EOSC-Life

EOSC-Life: Building a digital space for the life sciences

D6.2 – Common Provenance Model

WP6 – FAIRification and Provenance Services
Lead Beneficiary: BBMRI-ERIC
WP leader: Petr Holub and Isabelle Perseil
Contributing partner(s): BBMRI-ERIC, University of Würzburg (3rd party to BBMRI-ERIC), Medical University Graz (LTP to BBMRI-ERIC), CRS4 (LTP to BBMRI-ERIC), VLIZ

Authors of this deliverable: **Rudolf Wittner, Cecilia Mascia, Francesca Frexia, Heimo Müller, Jörg Geiger, Katrina Exter, and Petr Holub**

Contractual delivery date: **28 February 2021**
Actual delivery date: **20 April 2021**
H2020-INFRAEOSC-2018-2

Grant agreement no. 824087
Horizon 2020
Type of action: RIA

Table of Contents

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Executive Summary

The exchange of research data and physical specimens has become an issue of major importance for modern research. Many reports indicate problems with quality, trustworthiness and reproducibility of research results, mainly due to poor documentation of the data generation or the collection of specimens. The significant impact of flawed research results on health, economics and political decisions has frequently been stated. Consequently, professional societies and research initiatives call for improved and standardised documentation of the data and specimens used in research studies.

Provenance information documents the evolution of an object and can be used to assess its quality and reliability. This deliverable defines components of distributed provenance information to enable interlinking of provenance information generated in different organisations involved in the research process, such as biobanks, research centres, universities or analytical laboratories. The distributed provenance information model builds on an existing provenance information standard, W3C PROV, and follows a general provenance composition pattern. Both W3C PROV and provenance composition pattern is described in this document.

Since understanding of the term “provenance information” differs across different domains and research communities, this deliverable firstly harmonises this understanding by providing a general explanation of how provenance information is generated and used.

In particular, this deliverable defines a *connector*, that is a provenance component containing technical information to traverse through provenance information. The connector is subsequently added to provenance information generated by different organisations. This deliverable also defines how to interpret identifiers of provenance structures in a distributed environment and how to include and interpret persistent identifiers of documented objects.

This deliverable deals with the common provenance model developed as a part of a standardisation process in the International Organisation for Standardisation (ISO) technical committee “Biotechnology” ISO/TC 276, and which is registered as project ISO 23494 in the working group 5 “Data processing and Integration”. Because this work is copyrighted by ISO and cannot be published as a public deliverable, this text describes the essential design of the provenance model, and the actual ISO document is provided as a non-public supplement. This is in line with the work plan of EOSC-Life WP6 in order to support adoption of the standard both in academia and in industry. The Common Provenance Model has been accepted as a Preliminary Work Item under 23494 Part 2 and it is being proposed for moving it into the next phase, the New Work Item at the time of submitting the deliverable.

Project Objectives

With this deliverable, the project has reached/contributed to the following objectives:

a. Establish EOSC-Life by publishing FAIR life science data resources for cloud use

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

    - The deliverable describes the interoperable common provenance model undergoing standardisation as ISO 23494 Part 2.
  b. Create an ecosystem of innovative life-science tools in EOSC
    - ditto
- Call objective: Proposals will address the stewardship of data handled by the involved research infrastructures according to the FAIR principles.
- Call objective: This will include the definition of domain specific data policies (e.g. acquisition, deposit, curation, preservation, access, sharing and re-use) and address any legislative or interoperability issues which affect data handling across geographical and discipline borders.

Detailed Report on the Deliverable

1. Introduction

The exchange of research data and physical specimens has become an issue of major importance for modern research. Many reports indicate problems with quality, trustworthiness and reproducibility of research results, mainly due to poor documentation of the data generation or the collection of specimens [1–6]. The significant impact of flawed research results on health, economics and political decisions has frequently been stated [7–10]. Consequently, professional societies and research initiatives call for improved and standardised documentation of the data and specimens used in research studies [11–16]. Such documentation – also called provenance information – should not form a standalone description of research data, but rather be interlinked with documentation coming from other sources; to create an uninterrupted chain of description of the whole research process, starting from physical specimen acquisition or initial measurements and ending with data integration and processing at the end of the research process. Such distributed provenance information is created when a described object or its derivatives traverse through different organisations, and each organisation documents a part of its life-cycle. That this is done properly is important because of the dependency of the quality of research data on the reliability and quality of all the inputs from which the data was generated, in order to avoid “garbage in, garbage out” situation.

Provenance information documents the evolution of an object and can be used to assess its quality and reliability. In context of the FAIR initiative [17], it primarily supports reproducibility of research by enabling an effective evaluation of performed activities. To enable effective and meaningful use of provenance information, provenance information must be FAIR. Findability and accessibility of provenance information can be partially achieved by attaching provenance related information to particular data. On the other hand, findability of a complete provenance chain of a data back to its source (possibly to a particular human, plant or location) together with its interoperability is a current major challenge. In addition, the significant portion of research data comes from biological material, and including the provenance of these physical specimens is often neglected when implementing FAIR as data-only principles [18].

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

A barrier for creating a fully technically-interoperable, distributed provenance for research, is that different communities have a different understanding of what provenance information is specifically, ranging from various provenance models to dumped collections of log files. In addition, despite the fact that there is a widely-accepted general provenance information model in several domains, its adoption often involves inconsistent technical solutions providing a different granularity of information.

The main benefit of the proposed common provenance model is that it is domain and technology agnostic, by which it aims to be adopted by a wide diversity of research areas and research organisations. The model supports the creation of provenance information for both digital and physical objects, such as research data or biological samples respectively. The main contributions of this work are: a clear definition of what is meant by the term "provenance information", to harmonise its understanding within different research communities and domains; the definition of a "connector" as an extension of a W3C PROV [19] provenance model: that is, a piece of provenance information that enables the interconnection of provenance information generated within various environments to be made; the proposal of practical aspects related to distributed provenance, such as using shared identifiers in provenance information; and finally, the integration and interpretation of persistent identifiers within provenance.

2. Preliminaries

This section describes an existing groundwork that the deliverable is built on. In particular, it describes fundamentals of W3C PROV, which is a current standard for provenance information representation. The provenance composition pattern is built on top of the W3C PROV standard and defines an abstract way to connect provenance information coming from different sources. In the end, this deliverable addresses the gaps to create fully-interoperable interconnections of distributed provenance information.

2.1. W3C PROV

W3C PROV [19] is a family of specifications defining standard for provenance information. W3C PROV introduces a PROV-DM data model [20] and its respective serialisations. The core of the model lies in provenance structures – an entity, an activity and an agent.

1. An entity can be perceived as a snapshot of an object (which might be both physical or digital) with fixed attributes, thus the same object can be expressed in provenance information by multiple entities expressing different snapshots of the same object.
2. Activities express processes that act upon objects represented by entities.
3. Agents can be used to express responsibility for entities and activities. These structures can be interrelated using a predefined set of relations, such as *wasDerivedFrom*, *wasInformedBy*, or *wasGeneratedBy*.

The main parts of the model are depicted in Figure 1.

In addition, all PROV provenance structures and relations can be specialised using pre-defined extensibility points to express more precise semantics to be used when applying the model in

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

particular domain. Since the PROV data model puts minimal requirements on PROV instances, W3C PROV also introduces the PROV-CONSTRAINTS recommendation [21] to define the validity of provenance information. Validity of provenance is defined by application of various rules to create consistent descriptions of an object's history, and to enable meaningful logical reasoning over provenance.

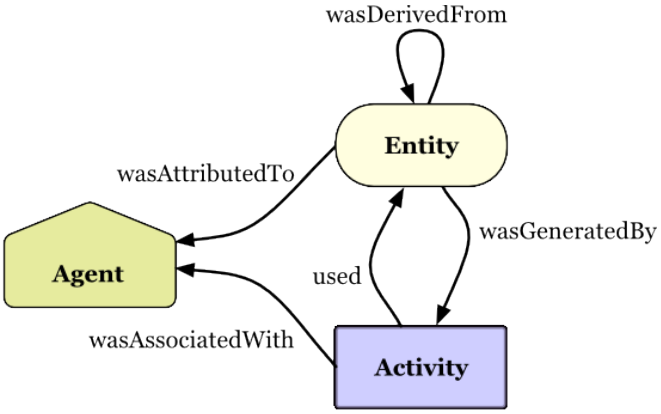

Figure 1: Conceptual structure of the W3C PROV-DM (cited from [20]).

Provenance information is structured in PROV in the following way: all provenance structures and relations are expressed in provenance information by provenance statements; provenance statements are then encapsulated in provenance documents, which consist of *bundles*; a *bundle* is defined as a named set of a provenance statements and can be represented in provenance information as an *entity* to its provenance.

2.2. Provenance Composition Pattern

This deliverable builds primarily on a conceptual proposal of a provenance composition pattern[1] [22]. The pattern provides the foundation for distributed provenance in PROV, and aims at documenting independent communicating processes. In this section, the relevant parts of the model and its shortage for direct practical applications are described.

In the provenance composition pattern, communication of two processes involving a sender and a receiver is expressed using a common *entity* called *message*. Both sender and receiver generate individual pieces of provenance information for their processes and encapsulate each in their own *bundle*. In the sender's provenance, the outcome of the sender's process is expressed as an *entity* (or multiple entities) and this outcome is to be sent to the receiver. The sender's outcome is then used as an input to the receiver's process, so the receiver includes *entities* that represent the inputs in provenance information. In particular, the *message* represents a pairing between what has been sent (sender's outputs) and what has been received (receiver's inputs), it is included in both the sender's and the receiver's provenance information, and it is related to respective *entities* using the *wasDerivedFrom* relation. The *message* contains additional attributes to enable navigation between particular *bundles*, and it can be also attributed to an *agent* to express

[1] "Provenance patterns" or "provenance recipes" are structures and mechanisms used to describe particular scenarios in provenance. Conceptually, they are similar to design patterns used in common programming languages.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

responsibilities for *message* creation and receipt. The pattern also suggests using *bundle* and *message* identifiers to locate particular pieces of provenance information.

This pattern can be used to interconnect distributed provenance information generated by different organisations. The pattern does not address practical aspects of distributed provenance, such as conventions for sharing identifiers or normative definition of attributes containing information to navigate through provenance bundles. The pattern also suggests updating the sender's provenance information to enable forward navigation by adding additional information. This will not always be feasible, e.g. in cases when the sender's provenance information is digitally signed already and updating the content of the original provenance would disturb the signature. In order to implement forward navigation while implementing signature-based trust mechanisms, a more sophisticated method is required to add additional information into a *bundle* that is part of distributed provenance.

3. Elements of distributed provenance information

Here we describe a refinement of the provenance composition pattern. In particular, we provide an updated definition of the *message*; describe how identifiers of *connectors* and other structures are generated, used, and interpreted; how to link provenance structures to described objects; and how to update content of an existing bundle, which can be referred to from provenance information held by different organisation.

3.1. Provenance information finalisation

The main goal of provenance information is not to include or replace existing documentation or logging infrastructure, but rather to provide an additional level of documentation, which is fully technically interoperable across different organisations and potentially fulfils additional requirements (e.g. recommendations listed in [23]).

This can be achieved by assembling the required information in the usual formats and generating provenance information when necessary: on request; at the end of specified process; or periodically after a predefined time interval. Specific examples are: generate provenance information on request during dataset request, at the end of a DNA sequencing process, or during monitoring of a physical environment of a biological material storage (e.g. temperature, humidity, etc.). This is referred to as a *provenance finalisation event*, to explicitly distinguish arbitrary documentation and log files from provenance information.

In particular, every piece of provenance information generated by an organisation during the f*inalisat*ion event is a valid W3C PROV instance and is encapsulated in a PROV *bundle* containing all relevant information in terms of *entities*, *activities* and *agents*. After the *finalisation*, particular information is considered to be a fixed snapshot of the current knowledge and is not further directly updated. This allows one to generate valid and complete provenance information, encapsulated in a *bundle* documenting a particular process. Such provenance information can then be digitally signed and archived to be used for auditability, accounting, or other purposes.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

3.2. A connector

In order to create an uninterrupted documentation of an object, particular provenance information coming from different organisations must be interlinked. For that purpose, a *connector* is introduced. *Connector* is a PROV *entity*, that represents an object exchanged between two organisations: a sender and a receiver. Both sender and receiver generate provenance information and encapsulates this in a *bundle*, so the *connector* technically interconnects respective *bundles*. The main focus of this definition is to express continuity of distributed provenance information (in contrast, the *message* defined in the provenance composition pattern represents a mapping between entities).

Any object, e.g. dataset or a biological specimen, that is sent from a sender, is represented in the sender's provenance information as the *connector*. An object that is received by the receiver is expressed in the receiver's provenance information as an *entity* that is derived from the *connector*. The *connector* is invalidated in the receiver's *bundle* to express that the *entity* that represented the object, that was sent from the sender, is not available anymore after the reception. Invalidation of the *connector* will also force consecutive *activities* to work with the *entity* that represents the received object. In addition, responsibility and additional information related to a particular organisation can be expressed by linking the *connector* with an *agent* using an *attribution* relation. This is depicted in Figure 2. Sending and receiving processes can be documented using *qualified derivation pattern* [24] between the *connector* and particular *entity* in both the sender's and receiver's provenance information.

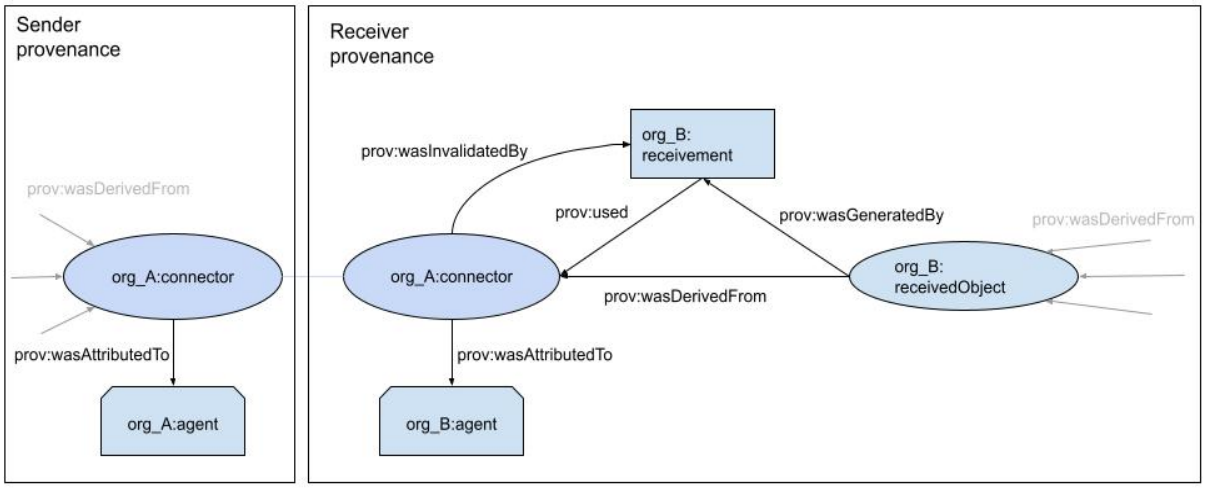

Figure 2: Using a connector to interconnect different bundles

If two provenance statements have the same identifier in PROV, they refer to a single provenance structure. With regard to this rule, the *connector* entity must have the same identifier in both the sender's and the receiver's provenance information. To make this work in the real world, the *connector* identifier should be globally unique to prevent potential collisions with identifiers from other provenance structures. This can be achieved using a format of identifiers in PROV in the following way:

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

1. The identifier in PROV is a qualified name, consisting of a prefix and a local name.
2. The prefix is associated with a URI. We suggest that all URIs that are used as a prefix for provenance structures and identifiers are assigned uniquely to an organisation, meaning that no organisations can share an identifier.
3. Then the global uniqueness of a provenance structure identifier can be achieved by combining the particular URI prefix with the local part of the particular identifier, which is unique locally within one organisation.

A URI can be generated by an authority, such as IANA[2], which will achieve its global uniqueness. This will prevent potential collisions between provenance information generated by different organisations. Assigning a URI to a qualified name prefix in the PROV document is depicted in Listing 1.

If applied on the *connector*, then the identifier of the *connector entity* is generated by the sender and contains the sender's URI as a prefix. As a result, the *connector* entity will have the sender's prefix also in the receiver's provenance information, which will indicate that this entity comes from a different organisation. Listing 2 depicts the *connector* entity in the receiver's *bundle*.

Assigning a unique identifier to a *connector* is only one part of what is needed to navigate through distributed provenance information. In addition to supporting fully machine-actionable provenance information navigation, information about where to find provenance of an entity with a given identifier is needed. The W3C PROV-AQ document [25] defines the terms *provenance URI* and *service URI*, which refer to a *URI denoting some provenance record*, and a *service that accesses provenance given a query containing a target-URI or other information that identifies the desired provenance*. The proposed model uses a pair of attributes to store these values in the *connector* in both the sender and receiver's provenance information to navigate to the other party (i.e., sender to the receiver and receiver to the sender respectively). At least one of the values is mandatory for the receiver. Since the sender may not know who is the receiver of the exchanged object at the moment of provenance information generation, inclusion of those values is optional.

In addition to the identification of provenance structures, it is also important to preserve a link between an object and the entity that represents it in provenance information. To do so, a new attribute *primaryId* is introduced in the proposed model. The value of the *primaryId* attribute is an identifier of the represented object. If there is a requirement to keep the identifier persistent, then it has a format of a qualified name and is interpreted as described in the following section. Usage of the *serviceUri* and *primaryId* attribute is also depicted in Listing 2. Having a specific attribute for external identifiers is a better option than using external identifiers for provenance structure identifiers. This will limit the possibility of the constraints of different identifier management policies within organisations (which are, e.g. in the biomedical or biotechnology research domain, strictly regulated by law), which could cause problems for creating fully interoperable provenance information.

Listing 1: Assigning a URI to a provenance structure identifier prefix

```
bundle sender:bundle1
prefix sender <http://sender.org/>
```

[2] https://www.iana.org/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

```
prefix cpm <http://commonprovenancemodel.org/>
#the connector in the sender's bundle
entity(sender:entity1, [prov:type='cpm:connector',
cpm:primaryId='dataset042'])
endBundle
```

Listing 2: The connector entity in the receiver's bundle

```
bundle receiver:bundle1
prefix sender <http://sender.org/>
prefix receiver <http://receiver.org/>
prefix cpm <http://commonprovenancemodel.org/>

#the connector in the receiver's bundle
entity(sender:entity1, [prov:type='cpm:connector',
cpm:primaryId='dataset042', cpm:senderBundleId='sender:bundle1',
serviceUri='http://sender.org/provService'])
endBundle
```

4. Resolving persistent identifiers

Provenance information contains descriptions of an object that can change over time, although updating provenance information to reflect the current state of described objects may not be necessary or feasible (e.g. because the provenance information handler would have to be notified about the change, which might be complicated in cases when someone else owns the object). Examples of such values are *provenanceURI* and *serviceURI* described in the Section "A connector", for which it must be ensured that their validity does not expire. In this section, a mechanism for dealing with changes of objects described in provenance information, without updating that provenance information, is defined. The proposed mechanism exploits use of a globally unique and resolvable persistent identifier (PIDs) [26].

A PID is an identifier that is globally unique and resolvable in long term. Resolvability of an identifier refers to a property of being able to access the identified object, its digital representation, or related information using the PID. A PID is generated and assigned by a third party (e.g. ePIC[3] or registrants appointed by International DOI Foundation for DOI identifiers), which provides assurance of the required PID properties. PIDs can be expressed in provenance information using qualified names (containing a prefix and local name, see Section "A connector"). Interpretation of a PID follows interpretation of any qualified name (e.g. provenance structures identifier) in PROV. The prefix of a qualified name is associated with a particular resolver in namespace declarations (as depicted in the Listing 3). The local part of the qualified name is a persistent identifier itself, which can be resolved using the resolver indicated by the prefix of the qualified name. This can then be used for any attribute value in provenance information that should be persistent.

Listing 3: Using a persistent identifier

```
bundle receiver:bundle1
prefix sender <http://sender.org/>
```

[3] https://www.pidconsortium.net/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

```
prefix receiver <http://receiver.org/>
prefix cpm <http://commonprovenancemodel.org/>
prefix resolver <http://pidresolver.org>

#the connector in the receivers's bundle
entity(sender:entity1, [prov:type='cpm:connector',
cpm:primaryId='dataset042', cpm:senderBundleId='sender:bundle1',
cpm:senderProvenanceUriPid='resolver:sf7s743sf8'])
endBundle
```

If a value represented by a particular PID changes, the new value is assigned to the PID without modifying provenance information at particular resolver. The ability to track the history of values for a particular PID is delegated to the authority, which is implied through the ”persistence” property. The distributed provenance model uses persistent identifiers for attribute values that must always be up to date, but can be in the custody of different organisations, not directly in the custody of an organisation generating a particular piece of provenance (e.g. laboratory generating data is responsible for the provenance generation, but the PIDs are provided by the university at which the laboratory resides, or by a research infrastructure with which the laboratory is affiliated). If the value of the PID changes, the responsible organisation notifies the PID resolver, and the new value is then ”propagated” to every provenance information containing that particular PID.

If we speak about integrity, authenticity, or non-repudiation of provenance information, it is important to point out that this mechanism does not affect it. The security properties are strictly bound to provenance information and it is out of their scope to cover what is outside provenance information. This is also important because the reliability and trustworthiness of provenance information, which is supported by integrity, authenticity, and non-repudiation, still lies on its truthfulness and correctness.

5. Bundles versioning

A mechanism for the versioning of *bundles* is proposed in this section. This can be used to update part of the distributed provenance information. The main feature of this mechanism is that it allows for the updating provenance information and does not break integrity-related properties of the provenance information, thus enabling updates and navigation through provenance information. First, every organisation involved in the distributed processing chain maintains a *meta-bundle*. The *meta-bundle* contains the so-called *provenance* of *provenance*, which is the provenance of existing bundles managed by a particular organisation. If provenance information is generated and encapsulated in a *bundle*, then the *bundle* is also represented in the *meta-bundle* as two *entities* - one representing general aspects of the *bundle* (with no details related to a specific version) and the second representing the specific version of the *bundle* (this is an application of a revision pattern described in Moreau and Groth, 2013 [27] and also loosely follows the semantics defined in the PAV ontology [28]). This is depicted in Figure 3.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

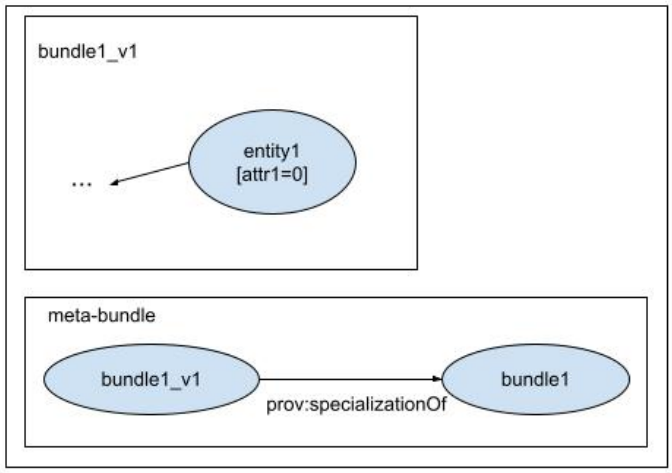

Figure 3: Content of the meta-bundle before an update of provenance information

To update original provenance information encapsulated in a bundle1_v1, a new copy of the bundle is created in a particular organisation, referred to as bundle1_v2. The bundle1_v2 is perceived as a replacement of the outdated bundle1_v1 and includes any modifications of the original bundle - additional, removed, or updated provenance information. The original bundle1_v1 is not deleted. In the *meta-bundle*, a new entity representing bundle1_v2 is created and related to the original *entity* representing bundle1_v1 and general aspects of the bundle using the *wasRevisionOf* and *specialisationOf* relations respectively. This is depicted in Figure 4.

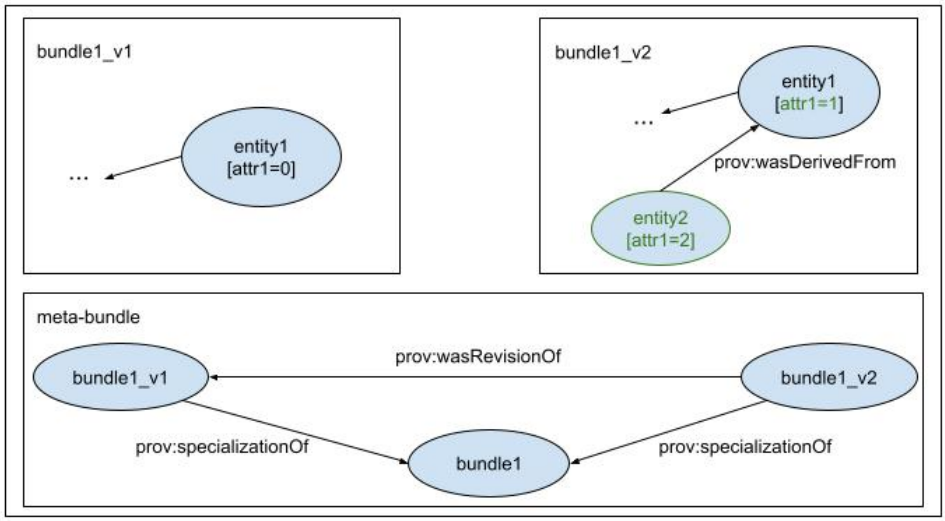

Figure 4: Content of the meta-bundle after an update of provenance information

If provenance information is updated in the way as described here, then the original provenance information is not modified, and hashes or digital signatures would not be invalidated. Since the original information is preserved, it is still potentially available during navigation through the distributed provenance. If updated provenance information is found during the navigation, an algorithm executing the navigation could be notified that an updated version exists and it can

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

potentially be provided to a user. This mechanism also supports auditability and accountability of past modifications of provenance information.

6. State of Standardisation

We aim for standardisation of the Common Provenance Model via International Organisation for Standardisation (ISO) technical committee “Biotechnology” ISO/TC 276, and which is registered as project ISO 23494 in the working group 5 “Data processing and Integration”. The reason for this is to stimulate industrial adoption of the standard as many devices such as laboratory automation and data generating equipment (from sequencers and spectroscopes to microscopes) are produced on a commercial basis and used by the research facilities. This is in line with the original work plan of EOSC-Life WP6. The domain-specific extensions of the standard also need to link to the standards for handling the biological material and data generation methods, which are standardised in the working groups 2 and 3 of the TC/276 respectively.

Because this work is copyrighted by ISO and cannot be published as a public deliverable, this text describes the essential design of the provenance model, and the actual ISO document is provided as a non-public supplement. The document is available in the EOSC-Life consortium as there has been established liaison between TC/276 WG5 and EOSC-Life. We are also actively exploring options to make the ISO standard open.

During the first 2 years of the EOSC-Life project, the Common Provenance Model has been accepted as a Preliminary Work Item under 23494-2 (Part 2 of ISO 23494) and it is being proposed for moving it into the next phase, the New Work Item at the time of submitting the deliverable.

References

1. Begley CG and Ioannidis JP. Reproducibility in Science. Circulation Research 2015;116:116–26.
2. Servick K and Enserink M. The pandemic’s first major research scandal erupts. Science 2020;368:1041–2.
3. Mobley A, Linder SK, Braeuer R, et al. A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. PLOS ONE 2013;8:1–4.
4. Morrison SJ. Time to do something about reproducibility. eLife 2014;3:1– 4.
5. Byrne JA, Grima N, Capes-Davis A, et al. The Possibility of Systematic Research Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Potential Solutions. Biomarker Insights 2019;14.
6. Prinz F, Schlange T, and Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 2011;10. Number: 9 Publisher: Nature Publishing Group:712–2.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

7. Freedman LP, Cockburn IM, and Simcoe TS. The Economics of Reproducibility in Preclinical Research. PLOS Biology 2015;13:1–9.
8. Mahase E. Covid-19: 146 researchers raise concerns over chloroquine study that halted WHO trial. BMJ 2020;369.
9. Chaplin S. Research misconduct: how bad is it and what can be done? Future Prescriber 2012;13. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/fps.88:5–76.
10. Committee on Responsible Science, Committee on Science, Engineering, Medicine, and Public Policy, Policy and Global Affairs, et al. Fostering Integrity in Research. Pages: 21896. Washington, D.C.: National Academies Press, 2017. doi : 10.17226/21896
11. Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. The Lancet 2014;383:166–75.
12. Freedman LP and Inglese J. The Increasing Urgency for Standards in Basic Biologic Research. Cancer Research 2014;74:4024–9.
13. Begley CG and Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature 2012;483:531–3.
14. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 2012;490. nature11556[PII]:187–91. 13
15. Consortium of European Taxonomic Facilities (CETAF) Code of Conduct and Best Practice for Access and Benefit-Sharing.
16. Benson EE, Harding K, and Mackenzie-dodds J. A new quality management perspective for biodiversity conservation and research: Investigating Biospecimen Reporting for Improved Study Quality (BRISQ) and the Standard PRE-analytical Code (SPREC) using Natural History Museum and culture collections as case studies. Systematics and Biodiversity 2016;14. Publisher: Taylor & Francis eprint: https://doi.org/10.1080/14772000.2016.1201167:525–47.
17. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 2016;3:160018.
18. Holub P, Kohlmayer F, Prasser F, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. Biopreservation and Biobanking 2018;16:97–105.
19. Groth P and Moreau L. PROV-Overview: An Overview of the PROV Family of Documents. 2013.
20. Moreau L et al. PROV-DM: The PROV Data Model. 2013.
21. De Nies T. Constraints of the PROV Data Model. 2013.
22. Buneman P, Caro A, Moreau L, et al. Provenance Composition in PROV. 2017.
23. Curcin V, Miles S, Danger R, et al. Implementing interoperable provenance in biomedical research. Future Generation Computer Systems 2014;34. Special Section: Distributed Solutions for Ubiquitous Computing and Ambient Intelligence:1–16.
24. Belhajjame K et al. PROV-O: The PROV Ontology. 2013.
25. Moreau L et al. Provenance Access and Query. 2013.
26. Valle M et al. A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC). 2020. doi: 10.2777/926037

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

27. Moreau L and Groth P. Provenance: An Introduction to PROV. Synthesis Lectures on the Semantic Web: Theory and Technology 2013;3:1–129.
28. Ciccarese P, Soiland-Reyes S, Belhajjame K, et al. PAV ontology: provenance, authoring and versioning. Journal of Biomedical Semantics 2013;4:37.

Delivery and schedule

The delivery is delayed: yes

The deliverable was slightly delayed due to capacity problems caused by COVID-19 (additional work and people on sick leave).

Adjustments made

None

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.