischemia (vessel ligation/arterial clamping time)

Ontology term(s):

Beginning timestamp: timestamp rounded to minutes

Ending timestamp: N/A

Parameters: –

Activity: Transportation after warm ischemia

Ontology term(s):

Beginning timestamp: timestamp rounded to minutes

Ending timestamp: timestamp rounded to minutes

Parameters:

- *Minimum transport temperature*
  Data type: float
  Permitted units: K *Level:* required
- *Maximum transport temperature*
  Data type: float
  Permitted units: K *Level:* required
- *Transport medium*
  Data type: enum (air, ice, dry ice, LN)
  Permitted units: N/A *Level:* optional

- *Deviations description*
  Data type: string
  Permitted units: N/A *Level:* optional

Activity: Material condition check

Ontology term(s):

Beginning timestamp: timestamp rounded to minutes

Ending timestamp: N/A

Parameters:

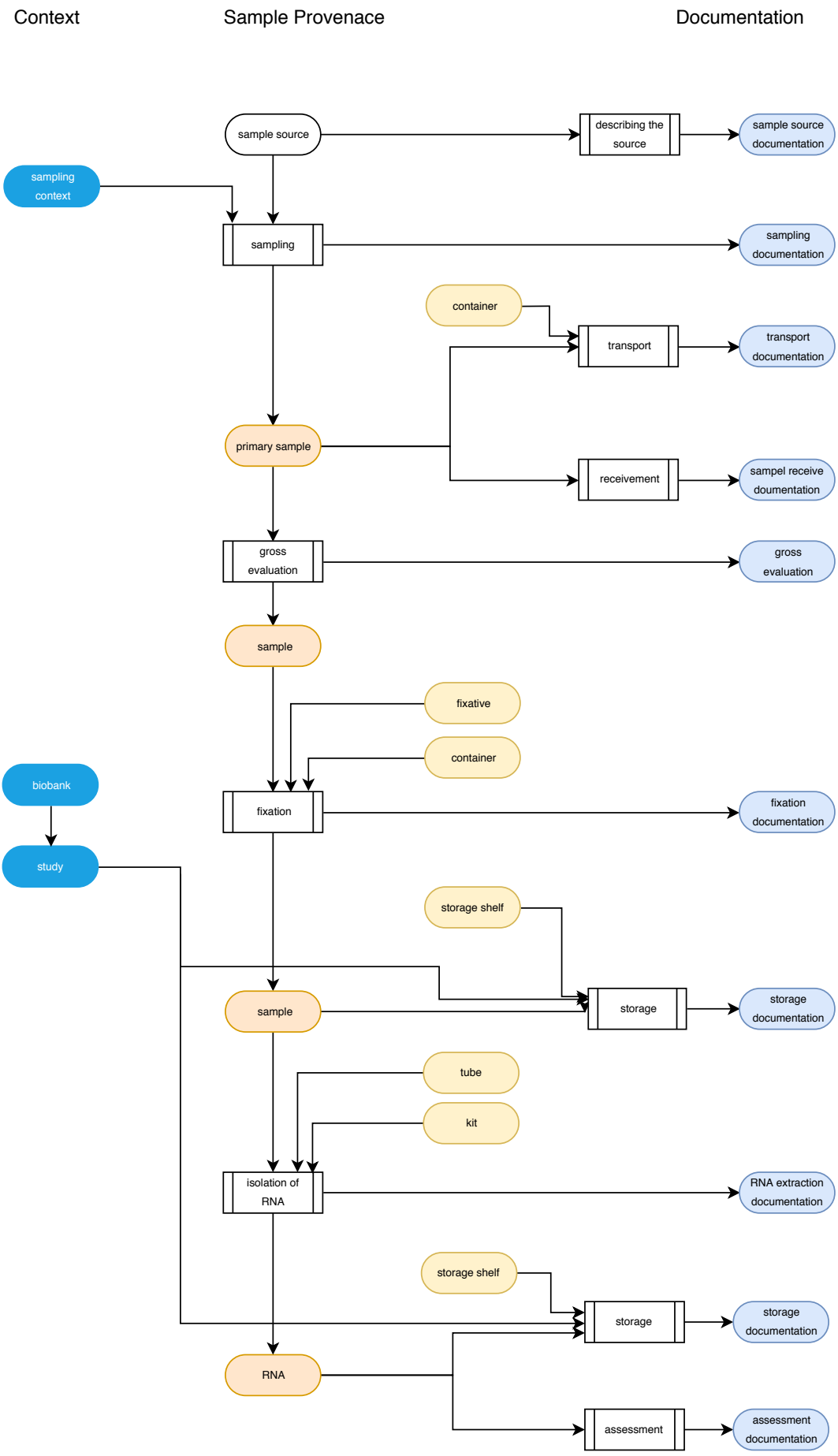
- *Labelling of transport container documented*
  Data type: enum (y/p/n)
  Permitted units: N/A *Level:* required
- *Leaking/breaking of transport container documented*
  Data type: enum (y/p/n)
  Permitted units: N/A *Level:* required
- *Transport conditions including temperature documented*
  Data type: enum (y/p/n)
  Permitted units: N/A *Level:* required
- *Tissue type and quantity in transport container documented*
  Data type: enum (y/p/n)
  Permitted units: N/A *Level:* required
- *Other deviation from the transport protocol documented*
  Data type: enum (y/p/n)
  Permitted units: N/A *Level:* required

A.3 Necessary links

A.3.1 Patient ID or link to previous patient ID (when recoding pseudonyms).

A.3.2 Link to health status documentation of the patient.

A.3.3 Link to the medical treatment documentation.


Context
Sample Provenace
Documentation
sample source
describing the source
sample source documentation
sampling context
sampling
sampling documentation
container
transport
transport documentation
primary sample
receivement
sampel receive doumentation
gross evaluation
gross evaluation
sample
fixative
container
biobank
fixation
fixation documentation
study
storage shelf
sample
storage
storage documentation
tube
kit
isolation of RNA
RNA extraction documentation
storage shelf
storage
storage documentation
RNA
assessment
assessment documentation

A.4 CEN/TS Provenance Example

```
:sampleSource
    a prov:Entity
    sprov:donorID                   ”someDonorID”;
    sprov:healthStatusICD10         ”C50.1”
    sprov:treatmentDocumentation    treamtmentEntity;
    sprov:startOfIchemia            ”2011-07-16T01:52:02Z”^^xsd:dateTime;

:Haahashtari
   a foaf:Person, prov:Agent, sprov:Medical;
   foaf:givenName       ”Haahashtari”;
   foaf:mbox            <mailto:haahashtari@example.org>;
   prov:actedOnBehalfOf :samplingContext;

:sampling
    a prov:Activity
    prov:wasAssociatedWith :Haahashtari;

:primarySample
    a prov:Entity
    prov:generatedAtTime   ”2011-07-16T01:52:02Z”^^xsd:dateTime;
    prov:wasGeneratedBy    :sampling

:sampleSourceDescription
    a prov:Activity
    prov:generatedAtTime   ”2011-07-16T01:52:02Z”^^xsd:dateTime;
    prov:wasAssociatedWith :Haahashtari;

:samplingDocumentation
    a prov:Entity
    prov:wasGeneratedBy    :sampleSourceDescription
    sprov:documentation    :samplingDocumentation

:Habakkuk,
   a foaf:Person, prov:Agent, sprov:Biobank;
   foaf:givenName       ”Habakkuk”;
   foaf:mbox            <mailto:habakkuk@example.org>;
   prov:actedOnBehalfOf :biobank;

:transportContainer
    a prov:Entity

:transport
    a prov:Activity
    prov:startedAtTime     ”2011-07-14T01:01:01Z”^^xsd:dateTime;
    prov:wasAssociatedWith :Habakkuk;
    prov:used              :primarySample;
    prov:used              :transportContainer;
    sprov:tempertureRange  ”below 20 degrees”;
    prov:endedAtTime       ”2011-07-14T02:02:02Z”^^xsd:dateTime;

:transportDocumentation
     a prov:Entity
     sprov:startOfTransport     ”2011-07-16T01:52:02Z”^^xsd:dateTime;
     sprov:transportSOP         :transportSOP
     sprov:deviationAtTransport ”None”;

:Hagar
   a foaf:Person, prov:Agent, sprov:Biobank;
   foaf:givenName       ”Hagar”;
   foaf:mbox            <mailto:hagar@example.org>;
```



```
   prov:actedOnBehalfOf :biobank;

:receivment
    a prov:Activity
    prov:wasAssociatedWith :Hagar;
    prov:startedAtTime    ” 2011-07-14T01:01:01Z”^^xsd:dateTime;
    prov:used              :primarySample;
    prov:used              :transportContainer;

:sampleReceivmentDocumentation
    a prov:Entity
    prov:wasGeneratedBy    :receivment;
    prov:value             ”No deviation to SOP”;


:Hareph
   a foaf:Person, prov:Agent, sprov:Biobank;
   foaf:givenName       ”Hareph”;
   foaf:mbox            <mailto:hareph@example.org>;
   prov:actedOnBehalfOf :pathology;

:grossEvaluation
    a prov:Activity
    prov:wasAssociatedWith :Hareph;

:grossEvaluationReport
    a prov:Entity
    prov:generatedAtTime   ”2011-07-16T01:52:02Z”^^xsd:dateTime;
    prov:wasGeneratedBy    :grossEvaluation


:fixation
    a prov:Activity

:storageOfFFPEtissue
    a prov:Activity

:isolaionOfRNA
    a prov:Activity

:storageOfRNAsample
    a prov:Activity

:evaluationOfRNAsample
    a prov:Activity
```

Appendix B
(informative)

Tissue Engineering in Research: Microfluidic Cell Culture Devices to Investigate Disease Models and Drug Response of Living Tissue Recapitulated In Vitro.

The aim of the study of Occhetta et al. [4] is to develop an in vitro scar-model using the microfluidic organs-on-chip technology to investigate the key phases that leads to the scar formation upon myocardial injury and how this process is affected by biochemical and mechanical factors in a setting as closely as possible to the native one. Such a model, which is possibly usable to develop anti-fibrosis therapeutic solutions, is the result of a series of steps involving biological material manipulation, data acquisition and processing that expose provenance related issues. With the aim of having a grounding knowledge of these activities in order to set out the definition of the provenance requirements, a brief description of the main phases is given below.

a) **Cell isolation and characterization.** Fibroblast, cells responsible for the synthesis of extracellular matrix and collagen, were isolated from neonatal rat hearts following a standardized procedure that consists in a series of steps in which the ventricles are bathed in different solutions featured by time, temperature and rpm. The cells were seeded and cultured iteratively using different growth medium until a predefined confluence is reached. Here, the initial cell population may be assessed by image analysis after immunofluorescence staining. Fibroblast were then frozen and stored in liquid nitrogen.

   Main provenance aspects are related to:

   - Physical tracking of objects: freezing and storage parameters.
   - Experimental activities: source of biological material (donor species, cell type); cell isolation protocol (procedure, time, temperature, solution used, etc.); culture parameters (time, growth medium, etc.); staining procedure (type of stain, e.g., DAPI).
   - Computational analysis: image acquisition and analysis software.

b) **Cell culture.** After thawing, fibroblasts were further cultured within incubators, under specific temperature, humidity and CO_2 level, and detached through a dissociation solution. A suspension of cells were finally injected within microfluidic devices following a procedure that involves the mixing of two ice-cold solutions and the incubation of the devices to promote the polymerization of the hydrogel contained in the central channel. The growth medium was supplemented via the side channels with or without a biochemical substance that support the transition of fibroblasts towards myofibroblast. The fibrin hydrogel was also exposed to a cyclic mechanical stimulation that affect the myocardial remodeling.

   Main provenance aspects are related to:

   - Experimental activities: incubation parameters (temperature, humidity, CO2 amount, growth medium, time); processing procedure (cell passaging, dissociation, suspension in hydrogel, injection, incubation); biochemical factor supplemented; mechanical stimuli (strain direction, magnitude, frequency, etc.).

c) **Immunofluorescence staining.** After some day of culture, samples within the devices were prepared for the immunofluorescence analysis (fixation, incubation to permeabilize cells and block nonspecific

bindings, treating with primary antibodies). Different types of fluorescently labeled secondary antibodies were used to evaluate fibroblast proliferation, phenotype switch and matrix deposition, while DAPI staining was used to recognize the nuclei.

Main provenance aspects are related to:

- Experimental activities: samples preparation (fixation method, permeabilization agent, blocking, incubation parameters, antibodies category, DAPI staining, etc.).

d) **Image analysis.** Sample images were taken using a confocal microscope and analyzed via software to quantify cellular density, the amount of proliferating cells and the phenotype switching. Data were presented all as mean plus standard deviation for representative images of three different tissue regions.

Main provenance aspects are related to:

- Computational analysis: image acquisition system, software used to quantify tissue parameters; calculation algorithm.

e) **Quantitative RT-PCR.** RNA was isolated from the synthesized micro-scar tissue and quantitative real-time reverse transcriptase PCR was performed to obtain cDNA and investigate the presence of some markers which are typical of the fibrotic tissue. The expression level of each gene were normalized and calculated according to specific methodology ($2^{\Delta Ct}$ method).

Main provenance aspects are related to:

- Experimental activities: sample preparation (RNA isolation, reverse-transcription protocol and kit).
- Computational analysis: system used; quantitative information for mRNA targets; algorithm(s) used.

f) **AFM mechanical characterization.** In order to characterize the construct from a mechanical point of view, the actuator and the membrane were removed from the device and the culture chamber was glued onto a plastic dish. Measurement of the local Young's modulus was performed through atomic force microscopy nano-indentation using a mechano-optical microscope. From each sample, three force maps (consisting in force-displacement curves) were recorded in different locations. A specific model for calculating the elastic modulus was used, and the results consists in a histogram and the corresponding mean and standard deviation.

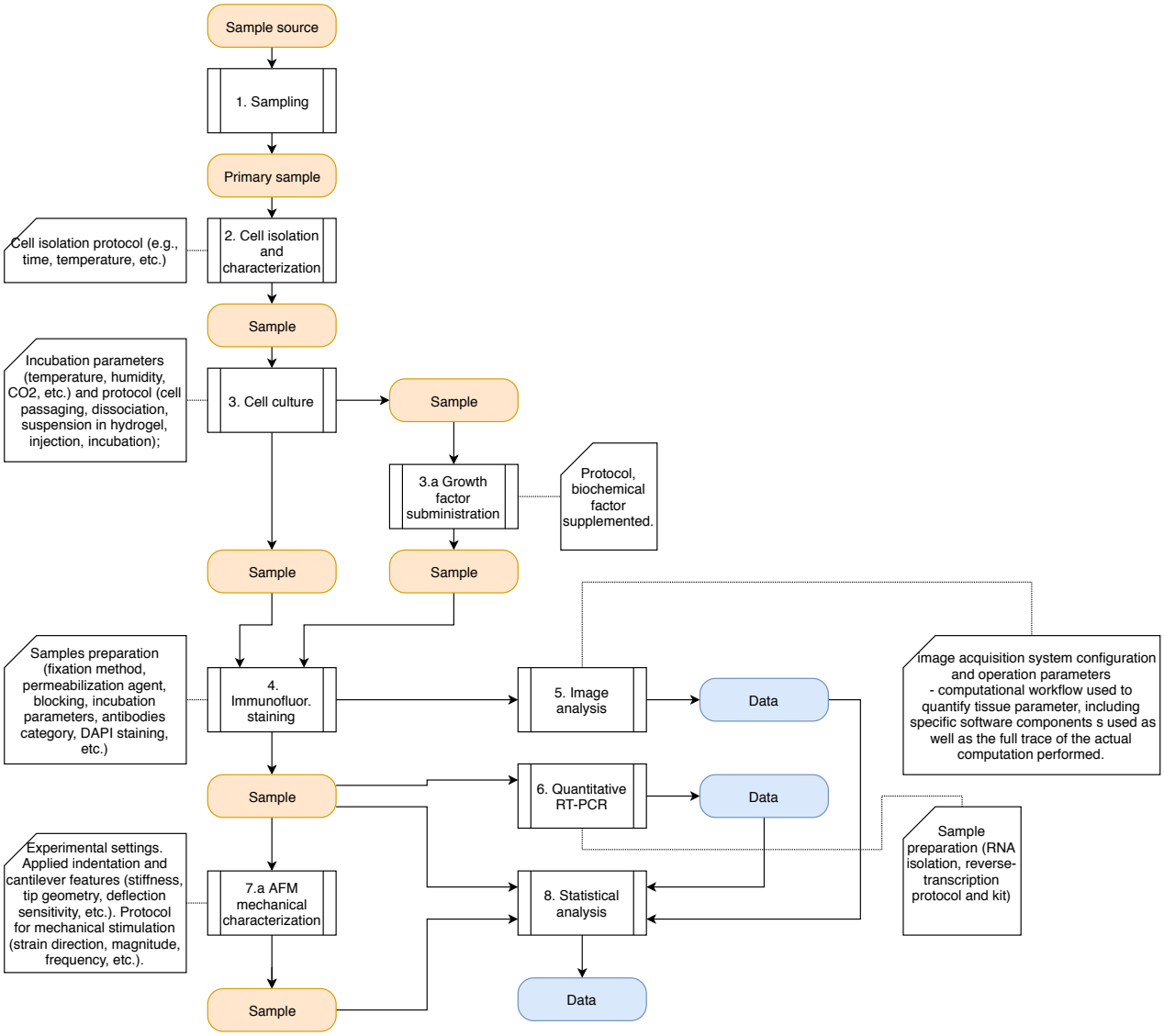
Main provenance aspects are related to:

- Experimental activities: experimental settings; applied indentation and cantilever features (stiffness, tip geometry, deflection sensitivity, etc.).
- Computational analysis: force map acquisition (location, size, resolution, curves); correction factors; calculation and normalization methods.

g) **Statistical analysis.** Different type of multiple comparison test was performed among different experimental groups for immunofluorescence analysis, quantitative RT-PCR and AFM measurement results. Mean and standard deviation were presented as results with the respective significance level for each of them.

Main provenance aspects are related to:

- Computational analysis: algorithm parameters.


Sample source
1. Sampling
Primary sample
Cell isolation protocol (e.g., time, temperature, etc.)
2. Cell isolation and characterization
Sample
Incubation parameters (temperature, humidity, CO2, etc.) and protocol (cell passaging, dissociation, suspension in hydrogel, injection, incubation);
3. Cell culture
Sample
3.a Growth factor subministration
Protocol, biochemical factor supplemented.
Sample
Sample
Samples preparation (fixation method, permeabilization agent, blocking, incubation parameters, antibodies category, DAPI staining, etc.)
4. Immunofluor. staining
5. Image analysis
Data
image acquisition system configuration and operation parameters - computational workflow used to quantify tissue parameter, including specific software components s used as well as the full trace of the actual computation performed.
Sample
6. Quantitative RT-PCR
Data
Sample preparation (RNA isolation, reverse-transcription protocol and kit)
Experimental settings. Applied indentation and cantilever features (stiffness, tip geometry, deflection sensitivity, etc.). Protocol for mechanical stimulation (strain direction, magnitude, frequency, etc.).
7.a AFM mechanical characterization
8. Statistical analysis
Sample
Data

Appendix C
(informative)

Body Fluids Example

Time data
Documentation of time stamps for each step is fundamental. For later analysis a consistent representation of the time data is required.

Questions to discuss:

— allow absolute and/or relative time data

— absolute: which format? Preferably numerical, eg. Unix time format, Julian day number

— relative: related to defined fix point in the process; common to all sample processing work-flows and exactly defined, eg. beginning of long term storage

Example use cases:

— sample collection time not exactly known, transport time documented, time stamp for arrival at laboratory and following processing steps until final storage available

— query for samples with less than 1h from sampling to freezing

Data representation

Representation of data (data types and formats) must be defined prior to definition of items.

At least the following questions should be solved:

— will binary data be supported and if so which formats should be supported (ISO, IEEE, others)

— will free text be supported

— will multiple formats for the same datatype be allowed

— will ranges for numeric data be defined

— will the number of decimal places be limited

Data quality indicators
For any data quality criteria and precision must be documented along with the data. Quality levels could be:

— measured, calibrated (metrological traceability)

— measured, verified

— measured uncalibrated

— inferred from other information

— estimated

Examples:

— temperature measured with thermometer calibrated against a reference

— temperature measured with thermometer calibrated against in house reference

— temperature measured with simple, uncalibrated thermometer

— inferred from thermostat setting

— temperature estimated from ambient temperature

Precision of data can be given as:

— standard deviation or coefficient of variation

— maximal deviation

— range

Example use cases:

— freezer set to -80°C, monitored with calibrated thermometer, accepted range: -90 .. -70°C

— sample transport without temperature monitoring, outdoor temperature 25°C

— sample processing in air conditioned laboratory thermostat set to 20°C

Units
For any numerical data an appropriate unit must be provided (dimensionless data should have a unit of 1)

At least the following questions should be solved:

— will multiple units for the same data be supported

— will non-standard units be supported

— if so: will these units be converted prior to loading

Prototypes
Several procedures during the processing of specimen are common to most, if not all workflows. Examples for prototypic procedures could be transport, transfer, storage, or preservation. For these procedures the attributes can be defined generally which would reduce redundancy and improve comparability. Possible attributes for these procedures could be:

— Transport: start, end, temperature, method (manual, car, conveyor system, pneumatic tube), speed, acceleration

— Transfer: method (manual, automated), temperature, container, material

— Storage: start, end, temperature, interruption

— Preservation: method (additives, speed of freezing

C.1 General remarks

Any procedure applied to a sample must be regarded relevant for documentation as it might affect sample properties and thus, fitness for future use. To enable a consistent and extensible data model a generalized concept for provenance information should be set up. An appropriate basis could be the SOSA ontology which provides a framework for a high level description of the sample processing work-flow.

C.2 Body fluids provenance information considerations

Body fluids are collected from the subject as native material and often immediately stabilized with additives. Provenance information may include information on the subject, which should, however, be limited to information relevant for preanalytics. Additional information like biometrics, health status etc. should not be included. The sample material should be specified according to the actual processing state. Subject related provenance information may include: preparation of the subject (tourniquet, disinfectant) fasting status, position during sample collection, sampling site. In my opinion information on the health status, disease, therapy etc. should not enter into provenance information. These are phenotypic data which are not always available and if available often of questionable quality. Actually we don't even have a clear definition for a "healthy" donor. The donor identification is only valid within a particular system. Outside the system the information is at best redundant. We should most probably better go for persistent unique sample identifiers.

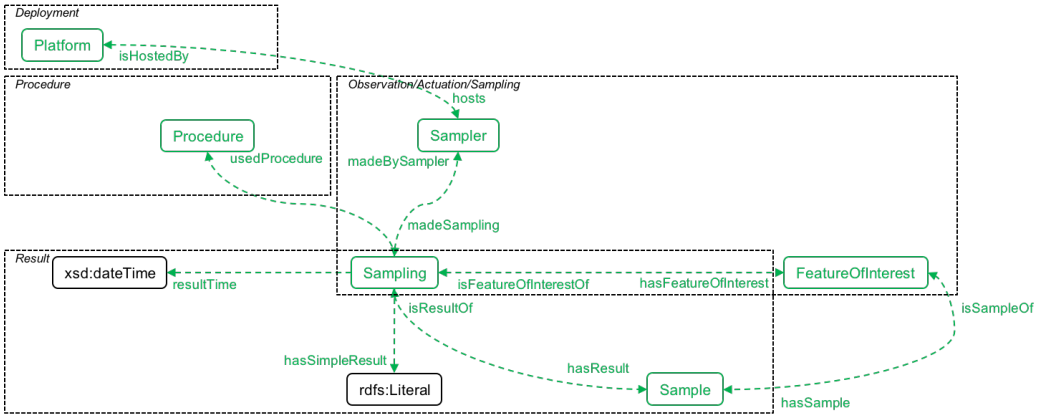


Figure 1: Overview of the SOSA classes and properties (observation perspective, https://www.w3.org/TR/vocab-ssn/images/SOSA-OntStructure-Observation.png)

Sampling procedure related provenance information may include: native material, method, sequence (for multiple samples), material (needle gauge), additives.

C.3 Body fluids provenance information elements

- Sample collection
  - Time (cf. above)
  - Place: type (institution, field ...); position: address, WGS 84 geographical coordinates, or identifier
  - Donor: identifier (pseudonym); properties: health data, status ..
  - Sample material: native material (blood, saliva, urine)
  - Method (dependent on sample material)
  - Material (dependent on method): consumables
- Transport (any change in the location and/or environmental conditions)
  - Start: Time
  - End: Time
  - Method/Means
  - Conditions: temperature, forces, radiation
- Processing (any procedure by which the sample material is modified and no longer identical to the input sample material)
  - Sample identifier
  - Start: Time
  - End: Time
  - Method
  - Conditions
  - Material: consumables, chemicals
  - Output sample material: eg. serum, plasma, cells

- Transfer (any procedure by which the sample is left unchanged but the sample container)
  - Start: Time
  - End: Time
  - Method
  - Conditions
  - Material: consumables
  - Input sample identifier
  - Output sample identifier
- Storage (any procedure by which the samples are left at a particular location for a significant (tbd.) duration
  - Start: Time
  - End: Time
  - Method
  - Conditions
  - Location

C.4 Example:

- Sample collection:
  - Time stamp: ca. 8.00h CET, 2018-02-01
  - Subject: overnight fasting, morning, sitting upright
  - Method: tourniquet venous stasis 60 mmHg, 1 min
  - Material: needle 20G, vacutainer, K-EDTA
  - Sample material:blood
- Transport:
  - Start: ca. 8.15h CET, 2018-02-01
  - End: ca. 2018-02-01 08:27:00
  - Temperature: ca. 20°C (air conditioned)
  - Method: pneumatic tube, speed 3m/s
- Transport:
  - Start: 2018-02-01 08:30:20 CET
  - End: 2018-02-01 08:44:30 CET
  - Temperature: mean 17°C, min: 14°C, max: 20°C
  - Method: bicycle courier
- Processing:
  - Start: 2018-02-01 08:49:20 CET
  - Method: centrifugation, 10min, 2000xg, 20°C (thermostatted)
  - Sample material: plasma

  - Additive: K-EDTA

- Transfer:
  - Start: 2018-02-01 09:01:20 CET
  - End: 2018-02-01 09:15:30 CET
  - Method: automated
  - Material: disposable pipette tips, .5ml cryo tubes, fluidX
  - Temperature 20˚C (air conditioned)

- Storage:
  - Start: 2018-02-01 09:19:20 CET
  - Temperature: -80˚C (calibrated monitoring, max: -70˚C)

*: required; v: recommended; o: optional

			human		animal			scientific	cells	b
Process	Record	Data	clinical	epidemiological	clinical	epidemiological	mutants	samples		
Acquisition:	timestamp		+	+	+	+	+	+	+	+
	sampling method		+	+	v	+	+	+	v	+
	collection site		+	+	+	+	+	v	+	+
	collection method	primary container type	+	+	v	+	+	v	+	+
		collection date and time	+	+	+	+	+	v	+	+
		additives, stabilizers,	+	+	+	+	+	v	+	+
		storage conditions prior to shipment	+	+	+	+	v	v	+	+
	specific properties	infectiousness	v	v	o	o	o	o	o	o
		transgenic/chimera/genetically modified etc.	v	v	v	v	+	+	+	+
Transport:	to processing/internal	mode of transportation	v	v	v	v	v	v	v	v
		temperature during transport	+	+	+	+	v	v	+	+
		temperature at reception	v	v	+	+	v	v	+	+
		duration of transport	+	+	o	o	v	v	o	o
		safety requirements.	+	v	o	o	+	+	o	+
Processing:	processing method	recording of timestamps	+	+	+	+	+	v	+	+
		monitoring temperature for critical steps	+	+	+	+	v	v	+	+
		cross-contamination	+	+	+	+	+	+	+	+
		sterility	o	o	o	+	v	v	+	+
		in-process control steps	+	+	+	+	v	o	+	+
	process output	storage container type	+	+	+	+	o	v	+	+
		sample homogeneity	+	+	+	+	+	+	+	+
		quantity	+	+	+	+	+	+	+	+
		number of aliquots	+	+	+	+	+	+	+	+
	temporary storage conditions	container type	+	+	+	+	o	v	+	+
		temperature	+	+	+	+	+	v	+	+
		duration	+	+	+	+	o	v	+	o
		humidity	o	o	o	o	o	v	+	+
		exposure to radiation (eg. light)	o	o	o	o	o	o	+	+
	preservation	cryopreservation technique used	+	+	+	+	v	v	+	+
		additives/preservatives	+	+	+	+	v	v	+	+
Testing:	Testing methods with regard to	integrity	+	+	+	+	+	+		
		cross-contamination	+	+	+	+	+	+	+	+
		sterility	o	o	o	o	+	+	+	+
		concentration	o	o	o	o	o	+	+	+
		purity	o	o	o	o	o	+	+	+
		identity	+	+	+	+	+	+	+	+
		viability	o	o	o	o	+	+	+	+
		composition of biological resources	o	o	o	o	+	+	o	o
	Requirements for the testing methods	measurement traceability	+	+	+	+	+	+	+	+
		monitoring of method	+	+	+	+	o	+	+	+
		external quality assessments	+	+	+	+	o	o	+	+
Storage:		environmental conditions	+	+	+	+	v	v	+	+
		duration	+	+	+	+	v	v	+	+
		access	+	+	+	+	v	v	+	+
		safety	+	+	+	+	v	v	+	+
		cross-contamination	+	+	+	+	+	+	+	+
		traceability	+	+	+	+	v	v	+	+
		freeze/thaw cycles	+	+	+	+	o	o	+	+
		correctness of inventory	+	+	+	+	+	+	+	+
Sample Recovery		Capacity to recover the sample	+	+	+	+	+	+	+	+
		recovery rate	o	o	o	o	o	o	+	+
		Verification of the correct genotype	o	o	o	o	+	o	+	+
		possible influence of storage conditions	v	v	v	v	v	v	+	+
		possible danger of contamination	v	v	v	v	v	v	+	+
Disposition (including destruction):		conditioning for transport	+	+	+	+	v	v	+	+
		compliance with regulatory and	+	+	+	+	+	+	+	+
		contractual requirements	+	+			+	+		
		safety	+	+	+	+	+	+	+	+
		shipment specifications	+	+	+	+	+	+	+	+
		approval by or information of	+	+	+	+	+	+	+	+
		customer (in case of destruction)	+	+						
		biological material and data consolidation	+	+	+	+	+	+	+	+
Quality control of data:		informed consent	+	+	+	+	+	+	v	v
		anonymization/pseudonymization	+	+	v	v			o	o
		clinical data	+	+	+	+			o	o
		biometric data	+	+			o	o		
		phenotypic data	+	+	+	+	o	o	+	+
		taxonomical data	o	o	+	+	+	+	+	+
		diagnosis, treatments	+	v	+	+	o	o	o	o
		epidemiological data	v	+	+	+	o	o	o	o
		environmental data of collection	v	v	+	+	v	v	o	o
		origin of specimen/sample (i.e.: inpatient,outpatient; scientific project-related; farmed species or wild)	v	+	+	+	+	+	o	o
		exposures: smoking status, diet	v	+	o	o			o	o
		demographical data:	v	+	+	+			o	o
		biological data	+	+	+	+	o	o	+	+
		biological material life cycle data	+	+	+	+	o	o	+	+
		biological material characterization data	+	+	+	+	+	+	+	+
		biosafety information	o	o	o	o	+	+	+	+

Appendix D
(informative)

Provenance Requirements from Nagoya Protocol

Appendix E
(informative, to be deleted)

Requirements on Provenance Model

E.1 Requirements on Common Provenance Model

DR.1 The provenance information model should capture, in a computable (machine-readable and processable) and reproducible way, all the events connected to the physical operations performed on the biological material and all the details of the data generation and data processing workflows, in order to allow the tracking and the backward reconstruction of the history related to sample processing and/or data generation and/or data processing.

DR.2 Provenance model should have "institution" entity, in order to capture institutional responsibility and also to support resolving distributed provenance. The model should, in a computable manner, define responsibility of institution and responsibility of individual persons (and possible delegation of responsibility from institution to a particular person).

DR.3 The provenance information model shall specify clear serialization guidelines as well as implementation guidelines to achieve interoperability of applications producing/consuming the provenance information.

DR.4 The provenance information model shall specify how to find and how to access the provenance information.

DR.5 The provenance information model should define the type (restrictions) on the provenance and its digital representations. Since provenance information can be represented as a graph, example of such a restriction is, e.g., a provenance graph shall be a directed acyclic graph.

DR.6 Provenance model shall support distributed provenance information, where different parts of it are stored and made accessible at a particular institution.

DR.7 Distributed provenance information must support both complete chain model (direct resolution of what are all responsible subjects) as well as predecessor model (only previous step is resolved). The predecessor model is meant for scenarios where the chain model by its nature can make certain sensitive information available (e.g., showing the complete chain how highly pathogenic material is transferred across institutions routinely).

DR.8 Distributed provenance information must support opaque sub-chains of the complete provenance chain that are resolvable *only* at a particular responsible subject.

DR.9 Provenance model must support non-repudiation of origin for all steps of processing.

DR.9.1 non-repudiation of origin of non-sensitive information shall be made directly available as open part of the provenance information;

DR.9.2 non-repudiation of origin of sensitive information must allow storing the sensitive information package only at the responsible subject, while still ensuring the non-repudiation of origin property (using opaque parts of the provenance chain).

DR.10 Provenance information management shall technically support traversal of provenance information from parent to children. This shall be implemented as an optional feature that is only practically enabled for certain scenarios.

E.2 Physical material and its processing

DR.11 The provenance information model shall include generic and extensible support for describing *(a)* acquisition; *(b)* processing; *(c)* storage; *(d)* transport of biological material.

DR.12 The provenance information model shall support the following standards and community standards:

DR.12.1 Support for processes defined in Working Groups (WGs) 2 (possibly also 3 & 4 if relevant) of ISO/TC 276 and ISO/TC 212 WG 4.

DR.12.2 Support for existing methods describing pre-analytical sample processing: *(a)* CEN/TS 16826-1:2015, CEN/TS 16826-2:2015; *(b)* CEN/TS 16827-1:2015, CEN/TS 16827-2:2015, CEN/TS 16827-3:2015; *(c)* CEN/TS 16835-1:2015, CEN/TS 16835-2:2015, CEN/TS 16835-3:2015; *(d)* CEN/TS 16945:2016; *(e)* CEN/TS 16945:2016; *(f)* Standard PREanalytical Code (SPREC) [5]; *(g)* BRISQ [6].

DR.12.3 Support standards coming from current H2020 SPIDIA4P project (which is input into CEN standardization).

DR.13 The provenance information model shall allow the tracking of all the operations involving the biological material being processed, even if they are not directly part of the experimental protocol (e.g., retrieval from a certain lab, transportation, storage, etc., might not be part of experimental protocol).

DR.14 Material processing (e.g., cell culture staining, mechanical stimulation, etc.): the provenance information model shall be able to identify *(a)* the entities (entity ID, primary sample, eventual originating/deriving samples); *(b)* the processing method; *(c)* a link to the reference processing protocol, the performer, the processing parameters, the device(s), the software version, timestamp; *(d)* physical conditions; *(e)* post-analytical conditions; *(f)* execution logs; *(g)* result confidence; *(h)* reference to the donor consent (only if processing human material and if consent is needed in particular legal settings – further denoted as “if applicable”)

DR.15 Material retrieval (e.g., thawing): the provenance information model shall be able to provide information about *(a)* the pre-analytical conditions; *(b)* donor identification; *(c)* entities identification; *(d)* location of the material to be retrieved; *(e)* the physical conditions at retrieval; *(f)* the physical conditions during transportation; *(g)* the performer, shipment details (sender, receiver, carrier), the timestamp; *(h)* reference to the donor consent (if applicable).

DR.16 Material storage: the provenance information model shall be able to locate the biological material, to provide *(a)* details about the physical storage conditions; *(b)* to detail the physical conditions at the arrival; *(c)* to identify the entities; *(d)* to record the physical conditions during transportation; *(e)* to identify the performer; *(f)* to provide the shipment details (sender, receiver, carrier), the timestamp; *(g)* reference to the donor consent (if applicable).

DR.17 The provenance information model shall contain link to the physical label identifying the biological material.

DR.18 Level of detail recorded in the provenance information (e.g., precision of timestamps or precision and frequency of temperature measurements) will depend on intended use of the biological material.

E.3 Data and its processing

DR.19 The provenance information model shall include generic and extensible support for data processing.

DR.20 The provenance information model shall support the following standards and community standards in the field of data processing:

DR.20.1 Support data generation and processing defined within WGs 3 & 4 of ISO/TC 276 and ISO/TC 212 WG 4.

DR.20.2 Support for workflow provenance, using commonly accepted workflow description language(s) such as Common Workflow Language (CWL).[1]

Note: Compatibility with ISO 8000-120:2016 [7] needs to be clarified.

DR.21 Data retrieval (e.g., directly obtained data, processed data, etc.): the provenance information model shall be able to provide details about *(a)* data authorship; *(b)* acquisition information; *(c)* information necessary to verify non-corrupted status of the data; *(d)* retrieval information in a computable manner; *(e)* reference to the donor consent (if applicable).

DR.22 Data generation (e.g., image acquisition, cell culture measurements, etc.): the provenance information model shall be able to computationally describe *(a)* the acquisition protocol; *(b)* the performer; *(c)* the execution parameters; *(d)* the device(s); *(e)* the software version; *(f)* the timestamp of the execution; *(g)* acquisition logs; *(h)* reference to the donor consent (if applicable).

DR.23 Data processing (e.g., quantitative RT-PCR, statistical analysis): the provenance information model should be able to computationally describe *(a)* the analysis protocol; *(b)* the performer; *(c)* the analysis parameters; *(d)* the device(s); *(e)* the software version; *(f)* the timestamp; *(g)* execution log; *(h)* result confidence; *(i)* reference to the donor consent (if applicable).

E.4 Privacy requirements

This section specifies privacy related requirements on provenance information management.

DR.24 The privacy requirements only sets minimum requirements in order to help implementing privacy protection compliance.

DR.25 If applicable, the provenance information model shall provide means to detach personally identifiable information of the research participant contributing biological material and/or data, so that only a trusted party or authorized institution may re-identify the person (for purposes such as incidental findings).

[1] http://www.commonwl.org/

Appendix F
(informative, to be deleted)

Design Considerations

— The proposal is based on the W3C PROV model, which makes it also compatible with HL7 FHIR (relevant for medical research).

— Serialization will be made based on PROV-O (RDF) for linked data applications and W3C PROV-XML (XML) for other applications.

— The proposal defines discovery mechanisms for the provenance information—linking similar to what is in HL7 FHIR and based on W3C PROV-Links.

— The proposal needs to specify requirements on persistent identifiers to be used.

— Appropriate ontologies shall be referred on basis of ISO/AWI 20691 with respect to requirements resulting from use cases in Parts 3–5.

— Opaque provenance components stored at the responsible subjects will be supported so that details do not need to be disclosed in order to access or chain the provenance links together.

— Access to an opaque provenance component will be subject to authorization decision done by a respective responsible subject.

Appendix G
(informative, to be deleted)

Overview of Use Cases (Legacy)

G.1 Scenarios

Three key scenarios involving the preservation of data provenance are discussed below. Essentially, they consist in all or almost all these steps: pre-analytic phase, biological material analysis, data processing.

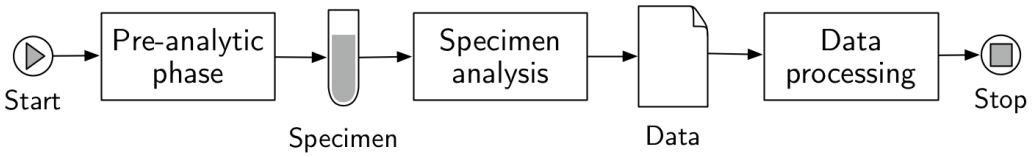


Figure 2: Block diagram of a generic specimen workflow

G.1.1 Clinical laboratory

Clinical laboratories are labs that process specimens for medical investigations (e.g., Clinical Pathology, Microbiology, Biochemistry, Molecular diagnostic, etc.). Typically, the sample lifetime starts somewhere externally to the laboratory in the collection point(e.g., hospital surgery) responding to a specimen request. The sample, possibly along with several others from the same source, arrives to the lab inside a container where it can be accepted or rejected. Each sample is identified, labeled and entered in the Laboratory Information System (LIS) accordingly to the lab's procedures, steps that can vary significantly depending on the level of lab automation. Specimens undergoes a processing phase that implements Standard Operating Procedures (SOPs) producing data output. Depending on the specific clinical analysis, output can be validated by a competent professional and delivered to the requestor as is. In other cases (e.g., genetic tests) raw data have to be further processed to extract clinically relevant information through complex computational steps of analysis.

The main aspects to be taken into account here concerned with provenance are:

— pre-analytical procedures: what happened to the sample (e.g., collection method, temperature history, timing, conservation means, etc.) from when it was collected to when it went to analysis;

— analysis: specific analysis protocols followed and what really happened to the sample (storage, retrieve, analysis);

— data processing: specific workflow applied (version, parameters, etc.) and additional resources involved (sources, version, etc.);

— propagation (link) of the provenance information to the clinical record;

— protect the privacy and confidentiality of the donor in compliance with applicable laws.

G.1.2 General research laboratory

General research laboratories are facilities that processes specimens for scientific research aims. Depending on the specific research activity, research and clinical lab may have a lot of commonalities, differing mainly in the goals. Thus, the same issues arise regarding the provenance concerns except for the privacy concept. Laboratories doing biomedical research on human samples indeed, typically have strict rules about donors' privacy requiring them a full anonymisation of specimen and data as first order of business. Moreover, research laboratory may follow experimental procedure that should be clearly defined and tracked as well.

G.1.3 Bioinformatic processing

Sometimes, the workflow of Figure 2 can start from the data processing block. Such a situation can occur when there is the necessity to reevaluate raw data in response to changes in analysis algorithms, updates in reference resources or, even, in light of new scientific discoveries or the availability of novel technologies. One typical example is that of next-generation sequencing (NGS) centers, since resources involved in genomic analysis are updated quite frequently.

The main issue related to the provenance are those linked to the data processing itself, in particular:

— data source: the origin of the data and how did they come into the processing;

— data processing: specific workflow applied (version, parameters, etc.) and additional resources involved (sources, version, etc.);

— propagation (link) of the provenance information to the clinical record or to further research steps;

G.2 Provenance data generation use cases

We will consider several specific provenance data generation scenarios.

G.2.1 Standardized data processing pipeline

This scenario is typical of high throughput data production facilities as, for instance, NGS centers. The data processing workflow, e.g., the processing from raw data to variant calling, follows a well defined, standardized logical flow. However, it is important to explicitly catch the details of all the steps of the processing pipeline since: (a) the output data could depend in subtle ways from the specific versions of the tools used and parameters chosen; (b) the are frequent changes, e.g., software version, reference databases, in the computational steps details.

The potential impact of these effects is exemplified in Figure 3, which highlights the degree of dissimilarity between sets of genomic variants identified from the exome sequencing of parent-child trios. All sets of variants were extracted from the same data using the same conceptual analysis protocol with different software tools and/or reference genomes.

The pattern shown is typical [8, 9, 10, 11] of data-intensive analysis pipelines and it illustrates the important point that the evolution of analysis algorithms and reference datasets can have a drastic impact on the results. This degree of instability in the results is the consequence of a number of factors, but may be mitigated by ensuring complete auditability and reproducibility of the analyses. It is therefore essential, to be able to effectively compare results, to maintain full tracing on the processing and analysis process that

produced them. Moreover, a functional abstraction of the analysis protocols is required to help cope with the constant evolution.

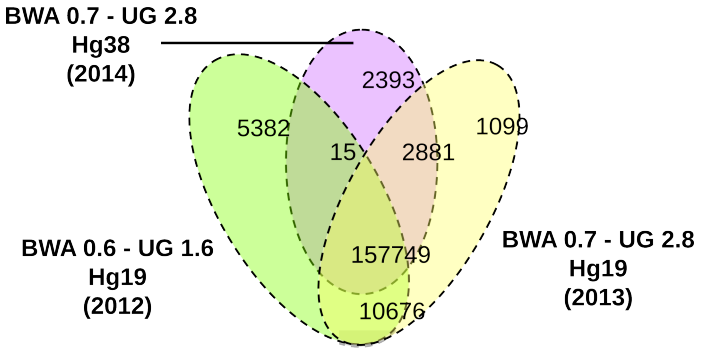


Figure 3: The degree of dissimilarity between the sets of genomic variants identified from exome sequencing of parent-child trios, extracted from the same exome sequencing data by the same abstract analysis protocol, but using different analysis software on the same dataset (P. Uva, unpublished).

There is an obvious dependence of the analysis results on the quality of the samples and how they have been processed and preserved [12, 13]. For instance, in the case of the extraction of dna from formalin-fixed, paraffin-embedded archival tissues, DNA yield and quality depend strongly on the specific extraction technology used [14, 15]. It is therefore important to have associated to the samples quality information (and how it has been measured) but also provenance trails that detail how the sample has been processed. This allows a degree of consistency checking between processing technologies and measured results. Analogously, the condition of acquisition, transport and preservation of the samples could have significant impact on the sample contents. For instance, there are important effects on DNA quality and phylogenetic composition of faecal microbiota derived from the permanence at room temperature of the transport vials of the faecal material [16, 17].

G.2.2 Semi-automatic data processing pipeline

This scenario is typical of situations where the data processing pipeline cannot be completely automated and it requires human intervention during the process. Here, the sequence of analysis steps that needs to be performed cannot be strictly defined *a priori*. This seems to be a common use case in metabolite analysis community, but more details need to be investigated about it.

G.2.3 Quality control trails

In this scenario, a facility—e.g., a biobank—is routinely performing quality controls by reanalyzing (and typically sacrificing) aliquots with a pipeline that involves measurements (e.g., nuclear magnetic resonance (NMR) spectra) and successive data analysis that involve the comparison with equivalent measures done at the moment of acquisition of the sample by the biobank.

G.3 Provenance data utilization use cases

G.3.1 Provenance Data Querying

In the simplest use case, the requester (searcher) searches/filters for samples or data sets given properties stored in the provenance metadata.

When searching for samples, properties of the sample acquisition, processing, or sample storage conditions can be used. Because the sample is also linked to the donating person, it should be possible to search at least based on identifier (code/pseudonym) of the donor (further donor data may or may not be present, as this is more of personal data rather then provenance data). The search based on person identifiers should support coded identities of persons (code/pseudonyms, typical in case of research purposes) and native identifiers (in case that the provenance is used directly as a part of healthcare).

In case of data set generated from samples, e.g., using NGS or any other omics analysis or imaging method, it may be any property of the originating sample as well as method used to generate the data, its configuration parameters, or environmental conditions (acting as external variables of the experiment – also called "context information" in ISO 14721:2012 [18]) during data generation which might have influence on the data generation process.

If different stakeholders keep different parts of the provenance information, these links need to be resolvable and—if the user has sufficient entitlements—the whole distributed provenance graph should be searchable.

G.3.2 "Meaningful" Data Integration

Data integration relies on comparable data to be integrated. It is known that diffent analytical methods may provide slightly different results even if producing seemigly similar results as discussed in Appendix G.2.1. Provenance information for each data set needs to contain description of method used for data generation, its specific configuration used for the particular analysis, and information on relevant enviromental conditions that might have influence on the analysis. Together with links to the provenance information for the source biological material, this should provide complete information to assess whether data coming from two different data sets are suitable for integration or not. It should also provide means to identify sources of gross errors and possibly to explain outliers in the data sets.

For example, RNA analysis of a biological material can be subject to the checks of integrity of RNA molecules and thus the overall conclusion may be that the results are meaningful. But the biological material, which is still "alive", may have responded to the artificial environmental conditions after its removal from its original environment, e.g., after surgical resection of a tumor[2] [19, Annex A]. As a consequence the material analyzed no longer represents the original material and its biological activity in the human body. While the analysis of such material is then performed in a technically correct manner, the results might not be meaningful. Accordingly, a prerequisite to obtain meaningful data, requires assessment of fitness of the biological material, based on the provenance information, for the purpose of the specific research and analytical method.

G.3.3 Open-ended data processing pipeline

In this scenario, we consider a data processing pipeline that uses external resources, typically from the web—such as large bioinformatics databases.

Performing NGS implies, typically, a significant use of external databases as references for genomic sequences, gene functions, effects of variations, computational tools, etc. The representation of this kind of information, however, is a non-trivial task because biological repositories undergo updating process and delivering of new releases quite often, making necessary the track of the particular revision of each instance. Thus, if one wants to mention a particular resource, for instance the gene "SLC35E2B" (HGNC symbol), two more information must also be provided: a link to that particular resource and a version. With reference to

[2] This behavior has been documented, e.g., in EU FP7 SPIDIA project (http://www.spidia.eu) and has become part of Annex A of CEN/TS 16826-1:2015 [19].

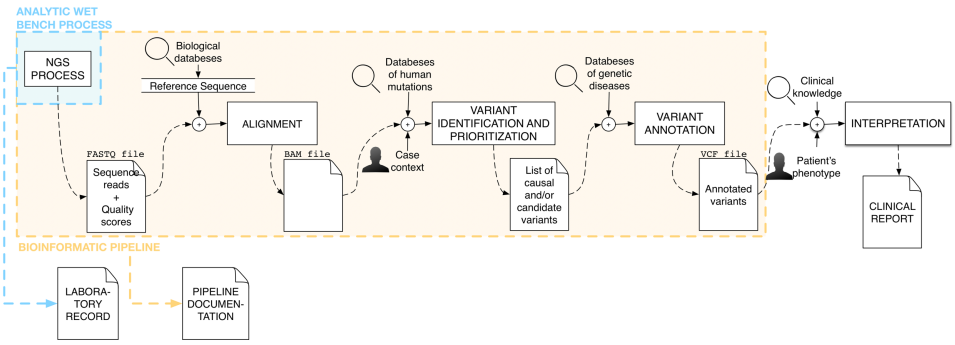


Figure 4: Bioinformatic pipeline - dependency on external resources.

the NCBI Reference Sequence Database (RefSeq), the accession (ID) of the resource is `NG_034044`, the version is `NG_034044.1` and the resource is available at http://www.ncbi.nlm.nih.gov/nuccore/NG_034044.1. Versioning and timestamping of identifiers has been explored also as a part of Persistent Identifiers Interest Group[3] (PID IG) and Data Citation WG[4] of Research Data Alliance (RDA). Versioned or timestamped identifiers are sufficient and necessary for the identification of each particular resource, whatever is inside a sequence database or in a repository of computational workflows.

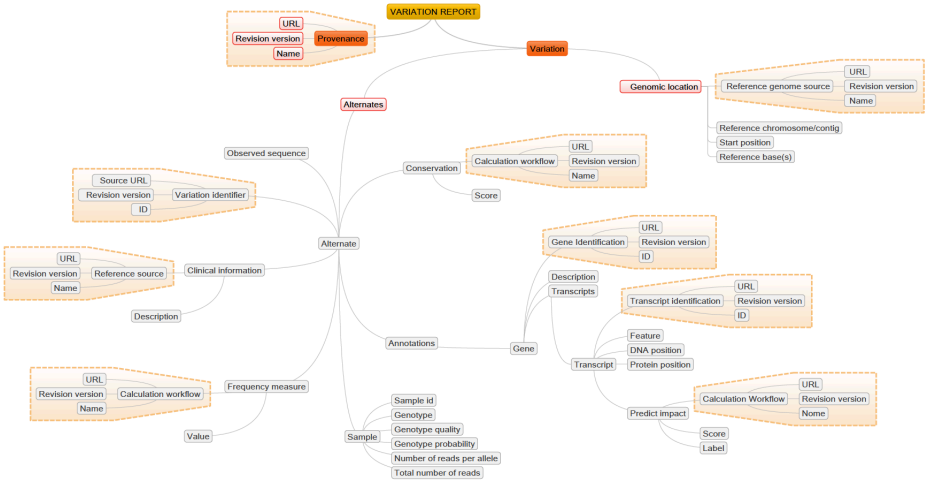


Figure 5: Dependency of a variation report on external resources.

G.4 Use cases specific for human material

G.4.1 Incidental Findings

An important use for when dealing with human biological samples and data in medical research is policies and their implementations for dealing with incidental findings, i.e., findings obtained during the research that might have substantial impact on an individual research participant, her health and life quality. For instance, a mutation is found in the given patient, for which there might be some preventive treatment relevant. Typical policies how to deal with such findings is to report them either to the research participant or her physician, in order to find an optimum way to how deal with the finding and its implications.

[3] https://www.rd-alliance.org/groups/pid-interest-group.html
[4] https://www.rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html

Provenance information in this case serves to trace the data set or samples back to the research participant, also enabling traversal of institutional border in a secure and privacy-respecting manner. Institutional border traversal is necessary when the data sets and biological samples are kept in a distributed way, which is very frequent in the medical research – the full contact information may only be available in the source clinical facitlity. The traversal typically includes: researcher → infrastructure keeping the research data → biobank keeping the biological samples → clinical facility with patient registry being able to identify the patient.

This scenario requires traversing the provenance information in the direction from the data to samples to the contributing research participant (c.f. Appendix G.4.2 for the opposite direction of traversal).

G.4.2 Informed Consent Withdrawal

Another important use case, when dealing with human biological samples and data in research, is when the research participant decides to change or even withdraw her informed consent, resulting in need to remove biological material from that person from the biobank repositories, and/or to remove person's data from the containing data sets where feasible.

This use case requires finding all the relevant data sets generated using the research participant's samples or data and removing them from repositories and upstream data sets where feasible.[5] This requires traversing the provenance information in the direction from the contributing research participant to samples or data (c.f. Appendix G.4.1 for the opposite direction of traversal). This is less common traversal direction, which requires additional measure to be taken: each derived sample or data set needs to be registered at its parent.

G.5 Use case specific for non-human material

G.5.1 Ocean microbial reference gene catalog

The *Tara Oceans Expedition* is a French foundation focused on the study of the marine ecosystems [20, 21]. Between 2009 and 2012 the schooner Tara, equipped with advanced sequencing and imaging technologies, collected over 30,000 samples of seawater and plankton at three depths from 210 station across all the world's oceans. Tara Oceans adopts the principle of open access, thus sequencing and oceanographic datasets were made available to the scientific community on the Internet. Raw short sequence reads and derived data (e.g., assemblies, annotations, etc.) are published in the European Nucleotide Archive,[6] while environmental data are archived in the PANGAEA, Data Publisher for Earth and Environmental Science.[7]

Here, the main aspects related to provenance are:

— geolocation of the sample extraction (e.g., specific position of the station, distance from coast, depth, etc.),

— environmental features during the sampling event (e.g., salinity, water temperature, etc.),

[5] This functionality may not be available for anonymized data sets according to the definition of anonymization in the upcoming European General Data Protection Regulation (GDPR), where links between the contributing research participants and the resulting data sets are permanently destroyed. Furthermore, even the GDPR assumes that when the removal is impossible, e.g., the data set has been already published publicly as a part of research results – see GDPR Article 7 §3 stating "withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal".

[6] http://www.ebi.ac.uk/ena/

[7] http://www.pangaea.de

— sample extraction (e.g., sampling date/time, device, methods, etc.),

— data generation and processing.

G.5.2 Genetic determinisms of plant responses to environmental conditions

The *PHENOME, French Plant Phenomic Network*,[8] is a web of seven facilities equipped with high throughput technologies aimed at the phenotypic and genotypic characterization of different species of plants. Two platforms works under controlled conditions: large collections of plants are subjected to environmental changes in soil water, temperature and CO_2 and the plant response (e.g., growth rates, biomass accumulation, transpiration, infra-red temperature, etc.) is recorded at different time intervals. Five of the facilities have fields platform equipped with soil and climate sensors that are able to control rainfall and temperature. Finally, omic analysis are centralized in two of the platform able to perform high throughput metabolic and structural characterization.

Peculiar aspects related to the preservation of provenance here come from the necessity to coordinate and manage large dataset of phenotypic data originating from different nodes. Here, each experiment must be thoroughly characterized in terms of:

— experimental setting (e.g., facility, platform, chamber, pot, plant number, etc.),

— environmental conditions (e.g., light, air temperature, soil water status, etc.),

— data acquisition (e.g., imaging system, metabolomic measurements, climatic sensor network, etc.),

— data processing (e.g., imaging segmentation, growth rate prediction, genomic analysis algorithm, etc.)

[8] https://www.phenome-fppn.fr/phenome_eng/

Bibliography

[1] *ISO 13888:2009 – Information technology – Security techniques – Non-repudiation*. 2018.

[2] *ISO 20387:2018 – Biobanking – General requirements for biobanking*. 2018.

[3] MOREAU, L. et al. *PROV-DM: The PROV Data Model*. 2013. Available also from: https://www.w3.org/TR/prov-dm/.

[4] OCCHETTA, P.; ISU, G.; LEMME, M.; CONFICCONI, C.; OERTLE, P.; RÄZ, C.; VISONE, R.; CERINO, G.; PLODINEC, M.; RASPONI, M.; MARSANO, A. A three-dimensional: In vitro dynamic micro-tissue model of cardiac scar formation. *Integrative Biology (United Kingdom)*. 2018, vol. 10, no. 3, pp. 174–183. ISSN 17579708. Available from DOI: 10.1039/c7ib00199a.

[5] BETSOU, F.; LEHMANN, S.; ASHTON, G.; BARNES, M.; BENSON, E. E.; COPPOLA, D.; DESOUZA, Y.; ELIASON, J.; GLAZER, B.; GUADAGNI, F.; HARDING, K.; HORSFALL, D. J.; KLEEBERGER, C.; NANNI, U.; ANIL, P.; SHEA, K.; SKUBITZ, A.; SOMIARI, S.; GUNTER, E.; INTERNATIONAL SOCIETY FOR BIOLOGICAL AND ENVIRONMENTAL REPOSITORIES (ISBER) WORKING GROUP ON BIOSPECIMEN SCIENCE. CEBP Focus : Biomarkers and Biospecimens Hypothesis / Commentary Standard Preanalytical Coding for Biospecimens : Defining the Sample PREanalytical Code. 2010, vol. 19, no. April, pp. 1004–1012. Available from DOI: 10.1158/1055-9965.EPI-09-1268.

[6] MOORE, H. M.; KELLY, A. B.; JEWELL, S. D.; MCSHANE, L. M.; CLARK, D. P.; GREENSPAN, R.; HAYES, D. F.; HAINAUT, P.; KIM, P.; MANSFIELD, E.; POTAPOVA, O.; RIEGMAN, P.; RUBINSTEIN, Y.; SEIJO, E.; SOMIARI, S.; WATSON, P.; WEIER, H.-U.; ZHU, C.; VAUGHT, J. Biospecimen reporting for improved study quality (BRISQ). *Journal of proteome research*. 2011, vol. 10, no. 8, pp. 3429–38. ISSN 1535-3907. Available from DOI: 10.1021/pr200021n.

[7] *ISO 8000-120:2016 – Data quality – Part 120: Master data: Exchange of characteristic data: Provenance*. 2016.

[8] HWANG, S.; KIM, E.; LEE, I.; MARCOTTE, E. M.; MCVEAN, G. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*. 2015, vol. 5, pp. 17875. ISSN 2045-2322. Available from DOI: 10.1038/srep17875.

[9] O'RAWE, J.; JIANG, T.; SUN, G.; WU, Y.; WANG, W.; HU, J.; BODILY, P.; TIAN, L.; HAKONARSON, H.; JOHNSON, W. E.; WEI, Z.; WANG, K.; LYON, G. J. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine*. 2013, vol. 5, no. 3, pp. 28. Available from DOI: 10.1186/gm432.

[10] KOFLER, R.; LANGMULLER, A. M.; NOUHAUD, P.; OTTE, K. A.; SCHLOTTERER, C. Suitability of Different Mapping Algorithms for Genome-wide Polymorphism Scans with Pool-Seq Data. *G3: Genes|Genomes|Genetics*. 2016. ISSN 2160-1836. Available from DOI: 10.1534/g3.116.034488.

[11] LAURIE, S.; FERNANDEZ-CALLEJO, M.; MARCO-SOLA, S.; TROTTA, J.-R.; CAMPS, J.; CHACÓN, A.; ESPINOSA, A.; GUT, M.; GUT, I.; HEATH, S.; BELTRAN, S. From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. *Human Mutation*. 2016, vol. 37, no. 12, pp. 1263–1271. ISSN 10597794. Available from DOI: 10.1002/humu.23114.

[12] SPENCER, D. H.; SEHN, J. K.; ABEL, H. J.; WATSON, M. A.; PFEIFER, J. D.; DUNCAVAGE, E. J.; MACKAY, A.; ASHWORTH, A.; PRITCHARD-JONES, K.; JONES, C.; KIBRIYA, M.; FENNELL, T.; KERNYTSKY, A.; SIVACHENKO, A.; CIBULSKIS, K.; GABRIEL, S.; ALTSHULER, D.; DALY, M.; SMITH, J.; MORGAN, G.; KNEBA, M.; MACINTYRE, E.; WESTERVELT, P.; DIPERSIO, J.; LINK, D.; MARDIS, E.; LEY, T.; WILSON, R.; GRAUBERT, T. Comparison of Clinical Targeted Next-Generation Sequence Data from Formalin-Fixed and Fresh-Frozen Tissue Specimens. *The Journal of Molecular Diagnostics*. 2013, vol. 15, no. 5, pp. 623–633. ISSN 15251578. Available from DOI: 10.1016/j.jmoldx.2013.05.004.

[13] KOFLER, R.; NOLTE, V.; SCHLÖTTERER, C. The impact of library preparation protocols on the consistency of allele frequency estimates in Pool-Seq data. *Molecular Ecology Resources*. 2016, vol. 16, no. 1, pp. 118–122. ISSN 1755098X. Available from DOI: 10.1111/1755-0998.12432.

[14] SENGÜVEN, B.; BARIS, E.; OYGUR, T.; BERKTAS, M. Comparison of methods for the extraction of DNA from formalin-fixed, paraffin-embedded archival tissues. *International journal of medical sciences.* 2014, vol. 11, no. 5, pp. 494–9. ISSN 1449-1907. Available from DOI: 10.7150/ijms.8842.

[15] HEDEGAARD, J.; THORSEN, K.; LUND, M. K.; HEIN, A.-M. K.; HAMILTON-DUTOIT, S. J.; VANG, S.; NORDENTOFT, I.; BIRKENKAMP-DEMTRÖDER, K.; KRUHØFFER, M.; HAGER, H.; KNUDSEN, B.; ANDERSEN, C. L.; SØRENSEN, K. D.; PEDERSEN, J. S.; ØRNTOFT, T. F.; DYRSKJØT, L. Next-Generation Sequencing of RNA and DNA Isolated from Paired Fresh-Frozen and Formalin-Fixed Paraffin-Embedded Samples of Human Cancer and Normal Tissue. *PLoS ONE*. 2014, vol. 9, no. 5, pp. e98187. ISSN 1932-6203. Available from DOI: 10.1371/journal.pone.0098187.

[16] HILL, C. J.; BROWN, J. R. M.; LYNCH, D. B.; JEFFERY, I. B.; RYAN, C. A.; ROSS, R. P.; STANTON, C.; O'TOOLE, P. W. Effect of room temperature transport vials on DNA quality and phylogenetic composition of faecal microbiota of elderly adults and infants. *Microbiome*. 2016, vol. 4, no. 1, pp. 19. ISSN 2049-2618. Available from DOI: 10.1186/s40168-016-0164-3.

[17] KNUDSEN, B. E.; BERGMARK, L.; MUNK, P.; LUKJANCENKO, O.; PRIEMÉ, A.; AARESTRUP, F. M.; PAMP, S. J. Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. *mSystems*. 2016, vol. 1, no. 5. Available also from: http://msystems.asm.org/content/1/5/e00095-16.

[18] *ISO 14721:2012 – Space data and information transfer systems – Open archival information system (OAIS) – Reference model*. 2012.

[19] *CEN/TS 16826-1:2015 – Molecular In Vitro Diagnostic Examinations - Specifications For Pre-Examination Processes For Snap Frozen Tissue – Part 1: Isolated RNA*. 2015.

[20] PESANT, S.; NOT, F.; PICHERAL, M.; KANDELS-LEWIS, S.; LE BESCOT, N.; GORSKY, G.; IUDICONE, D.; KARSENTI, E.; SPEICH, S.; TROUBLÉ, R.; DIMIER, C.; SEARSON, S.; ACINAS, S. G.; BORK, P.; BOSS, E.; BOWLER, C.; DE VARGAS, C.; FOLLOWS, M.; GORSKY, G.; GRIMSLEY, N.; HINGAMP, P.; IUDICONE, D.; JAILLON, O.; KANDELS-LEWIS, S.; KARP-BOSS, L.; KARSENTI, E.; KRZIC, U.; NOT, F.; OGATA, H.; PESANT, S.; RAES, J.; REYNAUD, E. G.; SARDET, C.; SIERACKI, M.; SPEICH, S.; STEMMANN, L.; SULLIVAN, M. B.; SUNAGAWA, S.; VELAYOUDON, D.; WEISSENBACH, J.; WINCKER, P. Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*. 2015, vol. 2, pp. 150023. ISSN 2052-4463. Available from DOI: 10.1038/sdata.2015.23.

[21] ZHANG, H.; NING, K. The Tara Oceans Project: New Opportunities and Greater Challenges Ahead. *Genomics, Proteomics & Bioinformatics*. 2015, vol. 13, no. 5, pp. 275–277. ISSN 16720229. Available from DOI: 10.1016/j.gpb.2015.08.003.

