



Linking provenance and its metadata in multi-organizational environments

Document Version

Submitted manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Wittner, R., Gallo, M., Frexia, F., Leo, S., Pireddu, L., Mascia, C., Plass, M., Soiland-Reyes, S., Müller, H., Geiger, J., & Holub, P. (2023). *Linking provenance and its metadata in multi-organizational environments*.

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.



Linking provenance and its metadata in multi-organizational environments

Rudolf Wittner^{1,2,3}, Matej Gallo², Francesca Frexia⁴, Simone Leo⁴, Luca Pireddu⁴, Cecilia Mascia⁴, Markus Plass⁵, Stian Soiland-Reyes^{6,7}, Heimo Müller^{1,5}, Jörg Geiger⁸, and Petr Holub^{1,2,3}

¹BBMRI-ERIC, Graz, AT

²Faculty of Informatics, Masaryk University, Brno, CZ

³Institute of Computer Science, Masaryk University, Brno, CZ

⁴CRS4 – Center for Advanced Studies, Research and Development in Sardinia, Pula, IT

⁵Diagnostic and Research Center for Molecular BioMedicine, Diagnostic & Research

Institute of Pathology, Medical University of Graz, Graz, Austria

⁶Department of Computer Science, The University of Manchester, UK

⁷Informatics Institute, University of Amsterdam, Amsterdam, NL

⁸Interdisciplinary Bank of Biomaterials and Data Würzburg (ibdw), University and University Hospital of Würzburg, Würzburg, DE

Corresponding author:

Rudolf Wittner¹

Email address: wittner@ics.muni.cz

ABSTRACT

Reproducibility issues are widely reported in life sciences. As a response, scientific communities have called for enhanced provenance information documenting the complete research life cycle, starting from biological or environmental material acquisition and ending with translating research results into practice. The integrity and trustworthiness of such provenance can be achieved by applying versioning mechanisms and cryptographic techniques, such as hashes or digital signatures, which are provenance metadata. However, the available provenance literature lacks an analysis of mechanisms for the exchange of provenance and its metadata between organizations as well as a grounded proposal of linking provenance and its metadata. In this work, we provide an in-depth analysis of the approaches for coupling provenance information and its metadata with documented research objects in the context of multi-organizational processes, leading to the categorization of possible approaches, description of their key properties, and derivation of requirements for underlying provenance models. We address the requirements by proposing a mechanism for linking provenance and its metadata by extending the Common Provenance Model, the open conceptual foundation for the ISO 23494 provenance standard series, currently under development. The concepts are demonstrated and validated on two complex use cases. This work is intended as a harmonized source of information on provenance coupling in the context of exchange of provenance between organizations, which can be used when designing or choosing a provenance solution. This type of usage is exemplified in the extension of the Common Provenance Model as another step toward a provenance standard for life sciences.

1 INTRODUCTION

The verifiability of existing scientific results is the cornerstone of research, since new scientific advances are typically built on existing ones (Mobley et al., 2013). Reproducibility entails the verification of the results by re-executing experiments, which could be performed – depending on the context, purpose, and precise definition of the term – by the original team or a different one, with the same or a different experimental setup. Despite the fact that the meaning of the term *reproducibility* varies in literature (Plessner, 2018; Freedman et al., 2015) (in this work we use the term broadly to include both replicability and aspects of reusability), there is a clear consensus that reproducibility is a way to verify scientific results. Reproducibility requires the research products and associated metadata to be traceable. These metadata

47 consequently enable us to assess whether the results fit the purpose of a new study, which is a standardized
48 understanding of the term quality (International Organization for Standardization (ISO), 2015). The
49 trustworthiness of the metadata enables us to rely on them and provides us with guarantees that the
50 information is authentic, truthful, and not fabricated.

51 However, problems with the quality, trustworthiness, and reproducibility of research results have
52 been often reported in life sciences (Begley and Ioannidis, 2015; Servick and Enserink, 2020; Mobley
53 et al., 2013; Morrison, 2014; Holzinger et al., 2023; Byrne et al., 2019; Prinz et al., 2011; Nickerson
54 et al., 2016), impacting health, economics, and political decisions (Freedman et al., 2015; Mahase, 2020;
55 Chaplin, 2012; National Academies of Sciences, Engineering, and Medicine, 2017). Poor documentation
56 of data precursors, such as biological or environmental specimens from which the data was generated, is a
57 significant reason for these issues. Consequently, improved and standardized documentation of data and
58 its precursors used in research studies is requested by professional societies and researchers (Ioannidis
59 et al., 2014; Freedman and Inglese, 2014; Begley and Ellis, 2012; Landis et al., 2012; Benson et al., 2016).

60 Generally speaking, *provenance* is information about the history of an object throughout its life-
61 time (Muniswamy-Reddy et al., 2010). During the last decades, provenance has been widely adopted in
62 scientific domains to support traceable lineage of research objects, such as biological material, workflows
63 or data. The purpose of provenance adoption varies, for instance, to support replication of conducted
64 experiments (Moreau, 2011; Korolev and Joshi, 2014), or to assess the quality of data (Buneman and
65 Davidson, 2010; Imran and Agrawal, 2017), the source of biological material (Holub et al., 2018), or
66 related processes. Research objects are frequently exchanged between organizations. For instance,
67 biological material in a clinical trial could be acquired from a patient in a clinical setting, and the resulting
68 samples processed in a laboratory (so called pre-analytical processing) and stored in a biobank (Müller
69 et al., 2017; Zatloukal et al., 2018). The samples can be handed over to another institution for analysis,
70 and resulting data – such as omics or images – can be further processed and passed on to another academic
71 or industrial user as input for successive studies. The data can also be pre-processed, analyzed, and
72 potentially integrated with data coming from other sources. In such a distributed environment, each
73 organization can provide provenance information only about a limited part of the described object’s life
74 cycle. As a result, a complete provenance chain documenting the whole research process involving the
75 object or its derivatives is formed by individual provenance components that correspond to the various parts
76 of the life cycle. These components may be generated, managed, and stored independently by different
77 heterogeneous organizations (Fig. 1).

78 The resulting provenance chain serves as documentation of the object’s history. Enabling queries over
79 such distributed provenance information is important to achieve tasks like tracing the history of the object,
80 or assessing the object’s fitness for purpose. For instance, given a trained AI model, we may request
81 information about how the data used for the training and validation of the model was curated or how the
82 original biological material – from which the data used for the model training derives – was acquired,
83 since this information has profound impact on usability of the AI model and affects to which other data
84 sets the trained model can be applied. Thus the ability to examine the whole provenance chain is critical
85 to assess the application domain and quality of the resulting AI model.

86 In order to establish trustworthy provenance chains, properties, such as provenance authenticity,
87 integrity, and non-repudiation, must be supported (Ametepe et al., 2021). In addition, the distributed
88 provenance framework must provide means to make permissible modifications to the chain without
89 breaking its integrity and validity, and these modifications must be transparent and traceable. This feature
90 is necessary, for instance, when an erroneous provenance component of the chain is detected and must
91 be corrected. These properties can be achieved by recording relevant metadata about provenance – i.e.,
92 *meta-provenance*, or *provenance of provenance* – such as attributions, version numbers, hashes, or digital
93 signatures, corresponding to a provenance component. As a result, each component of a provenance chain
94 has corresponding meta-information (Fig. 1).

95 One of the main goals of current research on distributed provenance is to enable the unified traversal,
96 processing, and analysis of distributed multi-organizational provenance chains. The current state-of-the-
97 art model for distributed provenance information – the Common Provenance Model (CPM) (Wittner
98 et al., 2022) – provides a groundwork that enables traversal through the chain and describes how meta-
99 information related to the versioning of components can be represented. Motivated by the reproducibility
100 issues in the life sciences, the model is primarily developed and piloted in the use cases from this field.
101 However, it is generic enough to be also applicable in other domains. The CPM also serves as an open

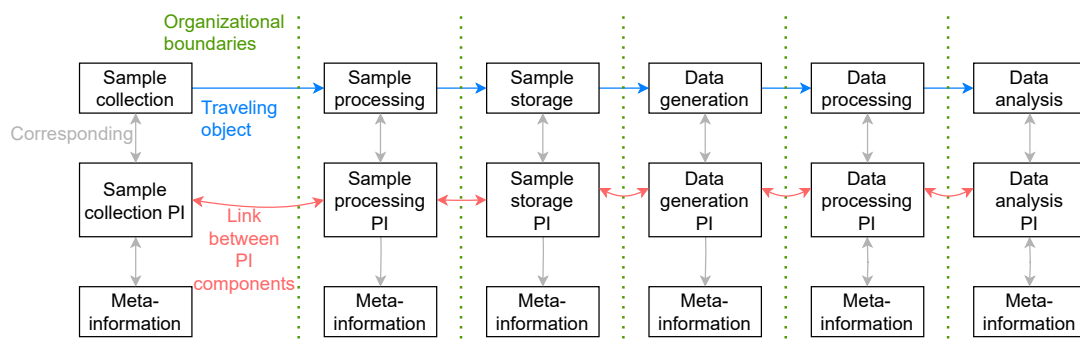


Figure 1. Illustration of a distributed provenance chain. Individual components of the provenance chain may be generated, managed, and stored independently by different heterogeneous organizations. Each component of the provenance chain has corresponding meta-information. For simplicity, here we show 1-to-1 cardinality of the correspondence between components and documented processes, however, generally, each process can be documented in multiple provenance components. "PI" stands for "provenance information".

102 foundation for the “ISO 23494 Biotechnology — Provenance information model for biological material
 103 and data” series (Wittner et al., 2023c), which is currently under development. However, the model does
 104 not prescribe how to create links between the components of a distributed provenance chain and their
 105 corresponding meta-information. Consequently, adopters of the model are free to decide what means
 106 are used to represent the links. This gap creates room for incompatibilities between implementations,
 107 potentially preventing the application of common general mechanisms to process the meta-information
 108 for tasks such as verification of provenance integrity and authenticity. In addition, determination of
 109 how provenance and corresponding meta-information is interlinked depends on the characteristics of the
 110 exchanged object and additional requirements that the attached provenance information must meet. These
 111 requirements may include whether meta-information about a provenance component can be distributed
 112 across multiple storage locations or whether meta-information about multiple provenance components
 113 may be integrated, which could affect the need for the uniqueness of identifiers used in meta-information.
 114 However, current literature on provenance lacks a thorough assessment of how the linking, characteristics
 115 of exchanged objects, and related requirements affect each other: the matter is either addressed marginally
 116 in the context of exchanged objects or not addressed at all.

117 Considering the wide range of areas where provenance information is being adopted, an in-depth
 118 analysis of such fundamental aspects may be of relevance to many domains where the research focus
 119 relies heavily on the analysis of data sets whose creation typically involves complex, distributed, and
 120 heterogeneous processes (Holzinger et al., 2023). Additionally, provenance models must be designed to
 121 integrate easily with existing domain-specific approaches (Curcin et al., 2014), so they are flexible enough
 122 to cover a wide range of use cases. For that reason, having a set of recommendations and guidelines
 123 related to the linking of provenance and meta-information in the context of exchanging research objects
 124 would be beneficial. As a response, multiple related research questions can be formulated:

- 125 1. How to represent and implement the domain- and scenario-agnostic links between components of a
 126 provenance chain and their related meta-information?
- 127 2. Can the links be used to traverse the distributed multi-organizational provenance chain using a single
 128 algorithm which is able to follow the links and exploit information stored in the meta-information
 129 during the traversal?

130 This work answers the above research questions and fills the aforementioned gaps by providing an
 131 in-depth analysis of the issue of coupling provenance information and meta-information with exchanged
 132 objects, resulting in a novel categorization of potential approaches for exchanging provenance and meta-
 133 information between organizations, and a description of key properties of the approaches. Based on this
 134 analysis, we derive general requirements for provenance models related to the interlinking of provenance
 135 and the corresponding meta-information, and we extend the CPM with a unified representation of these

136 links to address the identified requirements. Further, we demonstrate the feasibility of the proposed CPM
137 extension by using it to document two distinct use cases. Finally, we validate our results and demonstrate
138 that they fulfill the posed questions by implementing an algorithm that traverses the generated distributed
139 provenance chains, demonstrating that the CPM extension for harmonized representation of links between
140 provenance and meta-provenance supports not only the traversal of the chain itself, but also exploits the
141 links from provenance components to their meta-provenance to verify the integrity of each provenance
142 component using a hash stored in corresponding meta-provenance.

143 The main contributions of this work are the following.

- 144 • The introduction of *provenance exchange schemes* - a novel categorization of potential approaches
145 for exchanging between organizations provenance, meta-information, as well as descriptions of
146 their respective properties – including, for instance, the confidentiality of sensitive data.
- 147 • The formulation of general requirements for provenance models to enable the interlinking of
148 provenance and the corresponding meta-information.
- 149 • The extension of the CPM to handle the new requirements. In particular, the definition of how to
150 link provenance and meta-information in terms of the provenance model using persistent identi-
151 fiers (Hellström et al., 2020) (PIDs) and attributes of related provenance structures.
- 152 • The implementation of provenance and meta-provenance generation in accordance with the extended
153 CPM. The procedure was implemented for two distinct use cases: 1) a digital pathology use case
154 (also used as a running example throughout this article); 2) a ColoRectal Cancer (CRC) cohort
155 extension use case.

156 This paper is structured as follows. The Background section (Section 2) describes the concepts on
157 which this work builds. The Methods section (Section 3) describes how the analysis of methods for
158 coupling documented objects and their provenance was conducted, and how the CPM and its current
159 extension were developed and validated. The Related Work section (Section 4) provides a survey of
160 approaches for coupling provenance with documented objects, describes the current state-of-the-art
161 for distributed provenance information models and systems, and shows examples of how provenance
162 information can be exchanged. In the Results section (Section 5), the provenance exchange schemes are
163 defined, the requirements on links between provenance and its metadata are stated and addressed by an
164 extension of the CPM, and the implementation and validation of the concepts are described. Finally, the
165 Discussion section (Section 6) outlines various aspects of the proposed provenance model, such as the
166 importance and relevance of the presented work, practical aspects that must be considered when adopting
167 the CPM, and directions for future work.

168 In addition, the supplementary materials contain the technical description of how the proposed
169 concepts were implemented for the use cases, and the description of the traversal algorithm used for the
170 validation of our results. See the "Availability of Supporting Data and Materials" section to find references
171 to the code and related digital objects.

172 This manuscript was previously published as a preprint (Wittner et al., 2023a).

173 **2 BACKGROUND**

174 This section describes concepts we use in our work as a starting point. These include a running example
175 used to pilot the implementation of the presented contributions, approaches for coupling provenance and
176 described objects (at present mainly limited to the storage aspects), and relevant aspects of the CPM.

177 **2.1 Running Example**

178 The example, which has also been used for the development of the model and prototype implementation
179 of the proposed concepts, comes from the digital pathology domain. Digital pathology is a field in
180 which imaging technologies are applied to enable the acquisition, management and interpretation of
181 pathology information generated from digitized glass slides. In this context, machine learning supports
182 the development of systems based on trained AI models that consume clinical data and high-resolution
183 scans of histopathological biological material – i.e., Whole Slide Images (WSIs). The use case represents
184 a process involving several phases spread across different organizations – namely, a hospital, an analytical
185 laboratory, and a research group.

186 This use case consists of the following steps.

- 187 1. **Biological material acquisition** is done via surgery in a hospital. Primary samples are taken as
188 part of a medical treatment and sent to a pathology department for examination.
- 189 2. **Biological material processing, examination, and image data generation** are done as part of the
190 diagnostic process in a pathology laboratory. This phase consists of generation of tissue blocks,
191 cutting of tissue blocks into slices to be placed on glass slides, staining the slides, and scanning
192 them. The resulting WSIs are examined and annotated by a pathologist. The annotations depict
193 tumor areas and other morphological features. The annotated scans are provided to a research
194 group, where they are used as an input for an AI-based computational workflow.
- 195 3. **WSI data preprocessing**. The goal of the data preprocessing phase in this example is to split the
196 high-resolution WSIs into smaller segments, as the AI model can not process an entire WSI at once.
197 Prior to the splitting, each WSI is assigned either to a training or testing data set.
- 198 4. **AI model training**. The training data set, which includes a portion of the input WSIs and their
199 annotations, is provided as input to the AI model training process. The model is trained to detect the
200 presence of carcinoma cells in the WSIs. A portion of the training dataset is held out for validation.
201 The result of this step is a trained AI model (i.e., the model architecture with assigned weights with
202 the best classification performance on the validation set).
- 203 5. **AI model evaluation**. The trained AI model is applied to the testing data set to predict the presence
204 of carcinoma cells in the WSIs. The computed predictions are then compared with the original
205 annotations to evaluate the trained model's performance.

206 Trustworthy provenance information documenting the pipeline plays a crucial role in the application
207 of the trained model. For instance, the resulting provenance chain, and respective meta-information could
208 be used to prove compliance with regulations, such as the In Vitro Diagnostic Regulation (Spitzenberger
209 et al., 2022) and the Medical Device Regulation (MDR), to evaluate fitness-for-purpose of the trained AI
210 model (as the model may be trained for a specific category of biological samples), or to trace origins of
211 errors or inconsistencies in the input data set (Müller et al., 2022).

212 2.2 Coupling Provenance and Documented Objects

213 One of the fundamental aspects affecting the properties of provenance chains is provenance coupling –
214 i.e., whether provenance information is stored as part of a documented object or whether it is a standalone
215 piece of information linked with the object externally. The following three coupling schemes are currently
216 described in the literature.

- 217 • **Tight/high coupling**. Provenance is stored directly with the data for which provenance is
218 recorded (Glavic et al., 2007).
- 219 • **Loose coupling**. Provenance and data are stored in a single system but logically separated (e.g., by
220 storing data and their provenance in different tables of the same database (Pérez et al., 2018; Glavic
221 et al., 2007).
- 222 • **No coupling**. Provenance is stored in one or many repositories which are separate from the data
223 repository (Glavic et al., 2007).

224 Since determining a coupling scheme is an essential architectural question that affects where and how
225 the links between provenance and corresponding meta-information are stored, it significantly impacts the
226 properties of the resulting provenance chain and communication between organizations. However, the
227 coupling schemes are primarily considered storage methods, while how they affect provenance in transfer
228 is not described in literature.

229 In this work, we revise the coupling schemes in the context of distributed provenance and exchanged
230 research objects, describe their properties, and derive the requirements they pose on underlying provenance
231 models.

2.3 Distributed Provenance Information & CPM

The Common Provenance Model is a novel model for representing distributed multi-organizational provenance information. The main goal of the model is to enable the creation of distributed provenance chains across heterogeneous multi-organizational environments, with support for unified traversal and querying mechanisms, independently from particular processes or research objects documented by the provenance. Thus, the CPM directly addresses the traceability of research objects and their provenance in both backward and forward direction. Additionally, depending on the specific provenance content, the CPM supports the reproducibility of research results and related experiments.

The core concept of the model is that each organization involved in a documented object's life cycle generates standardized provenance information represented as a single component of a chain (called *bundle* in terms of the CPM and underlying W3C PROV data model (Belhajjame et al., 2013)), and links it to each existing provenance produced by other actors involved in the object's life cycle. In particular, a described object and its provenance are transferred between organizations – from a *sender* to a *receiver* (Fig 2). The exchanged provenance is generated during a *finalization event*, which is a specific time instance when available information from log files or information systems is translated into a data model that conforms to the CPM. The *finalized provenance information* is archived by the sender and provided to the receiver with the described object. This way, the receiver is provided with a standardized representation of provenance together with the described object and can use it to assess the object's fitness for purpose or for other purposes. The receiver, in turn, can use the object and generate additional finalized provenance information, which links to the previous provenance component of the chain that was previously archived by the sender (*backward link*). In addition to the backward link, optionally, the receiver can inform the sender about the new finalized provenance component to update the sender's finalized provenance information to include the *forward link* to the receiver's finalized provenance. This process results in a distributed provenance chain (Fig 2).

A part of a provenance chain documenting the computational steps of the running example is designed to include three provenance components (Fig. 3), each documenting an individual step of the example: data preprocessing, AI model training, and AI model evaluation. The decision to create three individual components for the example was based on the fact that despite that each of these steps is handled by the same research group, the steps can be executed at different moments in time, and there might be significant delays between the executions. Additionally, the output of one step can be re-used for multiple successive steps. For instance, one preprocessed data set can be used for multiple training actions (e.g., each with different model hyperparameters).

As the CPM is built on the W3C PROV standard, respective provenance is represented as a graph structure with annotated nodes and edges to express their semantics. The nodes represent activities, entities, or agents, and the edges represent their mutual relationships. The CPM extends the PROV model with definitions of specific semantics to link the provenance components of the chain and to represent

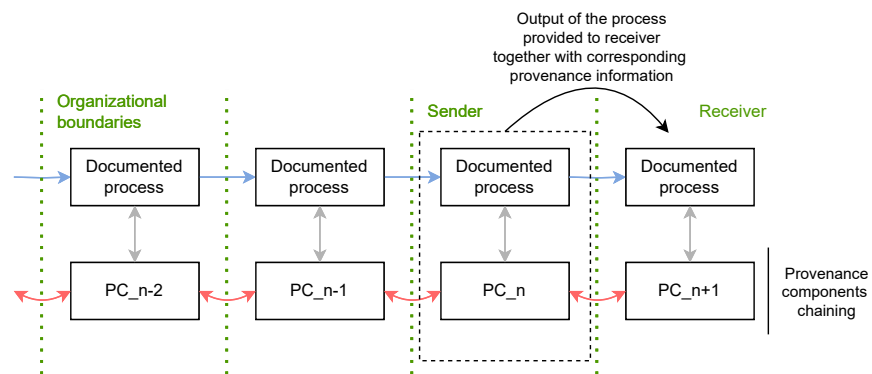


Figure 2. Output of a documented process and its provenance is passed from sender to receiver. The receiver uses the object as an input of its process, generates corresponding finalized provenance information, and links it to the previous provenance component of the chain. "PC" stands for "provenance component".

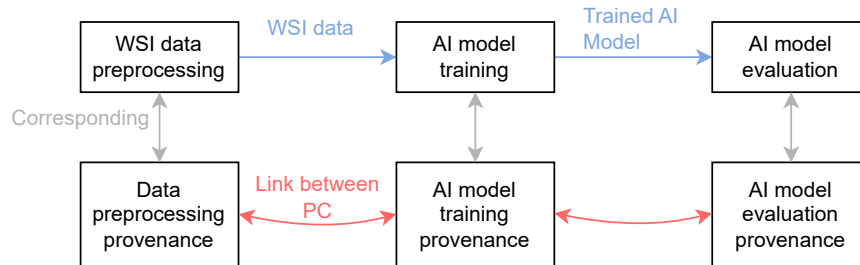


Figure 3. Illustration of a provenance chain documenting the AI pipeline execution. "PC" stands for "provenance component."

269 standardized derivation paths between inputs and outputs of a process. The forward and backward links
 270 between provenance components are implemented by a special type of entity, the *connector* (Fig. 4). A
 271 *forward connector* is a provenance structure that represents a snapshot of the described object at the time
 272 when it is sent from a sender to a receiver. The sender creates the forward connector, includes it in its
 273 finalized provenance, and provides it to the receiver. The receiver then includes this provenance structure
 274 (using the same id) as *backward connector* in its finalized provenance information and creates another
 275 provenance structure – the *current connector* entity – which represents the snapshot of the described
 276 object at the time of its receipt. Finally, the receiver creates an edge between these two structures to
 277 express the derivation path between the two states of the object. If the receiver provides the results of its
 278 process to another organization, a new forward connector is created in the sender’s finalized provenance.
 279 This forward connector is then related using the *Derivation* relation to the current connector and provided
 280 along with the described object and its provenance to the new receiver. This process is applied iteratively
 281 each time a described object is passed between organizations. This set of standardized derivation paths
 282 between inputs and outputs of documented processes is called *provenance backbone*, which forms the
 283 core of the resulting provenance chain. The CPM also prescribes additional provenance structures and
 284 a method to attach domain-specific information to the derivation paths. However, since this part of the
 285 CPM is irrelevant to this work, its description is omitted.

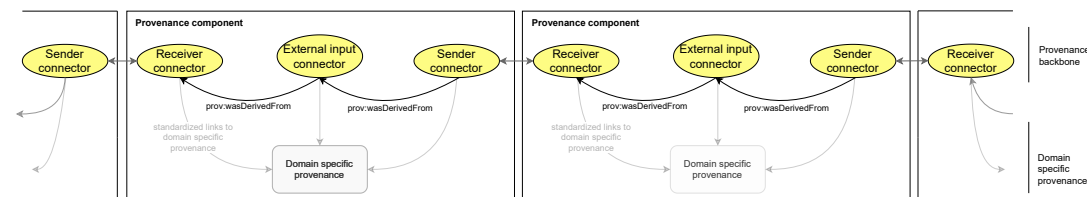


Figure 4. Illustration of a provenance backbone. The backbone is formed by standardized derivation paths between nodes of the underlying provenance graph and provides the core foundation for distributed provenance chains. In addition, the CPM prescribes a standardized method for attaching domain-specific information to the chain. In the original CPM work (Wittner et al., 2022), the terms *sender connector*, *receiver connector*, *external input* were used. In this work, we use the terms *forward connector*, *backward connector*, and *current connector* instead. This is because these terms were changed in the CPM specification since the publication of the original work.

286 The connectors introduced to document the use cases represent inputs and outputs, which are being
 287 exchanged between the steps of the documented processes. An illustration of connectors usage for the
 288 computational steps of the digital pathology use case is shown in Fig. 5 (all connectors are described in the
 289 "Supplemental Article 1.pdf"). For the WSI data preprocessing step, the input WSI data set is represented
 290 as a single backward connector, and the resulting training and testing data sets are represented as two
 291 separate forward connectors. For the AI training step, two backward connectors are present: 1) one linking
 292 to the original WSI dataset; 2) one representing the index table, that defines subset of the WSI dataset
 293 to be used for the training. The training step has a single output – the trained AI model – represented as a
 294 forward connector. The provenance component documenting the AI model evaluation step includes three

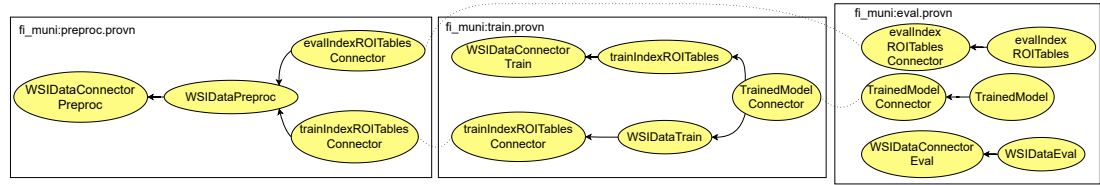


Figure 5. Illustration of a provenance backbone for the computational steps of the running example. The nodes in the figure are connectors that represent traceable objects, and are used to link components of a provenance chain. The arrows between the connectors represent `prov:wasDerivedFrom` relation. The dotted lines between connectors in different bundles represent a link between the two bundles, that is realized through particular connectors (backward and forward connectors).

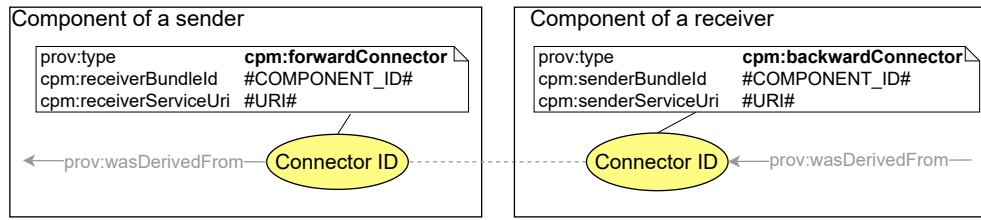


Figure 6. Illustration of connectors' attributes.

295 backward connectors – one linking to the trained model from the AI training step, the second one linking
 296 to the original WSI dataset, and the third one representing the index table defining subset of the WSI
 297 dataset to be used for the evaluation. No forward connector is present in the last provenance component
 298 to indicate that outputs of the evaluation step have not been used in any consecutive process yet, so this is
 299 the end of the provenance chain.

300 The links between distributed provenance components are implemented as attributes of the connectors
 301 (Fig. 6). In particular, the connectors must include an identifier of the destination provenance component
 302 and a service identifier where the corresponding provenance component can be requested.

303 In this work, we revise the attributes of connectors to better align with new requirements that emerged
 304 as a result of the presented work. Additionally, we extend the CPM with means to refer to meta-information
 305 about provenance components, complying with the new requirements, and suggest how to apply PIDs to
 306 connectors.

307 2.4 Versioning of Distributed Provenance in CPM

308 The CPM describes provenance versioning, a method to perform authorized changes of provenance
 309 chains, and prescribes how to represent the change in meta-information – *meta-provenance* in terms
 310 of the CPM. To update a chain component, a new component is created and linked with the original
 311 version in meta-provenance. The new version that supersedes the original version is considered as a
 312 replacement of the original component, and can contain new, reduced, or updated information. The
 313 original component must not be deleted but is kept archived in the original location to avoid disrupting
 314 the chain's integrity since other components may still refer to the original version. As a result, each
 315 provenance component of the chain may have multiple historical versions and refer to a specific version
 316 of another chain component (Fig. 7).

317 The CPM defines a standardized way of representing provenance component versions in meta-
 318 provenance (Fig. 8). The proposed mechanism for provenance components versioning loosely follows
 319 the semantics defined in the PAV ontology (Ciccarese et al., 2013) and is an application of a provenance
 320 revision pattern (Moreau and Groth, 2013). The resulting scheme is depicted in Fig. 8.

321 In this work, we build on the existing mechanism to include standardized links between meta-
 322 provenance and provenance.

323 2.5 Appending New Information To a Chain in CPM

324 If a documented research object is modified, the corresponding provenance component must be appended
 325 to the chain, and the corresponding meta-provenance must be generated. The CPM provides two general

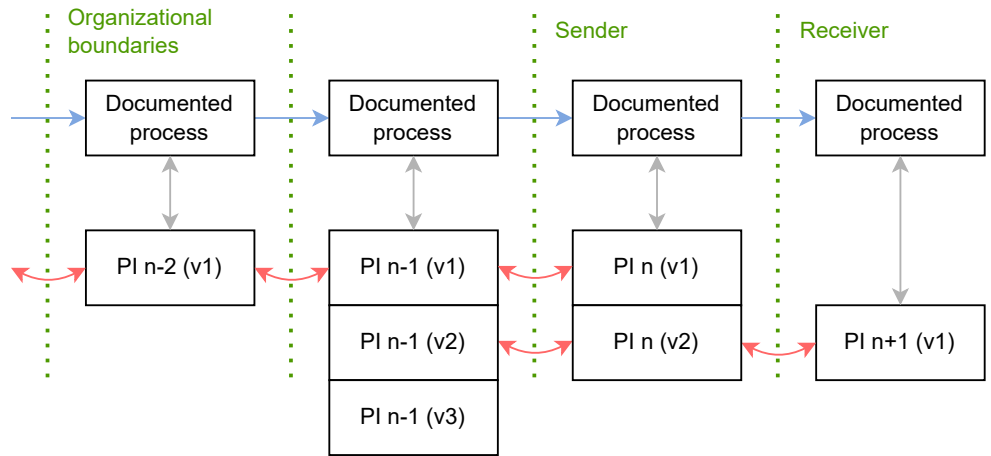


Figure 7. Illustration of different versions of provenance chain components. Different versions can be created as a result of error correction in provenance, inclusion of forward link in provenance components, or for other reasons.

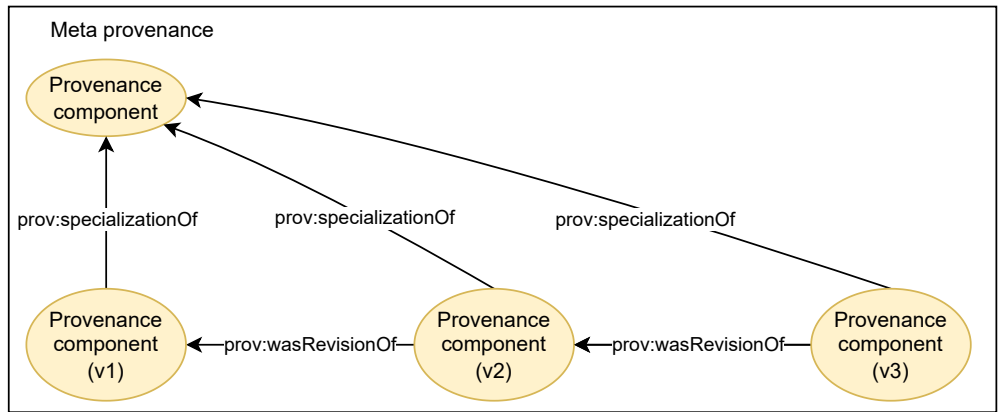


Figure 8. Schema of how different versions of a provenance component are represented in meta-provenance according to the CPM. In particular, each component version is represented in meta-provenance as a provenance structure of type *Bundle*. Each new component version is related to the previous version of the component using the *Revision* relation. Finally, all the versions are related to the common abstract entity that represents the common aspects of all the versions using the *Specialization* relation.

326 strategies for appending new information:

- 327 1. By the addition of a new provenance component to a provenance chain, whose connector refers to
328 the existing finalized provenance component (Section 2.3);
- 329 2. By replacing the latest provenance component in the chain with a new version, presuming this
330 operation does not modify existing provenance descriptions (Section 2.4).

331 Intuitively, if a described research object is modified and a new version of that object is obtained, the
332 versioning mechanism should be applied. On the other hand, if an object is modified and a new object
333 is obtained as a result, then a new provenance component should be appended to a provenance chain.
334 However, determining when an object can be considered a new version or a new object is not always
335 straightforward. Additionally, as it will be shown in this work, each appending strategy has implications,
336 and it may not always be applicable despite the original intuition. Being familiar with these properties is
337 essential to make informed decisions when designing or choosing a provenance solution. However, the
338 benefits of choosing one appending strategy over the other in the context of distributed provenance chains
339 are not explored.

340 In this work, we analyze the general appending strategies in the context of exchanged research objects
341 and list general aspects that affect decision-making.

342 **3 METHODS**

343 To answer the research questions and define a mechanism to link components of provenance in a meta-
344 provenance in a chain (research question one) traversable with a single algorithm (research question two),
345 the work was conducted in three phases: 1) the analysis of relevant literature about coupling described
346 objects, provenance and meta-provenance; (Section 3.1); 2) the extension of the Common Provenance
347 Model (Section 3.2); 3) the validation of the results (Section 3.3).

348 **3.1 Conducting an analysis for the linking of represented objects, provenance and meta- 349 provenance**

350 The goal of the investigation we conducted was to find information, analysis and recommendations
351 about how the coupling of described objects, provenance, and meta-provenance, mutually affect each
352 others' properties. Since literature surveys and systematic literature reviews are important in categorizing
353 different approaches, harmonizing related terminology and perspectives, and often provide a consolidated
354 overview of heterogeneous requirements which are originally fragmented across the literature, we used
355 them as a basis for finding the required information.

356 The Systematic review of provenance systems (Pérez et al., 2018) was used as a starting point
357 for our related work survey, as we consider it the most comprehensive provenance literature review.
358 The Systematic review identified 251 provenance systems and selected 105 papers as a basis for the
359 development of a taxonomy of provenance characteristics. The taxonomy was created as an extension of
360 a previous categorization (Cruz et al., 2009), which was considered the most complete one at the time of
361 publication of the systematic review. In this context, we have analyzed the systematic review, the original
362 categorization, and the surveys and reviews referred to from the systematic review, which were listed in
363 the storage dimension of the taxonomy (Dogan, 2016; Cruz et al., 2009; Glavic et al., 2007), as this one
364 contained a description of the provenance coupling schemes, which was the most relevant for our work.

365 Since the taxonomy was based on the work published between 2001 and July 2017, we wanted
366 to ensure that the analysis we sought was not conducted later. Therefore, we extended our survey to
367 work published between July 2017 and August 2022. In particular, we used Google Scholar (<https://scholar.google.com/>) and searched for recent work using the keywords “provenance survey”
368 and “provenance review” in the publication name, which were published in the range between 2017 and
369 2022, and checked the first five pages of the search results. Inclusion criteria were that the result was:
370 i) a survey or a systematic review related to provenance information; ii) available to us without further
371 payment; iii) an officially published article (not a preprint). In addition to this, we included the work
372 that cited the systematic review and was categorized as “taxonomy”, “review”, or “survey” to consider
373 additional provenance-related surveys, reviews, or taxonomies. As a result, 22 additional papers published
374 between July 2017 and August 2022 were identified as relevant and analyzed.
375

376 As this methodology restricts the survey to only what the original reviews' authors envisioned as
377 relevant before, we decided to analyse eight additional provenance systems that handle distributed
378 provenance information, namely: Karma (Simmhan et al., 2006, 2008), Chimera (Clifford et al., 2008;
379 Foster et al., 2002), Whips (Cui and Widom, 2000; Wiener et al., 1996), Buneman (Buneman et al., 2006),
380 Orchestra (Ives et al., 2008; Green et al., 2007), Lipstick (Amsterdamer et al., 2011), PLUS (Chapman
381 et al., 2011; Blaustein et al., 2008; Chapman et al., 2010; Allen et al., 2011, 2012), RAMP (Park et al.,
382 2011; Ikeda et al., 2011).

383 Options for handling provenance and meta-provenance with regard to exchanged objects, their
384 properties, and the derivation of general requirements on provenance models, are presented in the Results
385 section (Section 5). These results are based on the analysis of the related work, iterative deep discussions
386 of all the authors, and lessons learned during the process of refining the CPM summarized in the next
387 paragraph (Wittner et al., 2023d).

388 **3.2 Extending the Common Provenance Model**

389 Once the properties of the schemes and requirements on provenance models were formulated, we
390 started to answer research question one developing an extension of the CPM. The first draft of the
391 CPM was published as a deliverable (Wittner et al., 2021b) of the EOSC-Life project ([https://www.
392 eosc-life.eu/](https://www.eosc-life.eu/)). The initial version was later extended (Wittner et al., 2022) with the provenance
393 backbone concept, and further aspects were added, namely opaque provenance components, integrity, non-
394 repudiation, and support for missing provenance components. In order to further validate these introduced
395 concepts, we integrated the CPM with RO-Crate – an implementation of FAIR digital objects (De Smedt
396 et al., 2020) – which was the main trigger for the conducted analysis and further extensions of the CPM
397 presented in this work. In particular, several informed decisions were required in order to integrate the
398 CPM with the RO-Crate: whether to allow splitting a component into multiple files or having multiple
399 provenance components in a single file; how to represent meta-provenance in an RO-Crate; or how to
400 represent links between provenance components and meta-provenance. Consequently, formulating these
401 design questions led us to the formulation of the research questions presented in the Introduction.

402 The extension presented in this work was created to enable harmonized links between components of
403 a distributed provenance and their corresponding meta-provenance. The development of the extension
404 was piloted on the two use cases.

405 **3.3 Validation of the results**

406 To answer research question two, and to validate the extension of the CPM, we had to show that the
407 generated provenance information can be traversed using a single algorithm which is not affected by
408 what kind of process or object is documented, or from which organization the object comes from. In
409 addition, the traversal algorithm has to exploit the links between provenance components and their
410 meta-provenance. This was addressed in this work by proposing and implementing a provenance chain
411 traversal algorithm, which lists precursors of an object requested as an input to the algorithm. As the
412 generated provenance chains are composed of components from distinct organizations, the algorithm
413 traverses the components stored in different locations. Hashes of the provenance chain components are
414 stored as part of meta-provenance, and the algorithm uses this information to verify the integrity of the
415 provenance components, to verify that the proposed linking mechanism can be used to traverse the chain
416 and exploit information stored in meta-provenance.

417 **4 RELATED WORK**

418 The related work section is divided into three main parts. The first part provides an overview of literature
419 related to approaches for coupling provenance and documented objects. The second part focuses on
420 distributed provenance information and provides an overview of relevant provenance systems and models.
421 The third part presents examples of methods to exchange provenance information between organizations.

422 **4.1 Approaches for coupling provenance and documented objects**

423 A Systematic Review of Provenance Systems (Pérez et al., 2018) is the most comprehensive provenance-
424 related systematic literature review. The main contribution of the work is a unified taxonomy of provenance
425 systems characteristics. The taxonomy is based on 105 provenance-related papers but is not specific
426 to any domain or provenance management technique. The most relevant taxonomy dimension related

427 to our work is the storage dimension, since it includes the coupling mechanisms, affecting who and
428 how appends new provenance information to the chain. However, the taxonomy does not describe any
429 properties nor requirements on how the different coupling mechanisms – loose-coupling, no-coupling,
430 tight-coupling – behave when documented objects are exchanged between organizations. We presume
431 that these properties were not comprehensively described in the literature earlier. Otherwise, they would
432 be described or referenced from the review, which is not the case. The original work (Glavic et al.,
433 2007) that introduced the three coupling categories states that the no-coupling strategy can deal with
434 heterogeneous environments and that most annotation-based approaches use tight or loose coupling
435 strategy. Other literature related to the storage taxonomy dimension just repeats this information (Cruz
436 et al., 2009) or omits any description of the coupling schemes at all (Dogan, 2016). Although a deeper
437 analysis of the coupling schemes is not presented in the survey, some fragments can be found in the
438 literature. (Muniswamy-Reddy et al., 2006; Zafar et al., 2017) state that the benefit of the tight coupling
439 scheme is that it provides better support for ensuring consistency between provenance and data during
440 manipulation. On the contrary, separating provenance from data enables better separation of access
441 policies for provenance and data.

442 The systematic review covers work up to July 2017. Since then, several surveys and provenance-
443 related review papers that tackle coupling provenance and documented objects have been published.
444 (Herschel et al., 2017) provides an overview of what provenance is used for, what types of provenance
445 have been defined and captured for different applications, and which resources and system requirements
446 affect the choice of deploying a particular provenance solution. However, the Decoupling category
447 in this overview is not related to the coupling of provenance to documented objects but concerns the
448 coupling of provenance collection mechanisms with existing systems. In (Hu et al., 2020), the authors
449 propose a number of design requirements for data provenance in IoT and provide a deep-insight review of
450 existing schemes of IoT data provenance. However, as the concerned coupling mechanism is a property of
451 overall provenance architectural design rather than a requirement, a deeper discussion on this topic is not
452 included. On the other hand, the attachment of provenance information to data and separate provenance
453 information management are mentioned as two distinct approaches to provenance management, and the
454 authors identify flexible data provenance management as a future research direction. (Pimentel et al.,
455 2019) survey state-of-the-art for the provenance of scripts and propose a taxonomy for this field, which
456 includes three dimensions – provenance collection, provenance management, and provenance analysis.
457 The most relevant to our work, the Distribution category under the Management dimension, is dedicated
458 to means of distributing provenance information to consumers – local (OPM files, PROV files, logic
459 programming formats (e.g., Prolog or Datalog files) or graph formats (e.g., Graphviz files)) and remote
460 (repositories and web). Although these can be seen as potential options for implementing the exchange,
461 any discussion about how the dimensions relate to the coupling is not presented. (Khan et al., 2019)
462 reviews best-practice recommendations for workflow enactment metadata sharing and applies them in
463 CWL PROV specification, which results in a *CWLProv research object*, a standardized representation of
464 shareable data and metadata for workflow execution. The specification addresses the recommendations
465 related to preserving workflow-related information, such as execution parameters, inputs, intermediate
466 results, or provenance. However, as (Khan et al., 2019) is focused on the format and tooling, a deeper
467 analysis relevant to the provenance exchange schemes is not provided.

468 The rest of the recent provenance-related reviews and surveys do not tackle the coupling schemes or
469 exchanged objects at all. (Bai et al., 2021) provides an overview of security enhancements for provenance
470 in the Internet of Health Things domain. (Xu et al., 2020) focuses on analyzing the provenance of
471 human-computer interactions. (Tufek and Aktas, 2022) is a systematic literature review that aims to map
472 how provenance is handled in the Numerical Weather Prediction Models domain. (Gierend et al., 2021)
473 describes a protocol for a scoping review of provenance in biomedical data sets and workflows, but at the
474 time of conducting our survey the actual review was not available yet. (Oliveira et al., 2018) concerns
475 the problem of extracting useful information out of huge amounts of collected provenance information.
476 It surveys state-of-the-art work related to provenance analytics and proposes a taxonomy to categorize
477 related aspects. (Ametepe et al., 2021) surveys provenance collection methods and their security. (Kale
478 et al., 2023) provides a bibliometric analysis of explainable AI, trustworthy AI, and provenance-related
479 literature. (Rrmoku et al., 2022) presents a literature review of approaches and the influence that social
480 network analysis and data provenance have on recommender systems. (Zipperle et al., 2022) provides an
481 evaluation of research in the Provenance-based Intrusion Detection Systems field and proposes a novel

482 taxonomy for the systems.

483 **4.2 Distributed Provenance Information**

484 We understand the term distributed provenance to imply a possibly unlimited scope of use cases that can
485 be documented, as we may recursively ask for the provenance of any process inputs (sometimes called
486 open world provenance (David Allen et al., 2015)). In this sense, distinct parts of distributed provenance
487 can be generated, stored, and managed independently. As it will be shown, such a distributed provenance
488 has not attracted much attention so far. Here, we describe the distributed provenance information related
489 literature in two parts: 1) systems for distributed provenance collection and management; 2) provenance
490 information models for distributed provenance.

491 **4.2.1 Distributed Provenance Information Systems**

492 There are several systems to handle distributed provenance (Pérez et al., 2018). These systems consist of
493 *multiple logically interrelated repositories, which are distributed over a computer network.*

494 The PLUS system (Chapman et al., 2011; Blaustein et al., 2008; Chapman et al., 2010; Allen et al.,
495 2011, 2012) is the closest to meet the "open-world" presumption, as it intends to enable provenance
496 capture, storage and use across multi-organizational systems. It presumes the open world environment,
497 which consists of distributed heterogeneous environments with no assumption of control over systems
498 from which provenance is captured (including legacy systems). The system was designed as a centralised
499 database which is accessible via API to capture provenance. The API functions are invoked by so called
500 *coordination points* – such as Enterprise Service Bus – which is used for communication between the
501 systems. The coordination points contain a provenance collector which extracts relevant information
502 from the communication and capture it as provenance stored in the centralised database. The open-world
503 assumption is satisfied by the coordination point, through which heterogeneous systems can communicate.
504 In order to completely fulfil the open world requirement, another step towards is required – to enable
505 interconnection and traversal of multiple PLUS system instances, and a preliminary work has been done
506 in this direction (Allen et al., 2011). The PLUS system was deployed into a peer-to-peer network, where
507 traversing the provenance graph is realised using *advertisements* – a sort of catalogue that lists what
508 provenance descriptions are held by each network node. Traversing the graph is then realised by a
509 recursive function that gets available advertisements corresponding to the described object, and then
510 requesting the provenance at each node separately via an API. (David Allen et al., 2015) builds on the
511 experience with the PLUS system, and provides general engineering decisions for open world provenance
512 systems. The decisions concern identifiers of objects (content or context bound), provenance storage
513 (hierarchical files-based, relational, graph databases), or provenance protection. However, the work does
514 not describe decisions related to linking provenance and its meta-provenance.

515 There are several other systems that can handle distributed provenance. Chimera (Clifford et al.,
516 2008; Foster et al., 2002) collects provenance from computational workflows that are executed on a grid.
517 Karma (Simmhan et al., 2006, 2008) addresses collections of provenance from workflows composed of
518 grid and web services. The Lipstick framework (Amsterdamer et al., 2011) captures both fine and coarse
519 grained provenance from workflows that can span multiple organizations. Orchestra (Ives et al., 2008;
520 Green et al., 2007) is a system for data sharing among heterogeneous peers related by database schema
521 mappings with support for provenance, which is used to assess trust of the database systems updates.
522 (Buneman et al., 2006) proposed a system for fine grained provenance capture documenting moving of
523 data in the context of curated databases. RAMP (Park et al., 2011; Ikeda et al., 2011) supports provenance
524 capture and tracing for MapReduce workflows (Dean and Ghemawat, 2008), and is implemented as a
525 wrapper around Hadoop jobs and transformation. The Whips (Cui and Widom, 2000; Wiener et al., 1996)
526 system can capture provenance of collection, transformation, and integration of data in a data warehouse.
527 All of these systems either concern a limited scope of a specific area or use case, or break our distributed
528 and independent provenance management presumption. Finally, none of these works specifically focuses
529 on linking provenance and meta-provenance – the linking mechanism is implemented "somehow" without
530 a motivation to derive more general requirements, or describe general properties of available approaches,
531 or do not implement a linking mechanism at all.

532 **4.2.2 Distributed Provenance Information Models**

533 The first attempt to formalize distributed provenance (Buneman et al., 2016) as the provenance of two
534 independent communicating processes was built on ideas from graph grammars (Bao et al., 2012),

535 recursive state machines (Alur et al., 2005), graph rewriting (Rozenberg, 1997), and hypergraphs (Drewes
536 et al., 1997). The Provenance Composition Pattern (Buneman et al., 2017) is an implementation of the
537 concept, which applies the idea of shared identifiers for relevant provenance structures, and enables pasting
538 the provenance graphs over these structures (similarly to the JOIN operation in relational databases).
539 The idea of shared identifiers has already been used in implementations, for instance, in the context of
540 decentralized operating systems provenance (Ahmad et al., 2020), or has been recommended as a best
541 practice (Khan et al., 2019) for computational workflows provenance to navigate between provenance of
542 different granularity.

543 W3C PROV (Groth and Moreau, 2013) is a general provenance information standard that aims
544 to support interchange provenance information in heterogeneous environments using widely accepted
545 technologies and formats, such as XML or RDF. One of the main features of the PROV data model is
546 its wide applicability, so it can be adopted in various domains to describe any item – digital, physical,
547 or conceptual. In addition, the PROV introduces a concept of bundles, a named set of provenance
548 descriptions that can be used to pack provenance information, and provenance of which can be expressed
549 in terms of provenance of provenance, or meta-provenance in other words. The corresponding PROV-
550 LINKS document (Moreau and Lebo, 2013) that defines the bundles highlights the necessity of linking
551 provenance bundles coming from different sources and defines semantics to implement the links. However,
552 the mechanism cannot be directly applied to create bi-directional links between bundles (Wittner, 2022)
553 since their integration would lead to invalid provenance information. In addition, as one of the features of
554 the PROV model is its generality and wide applicability in different domains, its uncoordinated application
555 leads to incompatible solutions.

556 To address this gap, the Common Provenance Model (Wittner et al., 2022) has been designed. The
557 CPM is built on top of PROV and further specifies the required aspects of distributed provenance, which
558 were not addressed in PROV specifically. In comparison to PROV, it aims for advanced interoperability of
559 provenance by defining how to build provenance chains using a domain-agnostic provenance backbone
560 to which domain-specific information is attached in a standardized way and provides a standardized
561 groundwork for provenance components versioning, authenticity, integrity, and non-repudiation. In
562 particular, the CPM reuses the idea of shared identifiers and extends it with the definitions of standardized
563 derivation paths between entities with the shared identifiers – the connectors. The CPM is an open
564 conceptual foundation of ISO 23494 provenance standard series (Wittner et al., 2023c, 2021a), and can be
565 considered as the current state-of-the-art provenance model for distributed provenance. As the model is
566 being developed, open issues still have to be addressed to enable unified traversal and processing of the
567 provenance chains. As identified, presented, and addressed in this work, there was an existing gap related
568 to linking provenance components and their meta-provenance.

569 **4.3 Provenance Information Exchange**

570 PROV-AQ (Klyne et al., 2013) is a specification of how to exchange provenance information using
571 standard web protocols. In particular, it describes mechanisms of how provenance information can be
572 located, assessed, and queried. Accessing provenance can be implemented either by dereferencing a
573 URI to actual provenance content or through a provenance query service in cases where the documented
574 object can not be associated with a URI. In both cases, only a link to provenance is provided. The
575 PROV-AQ specification also defines how the URIs are embedded into HTML/RDF objects documented
576 by provenance. The existing specification could be extended to directly support the CPM so that it would
577 define means to link both provenance and meta-provenance and could utilize the PIDs of the connectors.

578 RO-Crate (Soiland-Reyes et al., 2022) is a lightweight domain-agnostic approach to pack research
579 artifacts, their metadata, and relationships between them, and it serves as a shareable digital research
580 object. The format can be used to encapsulate a wide range of items that contributed to a research
581 outcome, such as data, scripts, configuration files, or provenance, together with metadata that describes
582 them and their relationships with the other data entities and with contextual entities such as authors or
583 organizations. RO-Crate profiles are a mechanism to specialize the general RO-Crate model for specific
584 domains, purposes, or use cases. The CPM RO-Crate profile (Wittner et al., 2023d) specifies how to
585 identify the CPM-compliant provenance files within an RO-Crate object, providing a means to define the
586 standardized representation of links from the object to respective provenance chain components and their
587 meta-provenance. The provenance can be stored either internally within the crate or externally and just
588 referenced from the RO-Crate object.

589 There are several data format examples (Zafar et al., 2017) that embed provenance information directly
590 in the data files. These include, for instance, Astronomy's Flexible Image Transport (FITS) format (Pence,
591 W. D. et al., 2010), which enables data lineage entries as part of their metadata headers (Simmhan et al.,
592 2005).

593 The genomic domain is another domain where big data sets are commonly generated through a
594 sequence of complex processes handled in distributed heterogeneous environments, and where adoption
595 of the CPM is currently envisioned. ISO/IEC 23092 series (Voges et al., 2021), commonly known as
596 MPEG-G, is an interoperable solution for the encoding, compression, and management of sequencing
597 data built on the widely established MPEG technology. MPEG-G defines a file format for storing data
598 and a transport format for data streaming. An MPEG-G file is structured in a file header and various
599 layers (i.e., data set group and data sets) down to one or more access units, holding the actual compressed
600 sequencing data. Provenance information is embedded into the data. To each data structure, indeed,
601 two types of optional metadata can be attached: information and protection metadata. While protection
602 metadata offers tools to manage the confidentiality and integrity of the information, information metadata
603 provides general information about the data, such as the origin of the biological sample, a log of the
604 operations carried out on the data, and information associated with the preparation of the samples and the
605 sequencing process. Normative extension mechanisms are also defined to expand the defined set of core
606 elements and include new attributes.

607 Besides digital research objects, physical objects such as biological or environmental specimens are
608 also commonly exchanged between organizations. Provenance information documenting the specimens
609 needs to cover all steps of the specimen life cycle from their collection to analysis, including data
610 originating from analytical procedures applied to a specimen (Ezzelle et al., 2008). A general prescription
611 of the sequence of individual steps in a laboratory is provided in Standard Operating Procedures (Manghani,
612 2011), and the actual provenance information describing the executed steps is provided in lab books.
613 Historically, lab books were analog (e.g., in paper form), but with the expansion of computer based
614 systems, there are currently many solutions that enable creation of the lab books in a digital form.
615 However, the electronic lab books have still not been widely adopted in academia, mainly due to costs,
616 complexity of use, accessibility issues related to various device types and operating systems (Kanza et al.,
617 2017a), or wide range of options which make the selection difficult and confusing (Kwok, 2018). The
618 electronic lab books are often unstructured and have the form of a generic note-taking software, such
619 as OneNote (<https://www.onenote.com/>) or Evernote (<https://evernote.com/>), though
620 dedicated solutions exist (e.g., eLabFTW (CARP et al., 2017)). However, the solutions are typically not
621 interoperable (Kanza et al., 2017a,b), which is partially caused by lack of standards between different
622 ELNs manufacturers (Kanza et al., 2017a).

623 **5 RESULTS**

624 In this section, we: 1) introduce a categorization of how provenance information and meta-provenance
625 can be handled with regard to exchanged research objects; 2) describe properties of provenance and meta-
626 provenance related to the categorization; 3) derive requirements on links between provenance components
627 and meta-provenance; 4) extend the CPM with the support for linking provenance components and
628 meta-provenance. Finally, we briefly describe the implementation of the concept for the digital pathology
629 use case and explain how the results were validated respect to the research questions defined in the
630 Introduction (Section 1).

631 **5.1 Provenance exchange schemes**

632 The categorization of how provenance information and meta-provenance can be handled with regard to
633 exchanged research objects consists of three provenance exchange schemes: attached, semi-attached and
634 detached (Fig. 9). The schemes loosely follow the semantics of the coupling schemes, and we describe
635 how our categorization relates to the existing one. Each of the introduced schemes exhibits distinct
636 properties of provenance, which might be appropriate for different application scenarios.

637 ***Attached Scheme***

638 In the attached scheme, provenance information and meta-provenance are part of the communication
639 between a sender and a receiver. In particular, when the object is exchanged between organizations, the
640 copy of the corresponding provenance and meta-provenance is also exchanged. This may be done either

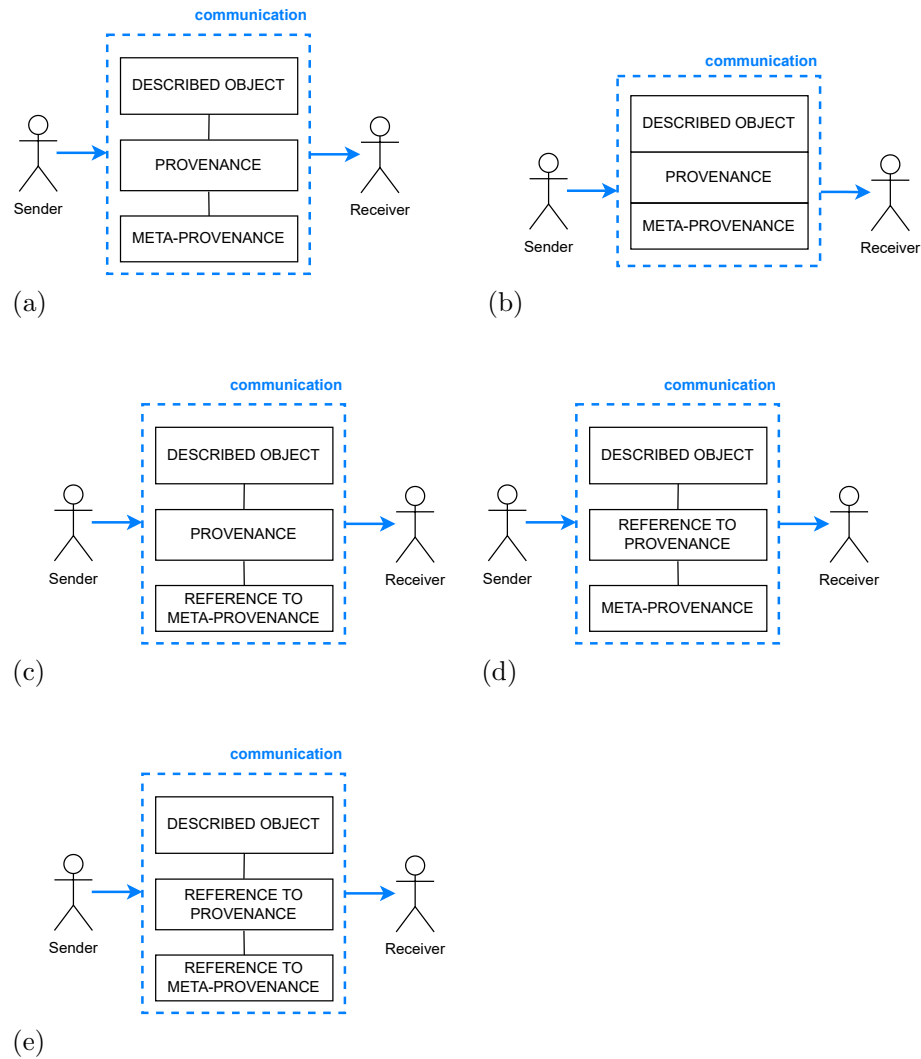


Figure 9. Illustration of the provenance exchange schemes: a) the attached scheme, where the provenance information and meta-information are not embedded within the exchanged object; b) the attached scheme, where the provenance information and meta-information are embedded within the exchanged object; c) the semi-attached scheme with a reference to meta-provenance; d) the semi-attached scheme with a reference to provenance; e) the detached scheme where both provenance and meta-provenance are referenced. The lines between squares express correspondence.

641 by providing it as a standalone piece of information outside the object (Fig. 9 (a)), or by embedding
 642 the information directly in the documented object (Fig. 9 (b)). As a result, when the process of the
 643 object exchange is finished, the receiver has an actual copy of the provenance and meta-provenance. For
 644 instance, FAIR digital objects encapsulating data with provenance or data formats that include provenance
 645 information in their header fall within this category.

646 **Semi-attached schemes**

647 In the semi-attached scheme, either provenance information or meta-provenance is not part of the
 648 communication between a sender and a receiver (only one is present). As a consequence, the receiver
 649 of the object has to make an additional request to get an actual copy of the missing part after the object
 650 exchange is finished. For instance, FAIR digital objects that encapsulate data with a reference to its
 651 provenance stored externally to the object fall within the semi-attached scheme.

652 1. **Semi-attached scheme with provenance attached:** Provenance information is part of the commu-
653 nication between a sender and a receiver, but the corresponding meta-provenance is not part of the
654 communication (Fig. 9 (c)).

655 2. **Semi-attached scheme with meta-provenance attached:** meta-provenance is part of the communi-
656 cation between a sender and a receiver, but provenance information is not part of the communication
657 (Fig. 9 (d)).

658 ***Detached scheme***

659 In the detached scheme, neither provenance nor meta-provenance is part of the communication between a
660 sender and a receiver (Fig. 9 (e)). As a consequence, the receiver of the object has to make an additional
661 request to get actual copies of both provenance and meta-provenance after the object exchange is finished.
662 For instance, FAIR digital objects that encapsulate data with a reference to its provenance and meta-
663 provenance stored externally to the object fall within the detached scheme (for an elaboration on possible
664 configurations of byte sequences and metadata references in FAIR Digital Objects we refer to (Lannom
665 et al., 2022)).

666 ***Properties of the schemes***

667 The available literature presents various properties of the provenance coupling schemes which can be
668 adopted for the presented provenance exchange schemes. These properties are amended and described in
669 the context of the provenance exchange schemes in Table 1.

670 Another important aspect is that the exchange schemes are defined with respect to a sender and a
671 receiver and communication between them. As a result, distinct parts of a distributed provenance chain
672 may correspond to different provenance exchange schemes.

673 ***Relating provenance exchange schemes and coupling schemes***

674 The main difference between the schemes is that while the provenance exchange schemes categorize
675 coupling of provenance with described objects while they are transferred, the provenance coupling
676 schemes categorize coupling of provenance with described objects while they are stored. Despite the
677 similarity between provenance exchange and coupling schemes, we have not identified any implicit
678 general relation between the schemes. In particular, none of the provenance coupling schemes generally
679 implies a specific exchange scheme and vice versa. The following example illustrates this conclusion.

680 Consider an object stored in conformance with tight coupling scheme – provenance (and meta-
681 provenance) are part of the object. This object can be generally sent according to all three exchange
682 schemes. In the first case, the object can be sent as it is, resulting in the attached scheme. In the second
683 case, both provenance and meta-provenance can be detached from the object. In this case, the detachment
684 of provenance and meta-provenance should be documented somewhere in provenance, as the "detached"
685 object differs from the originally stored tightly coupled object. There are two possible options to continue
686 with respect to the detached object: 1) The detached object is stored (according to the no-coupling or the
687 loose-coupling scheme) by the sender and sent to the receiver consequently; 2) the detached object is sent
688 to the receiver directly, without storing its detached version (so the last storage category of the object was
689 tight-coupling). In both cases, the provenance exchange scheme is determined by how the provenance and
690 meta-provenance is sent. If both provenance and meta-provenance are sent together with the object, this
691 conforms to the attached scheme. If only references are provided, this results in the detached scheme. If
692 only one of provenance or meta-provenance is provided directly, and only a reference is provided for the
693 other one, this results in the semi-attached scheme. This example shows how originally tightly coupled
694 object can be sent according to any of the three exchange schemes.

695 Consider the opposite now, having an object that is stored according to the loose-coupling or no-
696 coupling scheme. Depending on whether all three are sent to receiver together or only references to
697 provenance and meta-provenance are provided, this can result in all three provenance exchange schemes.
698 If the provenance (and meta-provenance) are attached to the object before it is sent, the attachment of
699 provenance and meta-provenance should be documented somewhere in provenance, as the "attached"
700 object differs from the originally stored loosely-/no- coupled object. If this version of the object is stored
701 by the sender before it is sent, we end up in the previous example. If the attached object is sent, this
702 results in the attached exchange scheme, despite having the object stored according to loose-/no- coupling
703 scheme originally.

704 This is summarized in Table 2.

Table 1. The table compares various provenance and meta-provenance-related properties in the attached, semi-attached, and the detached scheme.

Property of a scheme*	Attached scheme	Semi-attached and detached schemes
Accessibility	As the actual copy of provenance and meta-provenance** are immediately available to the receiver after the exchange of the object is finished, the scheme is less prone to accessibility errors in comparison with the detached schemes. The reason is that no additional query is needed to get the actual copies, and there is no need to maintain an additional reference to provenance or meta-provenance.	As the actual copy of provenance or meta-provenance is not immediately available to the receiver after the exchange of the object is finished, the scheme is more prone to accessibility errors than the attached scheme. The reason is that there is a need to make an additional request to get the actual copies, and a reference to the actual copy must be maintained.
Access control (Zafar et al., 2017)	Case-by-case access control may be difficult to achieve. Once the provenance and meta-provenance are attached to an exchanged object, each consecutive receiver will have access to it by design. Provenance encryption can be used to protect sensitive information, but this would introduce additional complexity related to encryption/decryption keys management and pose additional risks related to keys leakage or keys/scheme deprecation. The attached scheme does not allow for the separation of access control strategies for exchanged objects and provenance, as they are provided to a receiver together.	Case-by-case access control management is achievable. Once a receiver requests provenance or meta-provenance, the sender may decide the authorization result case-by-case and make individual decisions specific to each consecutive provenance receiver in the chain. The semi-attached and the detached scheme provide better support for the separation of access control strategies for exchanged objects and provenance, as they are exchanged between a sender and a receiver separately.
Distributed & heterogeneous environments (Cruz et al., 2009; Glavic et al., 2007)	For provenance information embedded within the object, the scheme requires a higher level of standardization than the semi-attached or the detached schemes, as different steps of a research pipeline must be able to deal with a single format of the exchanged object in which provenance is embedded.	The schemes require a lower level of standardization than the attached scheme, as provenance or meta-provenance are not embedded directly within the exchanged object.
Consistency (Muniswamy-Reddy et al., 2006; Zafar et al., 2017)	If provenance and meta-provenance are embedded directly within the exchanged object, they are less prone to accidental loss.	Provenance or meta-provenance is more prone to accidental loss than the attached scheme since the linked information can be corrupted, e.g., during backups, restoration, copies, etc.
Interoperability	If provenance and meta-provenance are embedded directly within the exchanged object, they must be stored in a standardized format to achieve interoperability between implementations.	Provenance or meta-provenance can be stored in an arbitrary format but must be provided to a receiver in a standardized format to enable its processing.
Size & Ease of Distribution	Since provenance and meta-provenance may be bigger than the exchanged object (e.g., for small data sets with very granular provenance descriptions), their inclusion inside the communication between a sender and a receiver may negatively affect the ease of their distribution.	The size of the communication is not affected by the corresponding provenance or meta-provenance size.
Non-repudiation (Trustworthiness (Zafar et al., 2017))	Since an exchanged object may pass through an untrusted environment, non-repudiation of provenance and meta-provenance would be practically unattainable.	The ability to achieve non-repudiation of provenance or meta-provenance is not directly affected by an untrusted environment through which an exchanged object is passed since it can be stored remotely in a secure environment.
Physical objects	Provenance and meta-provenance can not be part of physical objects like biological samples. For the description of physical objects, an attached scheme with the provenance information and meta-provenance outside the object, a semi-attached scheme, or the detached scheme applies.	The detached scheme can be used for the description of physical objects.

*Performance and scalability properties (Zafar et al., 2017) of queries over (de)coupled provenance information do not apply to provenance exchange schemes, as querying over provenance is unrelated to provenance exchange.

** Depending on which one is not included in the communication between a sender and a receiver. This comment applies to each “provenance or meta-provenance” phrase occurrence in this table.

Table 2. The table summarizing how coupling schemes relate to exchange schemes with respect to provenance. Any coupling scheme relates to a detached scheme where meta-information is just referenced.

Coupling scheme (storage)	Exchange scheme (transfer)	Description
Tight coupling	Attached scheme	The exchanged object is passed to a receiver as stored or is transformed to a different representation.
Tight coupling	Detached scheme	The exchanged object is separated from its provenance before it is passed to a receiver. A reference to the object’s provenance is provided.
Loose & no coupling	Attached scheme	A copy of the corresponding provenance information is passed to the receiver directly together with an exchanged object.
Loose & no coupling	Detached scheme	A reference to the corresponding provenance information is passed to the receiver together with an exchanged object.

705 Once the exchanged object and provenance and meta-provenance or a reference to it are received
706 and processed by a receiver, the receiver can decide where the provenance component corresponding
707 to its process will be stored. This can be done in principle according to one of the coupling schemes.
708 However, further aspects may affect the decision, such as requirements for the assignment of identifiers or
709 the exchanged object’s format. How these aspects affect the decision is described in the “Revision of the
710 Common Provenance Model appending strategies” section (Section 5.3).

711 5.2 Requirements on links between provenance and meta-provenance

712 In order to enable a provenance receiver to locate the “missing piece” in the semi-attached or the detached
713 scheme (provenance, meta-provenance, or both), a link to it must be provided. For the attached scheme, a
714 local link may be provided, e.g., referring to a part of the exchanged information intended for provenance or
715 meta-provenance representation. Standardized representation of links to provenance and meta-provenance
716 has the benefit of reducing the overhead of the underlying data formats otherwise necessary since format
717 developers do not necessarily need to develop their own format-specific links representation, as the
718 standardized representation would be available. Such standardized representation of links could be part
719 of the communication between a sender and a receiver, either as part of standardized provenance or
720 meta-provenance, embedded within the described object directly, or provided as a standalone piece of
721 information.

722 Each component of a provenance chain relevant to an exchanged object may be generated indepen-
723 dently by different organizations. If a detached scheme is used, each of the organizations should be
724 enabled to choose storage modalities for referenced provenance or meta-provenance. This is because the
725 enforcement of having prescribed provenance and meta-provenance storage for different components in
726 the chain – i.e., for distinct organizations – might be too restrictive, preventing the organizations from
727 adopting such a solution. Additionally, it would be very difficult to prescribe a centralized provenance and
728 meta-provenance storage for such a heterogeneous environment. Consequently, the underlying provenance
729 model should provide standardized means to include references to components of the provenance chain
730 from meta-provenance, each reference corresponding to a distinct part of the chain. Similarly, the under-
731 lying provenance model should provide standardized means to include references to meta-provenance
732 from provenance, each reference corresponding to a distinct part of the chain. These requirements are
733 summarized in Table 3.

734 Applying these requirements to the computational steps of the running example, each provenance
735 component – namely for the data preprocessing, AI model training, and AI model evaluation – is enabled
736 to link to its own meta-provenance by its own reference (Fig. 10 a)). In other words, if each of the steps
737 will be executed by a different organization, the requirements impose that the organizations would not be
738 forced (but are free) to use a shared meta-provenance location. Conversely, a reference to corresponding
739 provenance components should be enabled in each meta-provenance bundle. If the meta-provenance
740 bundle would be shared for all three provenance chain components, it is allowed to include three different

741 references in meta-provenance, for each corresponding provenance component individually (Fig. 10 b)).

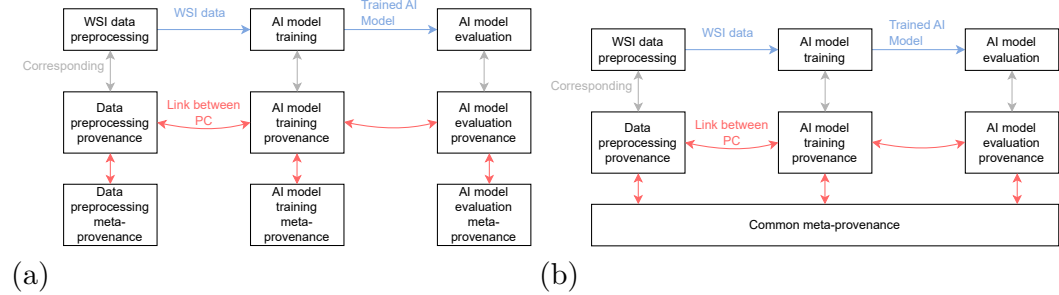


Figure 10. Illustration of how provenance components are enabled to link corresponding meta-provenance.

742 Some parts of the underlying provenance chain may have multiple versions, e.g., due to detected
 743 errors in provenance (Cheah and Plale, 2014, 2012), which have been later corrected (see the Background
 744 section – Section 2 – for the versioning mechanism description). If a new provenance component
 745 is created, two actions are required with regard to meta-provenance: 1) recording information about the
 746 new version into meta-provenance; 2) relating the new record with records about the previous versions. In
 747 this case, the corresponding meta-provenance related to a particular provenance component should not be
 748 fragmented (Figure 11) – so it can be referenced by a single reference – since the additional complexity
 749 potentially introduced by enabling the fragmentation of the meta-provenance outweighs potential benefits.
 750 An advantage of such functionality is that it provides more flexibility for implementers to decide where
 751 to store the new meta-provenance record. On the other hand, this could cause meta-provenance for
 752 different versions of the same provenance component to be fragmented into multiple storage locations and
 753 formats, which would overcomplicate meta-provenance generation process and representation. In addition,
 754 consistency and continuity of such fragmented meta-provenance should be achieved and verifiable.

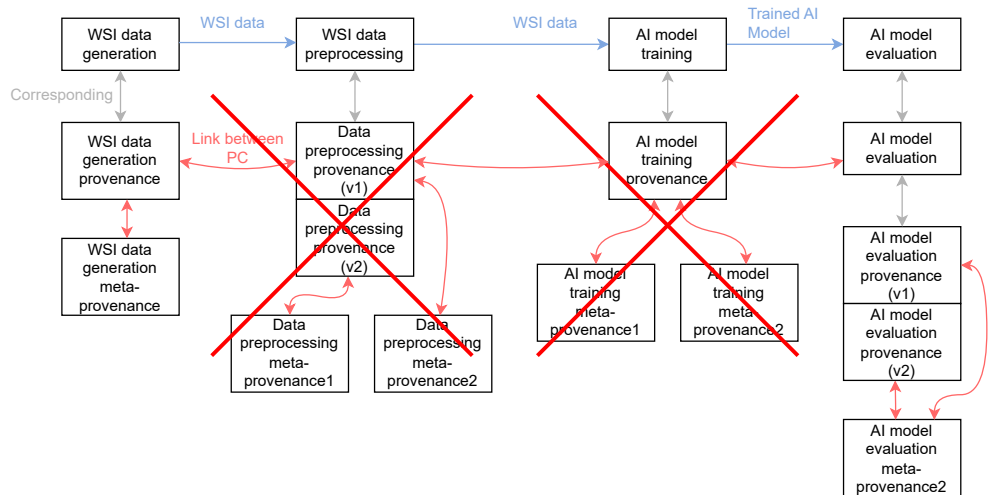


Figure 11. Illustration of how different versions of provenance components are enabled and not enabled to link corresponding meta-provenance. Meta-provenance of various versions of a provenance component should not be fragmented, and meta-provenance of a provenance component should not be fragmented.

755 **5.3 Revision of the Common Provenance Model appending strategies**

756 As was argued in the previous sections, bi-directional links between provenance components and corre-
 757 sponding meta-provenance should be supported in underlying provenance models. Additionally, meta-
 758 provenance corresponding to a provenance component should not be fragmented and potentially stored

Table 3. A summary of the derived requirements on the underlying provenance models related to the standardized representation of links between provenance and meta-provenance.

Derived requirements related to provenance and meta-provenance linking.	
1.	A standard way to represent links between provenance and meta-provenance.
2.	A standard way to include references to meta-provenance in the provenance and vice versa. Different components may have different references (Fig. 10).
3.	Avoid fragmentation of meta-provenance related to a provenance component. A single reference to meta-provenance is enabled per provenance component (Fig. 11).
4.	Avoid fragmentation of meta-provenance related to different versions of a provenance component. Only a single reference to meta-provenance is enabled for all the component versions (Fig. 11).

759 in distinct locations, so multiple versions of the same provenance component can link to a single meta-
760 provenance component. As a consequence, an organization creating a new version of a component
761 in a provenance chain should be provided with means to update the corresponding meta-provenance
762 component.

763 In the attached scheme, the two general appending strategies, i.e., versioning or adding provenance
764 components in the provenance chain, do not differ significantly. If an object and related meta-provenance
765 are meant to “travel together” (e.g., as part of the data file header), then a receiver of the object can always
766 re-write it, with no respect to which of the two appending strategies is used. On the other hand, if any
767 of the information is referenced in a semi-attached or the detached scheme, the properties of the two
768 appending strategies vary. If a new version of a provenance component is created in the chain, a receiver
769 must be able to append the new versioning information to the corresponding meta-provenance. On the
770 other hand, if a receiver adds a new provenance component to the provenance chain, he is not bound to
771 use any specific meta-provenance bundle.

772 One of the most important questions is whether a receiver of an object should even be able to create
773 a new version of an existing component that was created by another organization. Since provenance
774 components document part of a research object life cycle when a particular organization handles it, we
775 propose that this documentation should not be outdated by another organization, and suggest enabling the
776 creation of new versions by different organizations (different from the organization that originally created
777 the finalized provenance component) only in justified cases, e.g., when an organization ceases to exist and
778 an error in provenance is detected later. In this situation, it would be beneficial if another organization,
779 e.g., an authority, could create a corrected version. As a result, the versioning mechanism should not be
780 used to append new information to a chain without additional integrity assurances when an object passes
781 across organizational boundaries. The assurance must guarantee that the new version of the component
782 only appended a content, and that the original provenance content was not modified.

783 Another aspect that affects the determination of appending strategy is the intended usage of the
784 formats of exchanged objects. For instance, genomic data and their metadata is part of the MPEG-G
785 standard (Voges et al., 2021), which currently covers documentation of steps starting from raw sequence
786 reads up to their alignment to a reference sequence. Each time a new dataset is derived from an MPEG-G
787 file (e.g., a raw genomic dataset is stored as an MPEG-G file, and another dataset with aligned reads
788 will be derived from that raw genomic dataset), it is expected that the derived dataset is represented as
789 new MPEG-G file. In this scenario, the new file is a new object that should be documented in a new
790 provenance component.

791 The determination of provenance appending strategy might also be affected by the assignment
792 of identifiers for documented exchanged objects. For example, Zenodo (European Organization For
793 Nuclear Research and OpenAIRE, 2013), a popular open repository for storing digital research objects,
794 distinguishes identifiers for the objects themselves and for their specific versions. Using such identifiers
795 for the objects might indicate situations when appending a new provenance component is more appropriate
796 over the provenance versioning mechanism (or vice versa). In particular, when a derived object is assigned
797 the new object identifier, creating a new provenance component in a chain intuitively seem to be the
798 preferred version. On the contrary, if the described object is assigned an identifier of a new version of
799 another existing object, creating a new version of a provenance component may be the preferred option.

800 **5.4 Extending the CPM with the links between provenance and meta-provenance**

801 Distributed provenance chains are based on a provenance backbone (Section 2). However, the CPM does
802 not provide a description of how to create links between the provenance backbone and meta-provenance.
803 To achieve this, we propose the following mechanism based on the existing provenance backbone
804 structures and PIDs.

805 ***Linking from a provenance component to meta-provenance***

806 The provenance backbone contains three types of PROV entities: forward connector, backward connector,
807 and current connector. We suggest that each of these is identified with a PID within provenance information
808 and that the PIDs resolve to the following information:

- 809 1. The corresponding entity represented by the PID in any serialization (might be subject to a content
810 negotiation protocol);
- 811 2. Identifier of a corresponding provenance component that contains that entity:
 - 812 (a) For a backward connector, it is the identifier of the preceding component in the chain.
 - 813 (b) For a current connector, it is an identifier of the “current” provenance component.
 - 814 (c) For a forward connector, it is the identifier of the consecutive provenance component in the
815 chain.
- 816 3. Identifier of meta-provenance, where the provenance of the corresponding provenance component
817 is present:
 - 818 (a) For a backward connector, it is the identifier of the meta-provenance component correspond-
819 ing to the preceding component in the provenance chain.
 - 820 (b) For a current connector, it is an identifier of the meta-provenance component corresponding
821 to the “current” provenance component.
 - 822 (c) For a forward connector, it is the identifier of the meta-provenance component corresponding
823 to the consecutive provenance component in the chain.

824 As the forward connector and the corresponding backward connector are identified with the same
825 identifier (required by the CPM), the corresponding PID resolves to both provenance components identi-
826 fiers in which they are present (Fig. 12, green arrows). Consequently, the information the PID resolves to
827 must be updated each time a new component containing the forward/backward connector is added to the
828 provenance chain.

829 Following the general mechanism for provenance access described in the PROV-AQ (Klyne et al., 2013)
830 specification, we suggest that the provenance components identifiers and meta-provenance identifiers are
831 resolvable and that these resolve to particular content (presuming appropriate authorization), as depicted in
832 Fig. 12 (red arrows). As a result, the proposed mechanism implements links from provenance components
833 to the corresponding meta-provenance content through a pair of resolvable identifiers. Adopting the
834 PIDs for provenance components and meta-provenance is unnecessary since these identifiers are part of a
835 particular connector PID resolution.

836 Further details about the selected format and other properties of the information which connector
837 PIDs resolve to and new connector attributes to support the proposed functionality are described in the
838 supplementary file "Supplemental Article 1.pdf".

839 ***Linking from meta-provenance to provenance***

840 Each provenance component is identified in meta-provenance and is represented as a PROV entity. As
841 we suggested earlier, the provenance components and meta-provenance identifiers resolve to their actual
842 content. Consequently, the link from meta-provenance to corresponding provenance components is
843 established by the inclusion of that identifier in meta-provenance (see Section 2 for further details on how
844 specific component’s versions are expressed in meta-provenance).

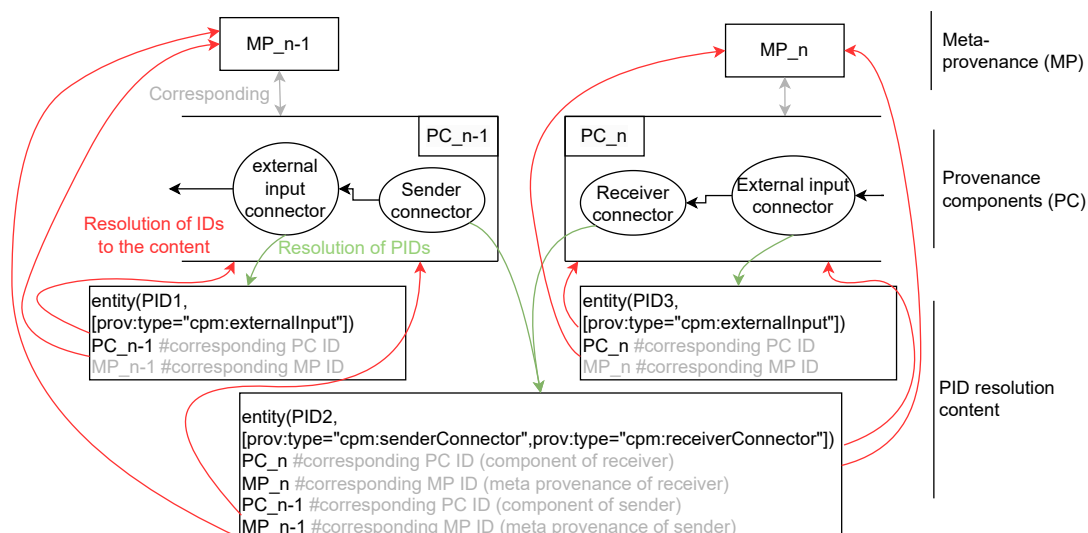


Figure 12. Schema of the connector PIDs resolution. Each of the PIDs resolves to the defined set of information: serialization of corresponding provenance structure and, depending on the connector type, identifiers of corresponding provenance components and meta-provenance. If these are resolvable, the connector PIDs can be used as a starting point to navigate to the actual content of respective provenance components and meta-provenance. Since the CPM requires that the respective sender and backward connector share the same identifier for the underlying provenance structure, the corresponding PID must resolve to the same information on the side of the sender's and receiver's finalized provenance.

845 5.5 Implementation

846 We have implemented the proposed mechanism for provenance and meta-provenance linking for two
 847 use cases: 1) the digital pathology use case; 2) the ColoRectal Cancer (CRC) cohort extension use
 848 case. As a result, the use cases are documented using multiple provenance components, and these
 849 components form two provenance chains – a chain for each use case. The provenance finalization scripts
 850 also generate corresponding meta-provenance files that are interlinked with the generated provenance
 851 chains. The PIDs are realised using Digital Object Identifiers (doi.org) generated by the DataCite
 852 (<https://datacite.org/>) registration agency. The associated code, inputs, resulting provenance,
 853 meta-provenance, and PID-resolved information, are either attached as supplementary materials and
 854 hosted on various Gitlab repositories according to the organization from which the provenance is coming.
 855 The research objects documented by the provenance chains are deposited in the Zenodo (European
 856 Organization For Nuclear Research and OpenAIRE, 2013) repository.

857 See the supplementary files for a detailed description of the use cases and the implementation, and
 858 the "Availability of Supporting Data and Materials" section for references to the code and related digital
 859 objects.

860 5.6 Validation of results

861 The proposed extension of the CPM was validated by implementation of an algorithm that traverses a
 862 provenance chain and processes part of the meta-provenance to demonstrate that the algorithm can exploit
 863 the links between provenance components and their meta-provenance. The algorithm was run on the two
 864 provenance chains presented in the Implementation section (Section 5.5).

865 The traversal algorithm is implemented using Java code, building on the top of the ProvTool-
 866 box (Moreau, 2016) library for the W3C PROV data model. Our algorithm performs a single task –
 867 it retrieves identifiers of precursors of a given object expressed on a provenance backbone and for each
 868 component of the chain, it check its integrity. In particular, the algorithm performs the following steps:

- 869 1. The algorithm takes two inputs: the identifier of a connector, for which the precursors are requested,
 870 and the identifier of a provenance component where the traversal is supposed to start.
- 871 2. As the component identifier is dereferencable, the algorithm uses the ID to retrieve the component's

- 872 content and finds the corresponding entity.
- 873 3. Starting from the node identified with the given entity ID, the algorithm traverses the provenance
874 backbone to find the identifiers of the precursors expressed in particular provenance component,
875 and to find backward connectors – links to preceding components of the chain.
 - 876 4. If a backward connector is found, its DOI is resolved to retrieve information about the referenced
877 component and its meta-provenance.
 - 878 5. Integrity of the referenced provenance component is checked by comparing the bundle’s hash stored
879 in meta-provenance and the hash calculated from the the referenced bundle content.
 - 880 6. If the hash verification succeeds, the algorithm runs recursively with the request for precursors of
881 the connector in the referenced bundle (Step 1).

882 Executing the algorithm on the generated provenance chains proves that the algorithm can successfully
883 find precursors of an object in the distributed multi-organizational provenance, demonstrating that the
884 proposed mechanism for linking provenance components and their meta-provenance is feasible.

885 See the supplementary files for more detailed description of the validation procedure, and the "Avail-
886 ability of Supporting Data and Materials" section for references to the code and related digital objects.

887 **6 DISCUSSION AND CONCLUSION**

888 The results presented in this work have a practical impact on the current state-of-the-art in the provenance
889 information domain. The conducted analysis and associated general requirements derivation are applicable
890 to a wide range of provenance models. While exploring possibilities for how the links between provenance
891 and meta-provenance can be designed, we have achieved specific architectural decisions related to
892 the matter, which enabled us to extend the current state-of-the-art provenance information model for
893 distributed multi-organizational provenance with new features. As several authors of this work are leading
894 and contributing to the development of ISO 23494 standard series (Wittner et al., 2023c), the presented
895 results will be integrated with the draft proposal of the standardized provenance model. Consequently, this
896 work is another step in global provenance information standardization. Without the proposed extension
897 for the standardized links representation and associated PIDs resolution presented in this work, adopters
898 of the provenance model would have to design the links by themselves, which can lead to incompatible
899 solutions. This has already been witnessed with the existing W3C PROV standard, which is too flexible
900 to enable unified traversal through distributed provenance chains (Wittner et al., 2022; Wittner, 2022).

901 Adoption of the CPM includes aspects that must be taken into consideration – i.e., determination
902 of specific semantics of the connectors, the granularity of provenance descriptions, and a level of
903 collaboration between organizations that handle the exchanged object (Wittner et al., 2022). In addition,
904 choosing an appropriate provenance exchange method depends on the specific use case. This work
905 describes the general properties of the provenance exchange schemes. However, the properties of the
906 resulting provenance chain are determined by the combination of all provenance exchange schemes
907 between different organizations in a chain and the application of provenance coupling schemes for
908 provenance storage within each organization. For example, if the whole chain adopts the attached
909 provenance exchange scheme and tight coupling of the object with provenance – e.g., all the provenance
910 information and documented object is present in an exchanged RO-Crate, which is iteratively appended –
911 each consecutive organization will have access to it. On the other hand, if a detached scheme is used for a
912 single segment of this chain, an authorized receiver can access related provenance through a reference and
913 can distribute this provenance on its own, providing access to other organizations in the chain similarly to
914 the attached scheme. This bottleneck can not be simply prevented by architectural decisions but must
915 be addressed, e.g., by contractual agreements between organizations involved in distributed provenance
916 handling.

917 In the case of the detached scheme, neither standardized provenance nor standardized meta-provenance
918 is part of the communication between a sender and a receiver. In this scheme, how the references are
919 designed and represented is within the constituency of the exchanged information format or commu-
920 nication protocol between a sender and a receiver. The analyses presented in this work may serve as
921 a starting point to design representations of references in the communication. They can be potentially

922 reused when designing interlinking provenance and meta-provenance outside the standardized provenance
923 information. For instance, if a given organization provides objects via /objects/<ID> in a REST API, it
924 could provide (meta)provenance via /(meta)provenance/<ID>. Another option could be to point to the
925 object, provenance link, and meta-provenance link from appropriate fields in the JSON object returned
926 by the API endpoint, which would serve as a higher-level wrapper around both the object and the CPM
927 artifacts.

928 One of the features of the proposed mechanism for linking from provenance to meta-provenance is that
929 the link is implemented only in cases when at least one connector is present in a provenance component.
930 On the other hand, if a connector is not present in a component, meta-provenance corresponding to the
931 component would not be linked. However, this case is irrelevant to our work because a component without
932 a connector is not part of a provenance chain, and these are not the subject of the CPM.

933 The resolvability of provenance component identifiers is an additional requirement related to prove-
934 nance chain components implementation. For instance, PROV bundles serialized into files can be made
935 directly accessible (with authorization) using web servers. However, implementing the components
936 using a graph database would require identifying the graph within the database and its extraction when
937 requested (Klyne et al., 2013).

938 Literature surveys and systematic literature reviews are important in categorizing different approaches,
939 harmonizing related terminology and perspectives, and defining future research directions. Importantly,
940 they often provide a consolidated overview of heterogeneous requirements, which are originally frag-
941 mented across the literature. This plays an important role since such a centralized and harmonized source
942 of requirements may be used when designing a solution for provenance, so designers can make informed
943 decisions when selecting or designing a provenance solution (Freire et al., 2008). However, the review
944 of approaches for coupling described objects, provenance, and meta-provenance has limitations. The
945 Systematic Literature Review (Pérez et al., 2018), the starting point for our related work survey, has
946 identified 251 published papers related to existing provenance systems, many of which are centralized
947 solutions, and only eight systems were categorized as solutions for distributed provenance. Among
948 these, as it was described in the Related work section (Section 4), the vast majority of them works as a
949 centralized solution that collects provenance from distributed and possibly heterogeneous environments.
950 The PLUS system is the most relevant, because it has tried to address distributed multi-organizational
951 provenance as it is understood in this work. However, after publishing the PLUS system as an initial work
952 addressing multi-organizational provenance, the further development in this direction did not continue,
953 and this is the main reason for bounding our literature review. In particular, the attention to research
954 objects exchange and multi-organizational provenance was brought only in recent years again, so we do
955 not presume that less recent work would have conducted such an in-depth analysis as presented in this
956 work for the current context. For this reason, we have decided to build the analysis mainly on available
957 taxonomies, literature surveys, and selected systems that are intended to work in distributed environments,
958 instead of reviewing all existing provenance systems.

959 **6.1 Future Work**

960 Our work can be divided into two main branches – development of the CPM to support various functional
961 and non-functional requirements, such as provenance authenticity or non-repudiation assurances, and
962 continuous validation of the model on various use cases.

963 Now, when a representation for the standardized links between provenance and meta-provenance com-
964 ponents is defined, we will continue with definitions of standardized representation for domain-agnostic
965 information included in meta-provenance. Similarly to the representation of provenance components
966 versioning and master bundles, we will work on the representation of security-related aspects of prove-
967 nance, namely authenticity, integrity, and non-repudiation. To achieve this, we will integrate the current
968 CPM with our previous work on provenance non-repudiation in the context of clinical decision support
969 systems (Fairweather et al., 2021). This direction aims to enable provenance chain traversal with support
970 for unified meta-provenance-related queries resolution. Examples of such queries are “Is the given
971 provenance component of a chain authentic, e.g., was it generated by the claimed organization?” or “Was
972 the given research object created as a result of an unreliable process, for which a trustworthy provenance
973 information component is not available?”. The results in this direction will be proposed to become an
974 input for the *ISO 23494-6 Biotechnology – Provenance Information Model for Biological Material and*
975 *Data – Part 6: Security Extensions*.

976 For the purpose of the model validation, we will apply it to document a wide range of use cases coming
977 from the life sciences, including optical microscopy experiments, genomic data compression, biological
978 samples handling, or computational workflow based experiments. The model will be provided to the ISO
979 23494 TC276 Biotechnology WG5 as a groundwork for the development of domain-specific provenance
980 standard parts, namely ISO 23494-3 (biological material provenance), ISO 23494-4 (data generation
981 provenance), and ISO 23494-5 (computational workflows provenance). The further development of the
982 CPM will be coordinated with the development of the MPEG-G standard (Voges et al., 2021) under
983 an ad hoc group under JTC1/SC29/WG08. The model is already being adopted in the BY-COVID
984 project (<https://by-covid.org/>), where it is being integrated with the Process Run Crate profile
985 specification (The Workflow Run RO-Crate working group, 2023) and several use cases related to
986 infectious disease. In this context, we await motivation for several possible extensions of the CPM.
987 These may include simplification of the provenance backbone structure, limitations of the number of
988 allowed connectors per provenance component, or definitions of additional types of connectors to support
989 advanced methods of provenance chain traversal.

990 **ADDITIONAL INFORMATION AND DECLARATIONS**

991 **Acknowledgements**

992 Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140)
993 supported by the Ministry of Education, Youth and Sports of the Czech Republic. We would like to thank
994 to MUDr. Rudolf Nenutil, CSc for providing us with an anonymized dump of records from the hospital
995 and laboratory information systems, which were used in the implementation for the digital pathology use
996 case. The implementation of the traversal algorithm is based on bachelor thesis of Tomáš Zobač (led by
997 R.W., to be defended at the Faculty of Informatics, Masaryk University, early in 2024).

998 This manuscript was previously published as a preprint (Wittner et al., 2023a).

999 **Authors' Contributions**

1000 R.W. is the main author of the presented work and has carried out the conceptualization, investigation,
1001 methodology, visualization, adoption of the code for provenance chains traversal, and writing the original
1002 draft and the supplementary files for the digital pathology use case and for the validation description. F.F.
1003 is the main author of the supplementary file for the (CRC) cohort extension use case. M.G., L.P., S.L.,
1004 C.M., F.F., M.P., S.S.-R., H.M., J.G., P.H. provided feedback to refine the proposed concepts or their
1005 implementation, and wrote and edited the manuscript and the supplementary files. M.G. and R.W. are
1006 the main authors of the implementation part for the digital pathology use case. F.F, S.L. and L.P. are the
1007 main authors of the implementation part for the CRC cohort extension use case. J.G., H.M., and P.H.
1008 supervised the work.

1009 **Availability of Supporting Data and Materials**

1010 **The digital pathology use case.** In this paper, we use representative images downloaded from the
1011 Camelyon16 dataset (Litjens et al., 2018), as using the real images from the laboratory information
1012 system poses privacy risks for regarded patients and donors. The representative images were linked to the
1013 anonymized data from the information systems by changing their identifier according to the pathology
1014 department internal conventions. More technical details about the implementation for the digital pathology
1015 use case are provided in the "Supplemental Article 1.pdf". The resulting AI model is publicly available in
1016 Zenodo. repository (Wittner et al., 2023b).

1017 **The ColoRectal Cancer-Cohort Cloudification use case.** The technical details about the imple-
1018 mentation for the CRC cohort extension use case are provided in the "Supplemental Article 2.pdf". The
1019 image considered to demonstrate the workflow-based conversion has been taken from a public repository
1020 (Collection: Cancer Moonshot Biobank (CMB-PCA) Slide ID: MSB-02917-01-02), as using one of the
1021 real images poses privacy risks for the donors. The resulting RO-Crate documenting the conversion is
1022 publicly available in Zenodo repository (Leo and Pireddu, 2023).

1023 **Validation of the results.** The technical details about the results validation procedure are provided in
1024 the "Supplemental Article 3.pdf".

1025 All the code used to generate finalized provenance information is attached as a supplementary material.

REFERENCES

- 1027 Ahmad, R., Jung, E., de Senne Garcia, C., Irshad, H., and Gehani, A. (2020). Discrepancy detection in
1028 whole network provenance. In *12th International Workshop on Theory and Practice of Provenance*
1029 (*TaPP 2020*). USENIX Association.
- 1030 Allen, M. D., Blaustein, B., Seligman, L. J., and Chapman, A. P. (2012). Capturing provenance data
1031 within heterogeneous distributed communications systems. US Patent App. 12/917,891.
- 1032 Allen, M. D., Chapman, A., Blaustein, B., and Seligman, L. (2011). Getting it together: enabling
1033 multi-organization provenance exchange. In *3rd USENIX Workshop on the Theory and Practice of*
1034 *Provenance (TaPP 11)*.
- 1035 Alur, R., Benedikt, M., Etessami, K., Godefroid, P., Reps, T., and Yannakakis, M. (2005). Analysis of
1036 recursive state machines. *ACM Trans. Program. Lang. Syst.*, 27(4):786–818.
- 1037 Ametepe, W., Wang, C., Ocansey, S. K., Li, X., and Hussain, F. (2021). Data provenance collection and
1038 security in a distributed environment: a survey. *International Journal of Computers and Applications*,
1039 43(1):11–25.
- 1040 Amsterdamer, Y., Davidson, S. B., Deutch, D., Milo, T., Stoyanovich, J., and Tannen, V. (2011). Putting
1041 lipstick on pig: Enabling database-style workflow provenance. *Proc. VLDB Endow.*, 5(4):346–357.
- 1042 Bai, B., Nazir, S., Bai, Y., and Anees, A. (2021). Security and provenance for internet of health things:
1043 A systematic literature review. *Journal of Software: Evolution and Process*, 33(5):e2335. e2335
1044 JSME-20-0163.R3.
- 1045 Bao, Z., Davidson, S. B., and Milo, T. (2012). Labeling workflow views with fine-grained dependencies.
1046 *Proc. VLDB Endow.*, 5(11):1208–1219.
- 1047 Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research.
1048 *Nature*, 483(7391):531–3.
- 1049 Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science. *Circulation Research*, 116(1):116–
1050 126.
- 1051 Belhajjame, K., B’Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T.,
1052 McCusker, J., Miles, S., Meyers, J., Sahoo, S., and Tilmes, C. (2013). PROV-DM: The PROV data
1053 model. *W3C Recommendation*.
- 1054 Benson, E. E., Harding, K., and Mackenzie-dodds, J. (2016). A new quality management perspective
1055 for biodiversity conservation and research: Investigating Biospecimen Reporting for Improved Study
1056 Quality (BRISQ) and the Standard PRE-analytical Code (SPREC) using Natural History Museum and
1057 culture collections as case studies. *Systematics and Biodiversity*, 14(6):525–547.
- 1058 Blaustein, B., Seligman, L., Morse, M., Allen, M. D., and Rosenthal, A. (2008). Plus: Synthesizing
1059 privacy, lineage, uncertainty and security. In *2008 IEEE 24th International Conference on Data*
1060 *Engineering Workshop*, pages 242–245.
- 1061 Buneman, P., Caro, A., and Murray-Rust, D. (2016). Composition and substitution in provenance and
1062 workflows. In *8th USENIX Workshop on the Theory and Practice of Provenance (TaPP 16)*, Washington,
1063 D.C. USENIX Association.
- 1064 Buneman, P., Chapman, A., and Cheney, J. (2006). Provenance management in curated databases. In
1065 *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD
1066 ’06, page 539–550, New York, NY, USA. Association for Computing Machinery.
- 1067 Buneman, P. and Davidson, S. B. (2010). Data provenance—the foundation of data quality. In *Workshop:*
1068 *Issues and Opportunities for Improving the Quality and Use of Data within the DoD, Arlington, USA*,
1069 pages 26–28.
- 1070 Buneman, P., Gascon Caro, A., Moreau, L., and Murray-Rust, D. (2017). Provenance composition in
1071 prov. Workingpaper.
- 1072 Byrne, J. A., Grima, N., Capes-Davis, A., and Labbé, C. (2019). The Possibility of Systematic Re-
1073 search Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Potential Solutions.
1074 *Biomarker Insights*, 14.
- 1075 CARP, N., Minges, A., and Piel, M. (2017). elabftw: An open source laboratory notebook for research
1076 labs. *J. Open Source Softw.*, 2(12):146.
- 1077 Chaplin, S. (2012). Research misconduct: how bad is it and what can be done? *Future Prescriber*,
1078 13(1):5–76.
- 1079 Chapman, A., Allen, M. D., Blaustein, B., Seligman, L., Wolf, C., Morse, M., and Rosenthal, A. (2010).
1080 Plus: Provenance for life, the universe and stuff. In *VLDB*, volume 10, pages 13–17.

- 1081 Chapman, A., Blaustein, B. T., Seligman, L., and Allen, M. D. (2011). Plus: A provenance manager for
1082 integrated information. In *2011 IEEE International Conference on Information Reuse & Integration*,
1083 pages 269–275.
- 1084 Cheah, Y.-W. and Plale, B. (2012). Provenance analysis: Towards quality provenance. In *2012 IEEE 8th*
1085 *International Conference on E-Science*, pages 1–8.
- 1086 Cheah, Y.-W. and Plale, B. (2014). Provenance quality assessment methodology and framework. *J. Data*
1087 *and Information Quality*, 5(3).
- 1088 Ciccacese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A. J., Goble, C., and Clark, T. (2013). Pav
1089 ontology: provenance, authoring and versioning. *Journal of Biomedical Semantics*, 4(1):37.
- 1090 Clifford, B., Foster, I., Voekler, J.-S., Wilde, M., and Zhao, Y. (2008). Tracking provenance in a virtual
1091 data grid. *Concurrency and Computation: Practice and Experience*, 20(5):565–575.
- 1092 Cruz, S. M. S. d., Campos, M. L. M., and Mattoso, M. (2009). Towards a taxonomy of provenance in
1093 scientific workflow management systems. In *2009 Congress on Services - I*, pages 259–266.
- 1094 Cui, Y. and Widom, J. (2000). Practical lineage tracing in data warehouses. In *Proceedings of 16th*
1095 *International Conference on Data Engineering (Cat. No.00CB37073)*, pages 367–378.
- 1096 Curcin, V., Miles, S., Danger, R., Chen, Y., Bache, R., and Taweel, A. (2014). Implementing interoperable
1097 provenance in biomedical research. *Future Generation Computer Systems*, 34:1–16. Special Section:
1098 Distributed Solutions for Ubiquitous Computing and Ambient Intelligence.
- 1099 David Allen, M., Chapman, A., and Blaustein, B. (2015). Engineering choices for open world provenance.
1100 In Ludäscher, B. and Plale, B., editors, *Provenance and Annotation of Data and Processes*, pages
1101 242–253, Cham. Springer International Publishing.
- 1102 De Smedt, K., Koureas, D., and Wittenburg, P. (2020). FAIR Digital Objects for science: From data
1103 pieces to actionable knowledge units. *Publications*, 8(2).
- 1104 Dean, J. and Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Commun.*
1105 *ACM*, 51(1):107–113.
- 1106 Dogan, G. (2016). A survey of provenance in wireless sensor networks. *Adhoc & Sensor Wireless*
1107 *Networks*, 30.
- 1108 Drewes, F., Kreowski, H.-J., and Habel, A. (1997). Hyperedge replacement graph grammars. In *Handbook*
1109 *of Graph Grammars and Computing by Graph Transformation*, pages 95–162.
- 1110 European Organization For Nuclear Research and OpenAIRE (2013). Zenodo.
- 1111 Ezzelle, J., Rodriguez-Chavez, I., Darden, J., Stirewalt, M., Kunwar, N., Hitchcock, R., Walter, T., and
1112 D’souza, M. (2008). Guidelines on good clinical laboratory practice: bridging operations between
1113 research and clinical research laboratories. *Journal of pharmaceutical and biomedical analysis*,
1114 46(1):18–29.
- 1115 Fairweather, E., Wittner, R., Chapman, M., Holub, P., and Curcin, V. (2021). Non-repudiable provenance
1116 for clinical decision support systems. In Glavic, B., Braganholo, V., and Koop, D., editors, *Provenance*
1117 *and Annotation of Data and Processes*, pages 165–182, Cham. Springer International Publishing.
- 1118 Foster, I., Vockler, J., Wilde, M., and Zhao, Y. (2002). Chimera: a virtual data system for representing,
1119 querying, and automating data derivation. In *Proceedings 14th International Conference on Scientific*
1120 *and Statistical Database Management*, pages 37–46.
- 1121 Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The Economics of Reproducibility in
1122 Preclinical Research. *PLOS Biology*, 13(6):1–9.
- 1123 Freedman, L. P. and Inglese, J. (2014). The Increasing Urgency for Standards in Basic Biologic Research.
1124 *Cancer Research*, 74(15):4024–4029.
- 1125 Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: A survey.
1126 *Computing in Science & Engineering*, 10(3):11–21.
- 1127 Gierend, K., Krüger, F., Waltemath, D., Fünfgeld, M., Ganslandt, T., and Zeleke, A. A. (2021). Approaches
1128 and criteria for provenance in biomedical data sets and workflows: Protocol for a scoping review. *JMIR*
1129 *Res Protoc*, 10(11):e31750.
- 1130 Glavic, B., Dittrich, K. R., Kemper, A., Schöning, H., Rose, T., Jarke, M., Seidl, T., Quix, C., and
1131 Brochhaus, C. (2007). Data provenance: a categorization of existing approaches. *BTW’07: Daten-*
1132 *bankssysteme in Buisness, Technologie und Web*, (103):227–241.
- 1133 Green, T. J., Karvounarakis, G., Ives, Z. G., and Tannen, V. (2007). Update exchange with mappings and
1134 provenance.
- 1135 Groth, P. and Moreau, L. (2013). PROV-overview. *W3C Working Group Note*.

- 1136 Hellström, M., Heughebaert, A., Kotarski, R., Manghi, P., Matthews, B., Ritz, R., Sparre Conrad,
1137 A., Valle, M., Weigel, T., and Wittenburg, P. (2020). A persistent identifier (PID) policy for the
1138 european open science cloud (eosc). [https://op.europa.eu/en/publication-detail/
1139 -/publication/35c5ca10-1417-11eb-b57e-01aa75ed71a1](https://op.europa.eu/en/publication-detail/-/publication/35c5ca10-1417-11eb-b57e-01aa75ed71a1).
- 1140 Herschel, M., Diestelkämper, R., and Ben Lahmar, H. (2017). A survey on provenance: What for? what
1141 form? what from? *The VLDB Journal*, 26(6):881–906.
- 1142 Holub, P., Kohlmayer, F., Prasser, F., Mayrhofer, M. T., Schlünder, I., Martin, G. M., Casati, S., Koumakis,
1143 L., Wutte, A., Kozera, Ł., Strapagiel, D., Anton, G., Zanetti, G., Sezerman, O. U., Mendy, M., Valfk, D.,
1144 Lavitrano, M., Dagher, G., Zatloukal, K., van Ommen, G. B., and Litton, J.-E. (2018). Enhancing Reuse
1145 of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. *Biopreservation
1146 and Biobanking*, 16(2):97–105.
- 1147 Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K., and Müller, H. (2023). AI for life: Trends in
1148 artificial intelligence for biotechnology. *New Biotechnology*, 74:16–24.
- 1149 Hu, R., Yan, Z., Ding, W., and Yang, L. T. (2020). A survey on data provenance in IoT. *World Wide Web*,
1150 23(2):1441–1463.
- 1151 Ikeda, R., Park, H., and Widom, J. (2011). Provenance for generalized map and reduce workflows.
- 1152 Imran, A. and Agrawal, R. (2017). Data provenance. In Schintler, L. A. and McNeely, C. L., editors,
1153 *Encyclopedia of Big Data*, pages 1–4. Springer International Publishing, Cham.
- 1154 International Organization for Standardization (ISO) (2015). ISO/IEC 9001:2015 – quality management
1155 systems – requirements.
- 1156 Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., and
1157 Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis.
1158 *The Lancet*, 383(9912):166–175.
- 1159 Ives, Z. G., Green, T. J., Karvounarakis, G., Taylor, N. E., Tannen, V., Talukdar, P. P., Jacob, M., and
1160 Pereira, F. (2008). The orchestra collaborative data sharing system. *SIGMOD Rec.*, 37(3):26–32.
- 1161 Kale, A., Nguyen, T., Harris, Frederick C., J., Li, C., Zhang, J., and Ma, X. (2023). Provenance
1162 documentation to enable explainable and trustworthy AI: A literature review. *Data Intelligence*, pages
1163 1–24.
- 1164 Kanza, S., Willoughby, C., Gibbins, N., Whitby, R., Frey, J. G., Erjavec, J., Zupančič, K., Hren, M., and
1165 Kovač, K. (2017a). Electronic lab notebooks: can they replace paper? *Journal of Cheminformatics*,
1166 9(1):31.
- 1167 Kanza, S., Willoughby, C., Whitby, R. J., Erjavec, J., Zupančič, K., Hren, M., and Kovač, K. (2017b).
1168 Dataset for: Electronic lab notebooks: Can they replace paper?
- 1169 Khan, F. Z., Soiland-Reyes, S., Sinnott, R. O., Lonie, A., Goble, C., and Crusoe, M. R. (2019). Sharing
1170 interoperable workflow provenance: A review of best practices and their practical application in
1171 CWLProv. *GigaScience*, 8(11). giz095.
- 1172 Klyne, G., Groth, P., Moreau, L., Hartig, O., Simmhan, Y., Myers, J., Lebo, T., Belhajjame, K., Miles, S.,
1173 and Soiland-Reyes, S. (2013). Prov-aq: provenance access and query. *W3C Working Group Note*.
- 1174 Korolev, V. and Joshi, A. (2014). PROB: A tool for tracking provenance and reproducibility of big data
1175 experiments. *Reproduce'14. HPCA 2014*.
- 1176 Kwok, R. (2018). Lab notebooks go digital. *Nature*, 560(7717):269–270.
- 1177 Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., Crystal, R. G.,
1178 Darnell, R. B., Ferrante, R. J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman, H. E., Golub, R. M.,
1179 Goudreau, J. L., Gross, R. A., Gubitza, A. K., Hesterlee, S. E., Howells, D. W., Huguenard, J., Kelner,
1180 K., Koroshetz, W., Krainc, D., Lazic, S. E., Levine, M. S., Macleod, M. R., McCall, J. M., Moxley,
1181 Richard T, r., Narasimhan, K., Noble, L. J., Perrin, S., Porter, J. D., Steward, O., Unger, E., Utz, U., and
1182 Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical
1183 research. *Nature*, 490(7419):187–191.
- 1184 Lannom, L., Peters-von Gehlen, K., Anders, I., Pfeil, A., Schlemmer, A., Trautt, Z., and Wittenburg, P.
1185 (2022). FDO configuration types. PR-ConfigurationTypes-2.1-20221017.
- 1186 Leo, S. and Pireddu, L. (2023). Workflow run ro-crate capturing provenance from wsi conversion. Zenodo.
- 1187 Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A.,
1188 Hermsen, M., van de Loo, R., Vogels, R., Manson, Q. F., Stathonikos, N., Baidoshvili, A., van Diest, P.,
1189 Wauters, C., van Dijk, M., and van der Laak, J. (2018). Supporting data for "1399 H&E-stained sentinel
1190 lymph node sections of breast cancer patients: the CAMELYON dataset". *GigaScience Database*.

- 1191 Mahase, E. (2020). Covid-19: 146 researchers raise concerns over chloroquine study that halted who trial.
1192 *BMJ*, 369.
- 1193 Manghani, K. (2011). Quality assurance: Importance of systems and standard operating procedures.
1194 *Perspect Clin Res*, 2(1):34–37.
- 1195 Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M., and Zwelling, L. (2013). A survey on data
1196 reproducibility in cancer research provides insights into our limited ability to translate findings from
1197 the laboratory to the clinic. *PLOS ONE*, 8(5):1–4.
- 1198 Moreau, L. (2011). Provenance-based reproducibility in the semantic web. *Journal of Web Semantics*,
1199 9(2):202–221. Provenance in the Semantic Web.
- 1200 Moreau, L. (2016). ProvToolbox—Java library to create and convert W3C PROV data model representa-
1201 tions.
- 1202 Moreau, L. and Groth, P. (2013). Provenance: An introduction to prov. *Synthesis Lectures on the Semantic*
1203 *Web: Theory and Technology*, 3(4):1–129.
- 1204 Moreau, L. and Lebo, T. (2013). Linking across provenance bundles. *W3C Working Group Note*.
- 1205 Morrison, S. J. (2014). Time to do something about reproducibility. *eLife*, 3:1–4.
- 1206 Müller, H., Malservet, N., Quinlan, P., Reihls, R., Penicaud, M., Chami, A., Zatloukal, K., and Dagher, G.
1207 (2017). From the evaluation of existing solutions to an all-inclusive package for biobanks. *Health and*
1208 *Technology*, 7(1):89–95.
- 1209 Muniswamy-Reddy, K.-K., Holland, D. A., Braun, U., and Seltzer, M. (2006). Provenance-aware storage
1210 systems. In *Proceedings of the Annual Conference on USENIX '06 Annual Technical Conference*,
1211 ATEC '06, page 4, USA. USENIX Association.
- 1212 Muniswamy-Reddy, K.-K., Macko, P., and Seltzer, M. I. (2010). Provenance for the cloud. In *FAST*,
1213 volume 10, pages 15–14.
- 1214 Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., and Zatloukal, K. (2022). Explainability and
1215 causability for artificial intelligence-supported medical image analysis in the context of the european in
1216 vitro diagnostic regulation. *New Biotechnology*, 70:67–72.
- 1217 National Academies of Sciences, Engineering, and Medicine (2017). *Fostering Integrity in Research*.
1218 National Academies Press, Washington, D.C.
- 1219 Nickerson, D., Atalag, K., de Bono, B., Geiger, J., Goble, C., Hollmann, S., Lonien, J., Müller, W.,
1220 Regierer, B., Stanford, N. J., Golebiewski, M., and Hunter, P. (2016). The Human Physiome: how
1221 standards, software and innovative service infrastructures are providing the building blocks to make it
1222 achievable. *Interface Focus*, 6(2):20150103. 00001.
- 1223 Oliveira, W., Oliveira, D. D., and Braganholo, V. (2018). Provenance analytics for workflow-based
1224 computational experiments: A survey. *ACM Comput. Surv.*, 51(3).
- 1225 Park, H., Ikeda, R., and Widom, J. (2011). RAMP: A System for Capturing and Tracing Provenance in
1226 MapReduce Workflows. *Proc. VLDB Endow.*, 4(12):1351–1354.
- 1227 Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., and Stobie, E. (2010). Definition of the flexible
1228 image transport system (fits), version 3.0. *A&A*, 524:A42.
- 1229 Pérez, B., Rubio, J., and Sáenz-Adán, C. (2018). A systematic review of provenance systems. *Knowledge*
1230 *and Information Systems*, 57(3):495–543.
- 1231 Pimentel, J. a. F., Freire, J., Murta, L., and Braganholo, V. (2019). A survey on collecting, managing, and
1232 analyzing provenance from scripts. *ACM Comput. Surv.*, 52(3).
- 1233 Plessner, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology.
1234 *Frontiers in Neuroinformatics*, 11.
- 1235 Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published
1236 data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712.
- 1237 Rozenberg, G. (1997). *Handbook of Graph Grammars and Computing by Graph Transformation*.
1238 WORLD SCIENTIFIC.
- 1239 Rrmoku, K., Selimi, B., and Ahmedi, L. (2022). Provenance and social network analysis for recommender
1240 systems: a literature review. *International Journal of Electrical and Computer Engineering*, 12(5):5383.
- 1241 Servick, K. and Enserink, M. (2020). The pandemic’s first major research scandal erupts. *Science*,
1242 368(6495):1041–1042.
- 1243 Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance techniques. *Computer*
1244 *Science Department, Indiana University, Bloomington IN*, 47405:69.
- 1245 Simmhan, Y. L., Plale, B., and Gannon, D. (2006). A framework for collecting provenance in data-centric

1246 scientific workflows. In *2006 IEEE International Conference on Web Services (ICWS'06)*, pages
1247 427–436.

1248 Simmhan, Y. L., Plale, B., and Gannon, D. (2008). Karma2: Provenance management for data-driven
1249 workflows. *International Journal of Web Services Research (IJWSR)*, 5(2):1–22.

1250 Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L. J., Coppens, F., Fernández, J. M., Garijo, D., Grüning,
1251 B., La Rosa, M., Leo, S., Ó Carragáin, E., Portier, M., Trisovic, A., RO-Crate Community, Groth, P.,
1252 and Goble, C. (2022). Packaging research artefacts with RO-Crate. *Data Science*, 5(2):97–138.

1253 Spitzenberger, F., Patel, J., Gebuhr, I., Kruttwig, K., Safi, A., and Meisel, C. (2022). Laboratory-
1254 developed tests: Design of a regulatory strategy in compliance with the international state-of-the-art
1255 and the regulation (eu) 2017/746 (eu ivdr [in vitro diagnostic medical device regulation]). *Therapeutic
1256 Innovation & Regulatory Science*, 56(1):47–64.

1257 The Workflow Run RO-Crate working group (2023). Process run crate. Accessed 30 March 2023.

1258 Tufek, A. and Aktas, M. S. (2022). A systematic literature review on numerical weather prediction models
1259 and provenance data. In Gervasi, O., Murgante, B., Misra, S., Rocha, A. M. A. C., and Garau, C.,
1260 editors, *Computational Science and Its Applications – ICCSA 2022 Workshops*, pages 616–627, Cham.
1261 Springer International Publishing.

1262 Voges, J., Hernaez, M., Mattavelli, M., and Ostermann, J. (2021). An introduction to mpeg-g: The first
1263 open iso/iec standard for the compression and exchange of genomic sequencing data. *Proceedings of
1264 the IEEE*, 109(9):1607–1622.

1265 Wiener, J. L., Gupta, H., Labio, W., Zhuge, Y., Garcia-Molina, H., and Widom, J. (1996). A system
1266 prototype for warehouse view maintenance. In *VIEWS*, pages 26–33.

1267 Wittner, R. (2022). Distributed provenance information model for sensitive data in life sciences. <https://is.muni.cz/th/ed52n/>.

1268
1269 Wittner, R., Gallo, M., Leo, S., Mascia, C., Frexia, F., Plass, M., Soiland-Reyes, S., Müller, H., Geiger,
1270 J., and Holub, P. (2023a). Preprint: Linking provenance and its metadata in multi-organizational
1271 environments of life sciences. <https://s11.no/2023/phd/linking-provenance/>.

1272 Wittner, R., Gallo, M., Leo, S., and Soiland-Reyes, S. (2023b). Linking provenance and its metadata for
1273 an ai-based computation using cpm and ro-crate. Zenodo.

1274 Wittner, R., Holub, P., Mascia, C., Frexia, F., Müller, H., Plass, M., Allocca, C., Betsou, F., Burdett, T.,
1275 Cancio, I., Chapman, A., Chapman, M., Courtot, M., Curcin, V., Eder, J., Elliot, M., Exter, K., Goble,
1276 C., Golebiewski, M., Kisler, B., Kremer, A., Leo, S., Lin-Gibson, S., Marsano, A., Mattavelli, M.,
1277 Moore, J., Nakae, H., Perseil, I., Salman, A., Sluka, J., Soiland-Reyes, S., Strambio-De-Castillia, C.,
1278 Sussman, M., Swedlow, J. R., Zatloukal, K., and Geiger, J. (2023c). Toward a common standard for
1279 data and specimen provenance in life sciences. *Learning Health Systems*, n/a(n/a):e10365.

1280 Wittner, R., Holub, P., Müller, H., Geiger, J., Goble, C., Soiland-Reyes, S., Pireddu, L., Frexia, F., Mascia,
1281 C., Fairweather, E., Swedlow, J. R., Moore, J., Strambio, C., Grunwald, D., and Nakae, H. (2021a). Iso
1282 23494: Biotechnology – provenance information model for biological specimen and data. In Glavic,
1283 B., Braganholo, V., and Koop, D., editors, *Provenance and Annotation of Data and Processes*, pages
1284 222–225, Cham. Springer International Publishing.

1285 Wittner, R., Mascia, C., Frexia, F., Müller, H., Geiger, J., Exter, K., and Holub, P. (2021b). EOSC-life
1286 common provenance model.

1287 Wittner, R., Mascia, C., Gallo, M., Frexia, F., Müller, H., Plass, M., Geiger, J., and Holub, P. (2022).
1288 Lightweight distributed provenance model for complex real-world environments. *Scientific Data*,
1289 9(1):503.

1290 Wittner, R., Soiland-Reyes, S., and Leo, S. (2023d). Common provenance model ro-crate profile. Accessed
1291 30 March 2023.

1292 Xu, K., Ottley, A., Walchshofer, C., Streit, M., Chang, R., and Wenskovitch, J. (2020). Survey on the
1293 analysis of user interactions and visualization provenance. *Computer Graphics Forum*, 39(3):757–783.

1294 Zafar, F., Khan, A., Suhail, S., Ahmed, I., Hameed, K., Khan, H. M., Jabeen, F., and Anjum, A. (2017).
1295 Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes. *Journal of
1296 Network and Computer Applications*, 94:50–68.

1297 Zatloukal, K., Stumptner, C., Kungl, P., and Mueller, H. (2018). Biobanks in personalized medicine.
1298 *Expert Review of Precision Medicine and Drug Development*, 3(4):265–273.

1299 Zipperle, M., Gottwalt, F., Chang, E., and Dillon, T. (2022). Provenance-based intrusion detection systems:
1300 A survey. *ACM Comput. Surv.*, 55(7).