Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220489

The Common Provenance Model: Capturing Distributed Provenance in Life Sciences Processes

Francesca FREXIA[a,1], Cecilia MASCIA[a], Rudolf WITTNER[b,c], Markus PLASS[d], Heimo MÜLLER[d], Jörg GEIGER[e] and Petr HOLUB[b]
[a] *CRS4 – Center for Advanced Studies, Research and Development in Sardinia, Italy*
[b] *BBMRI-ERIC, Austria*
[c] *Masaryk University, Czech Republic*
[d] *Diagnostic and Research Institute of Pathology, Medical University of Graz, Austria*
[e] *Interdisciplinary Bank of Biomaterials and Data Würzburg (ibdw), Germany*

Abstract. The distributed nature of modern research emphasizes the importance of collecting and sharing the history of digital and physical material, to improve the reproducibility of experiments and the quality and reusability of results. Yet, the application of the current methodologies to record provenance information is largely scattered, leading to silos of provenance information at different granularities. To tackle this fragmentation, we developed the Common Provenance Model, a set of guidelines for the generation of interoperable provenance information, and to allow the reconstruction and the navigation of a continuous provenance chain. This work presents the first version of the model, available online, based on the W3C PROV Data Model and the Provenance Composition pattern.

Keywords. Provenance information, distributed processes, W3C PROV, Provenance Composition, Common Provenance Model

1. Introduction

Knowing the origin and evolution of data and samples is crucial to assess and improve the reliability and reusability of the results generated from their processing, as recognised also by the FAIR initiative [1]. Some approaches to express provenance information (PI) exist, but the lack of an overall coordination severely limit their impact, especially in *distributed* processes (e.g., large-scale analysis, AI applications), generally involving heterogeneous and complex steps, often performed by several groups at different times. We present here the Common Provenance Model (CPM) [2], a set of methodological recommendations we developed to guide the creation and exchange of interoperable PI.

2. Methods

The CPM design followed a domain-agnostic approach, considering different research processes in the life sciences within the EOSC-Life Project (involving thirteen European research infrastructures in the life sciences) and regional initiatives in Sardinia. The

[1] Corresponding Author, Francesca Frexia, CRS4, Italy; E-mail: francesca.frexia@crs4.it.

model focuses on three fundamental dimensions of the provenance lifecycle: the formalisation, linking, and navigation of a chain of provenance information. For the generation of provenance descriptions, we have adopted the W3C PROV formalism [3], expressing the provenance via instances formed by simple structures (*entity, agent, activity*) linked to relevant domain ontologies, to capture accurately the semantic of each use case. To support linking and navigation of PI in a distributed chain, we revised the Provenance Composition, a conceptual methodology connecting the PI records of two communicating processes through a shared entity [4]. The CPM extends the Provenance Composition approach to couple adjacent tasks, by defining practical recommendations for the PI navigation.

3. Results and Conclusions

The Common Provenance Model prescribes that the PI to be stored or further shared is generated in the form of PROV instances during *finalisation events*, periodically or on request, from the combination of all the relevant details recorded during the execution of a process (e.g., logs, metadata, adopted protocols). The transfer of "objects" – samples or data – is represented in the model introducing a specific PROV entity, the *connector*, to link the sender and the receiver through URIs univocally assigned. The model includes guidelines about fundamental elements to generate PI of good quality, like a versioning mechanism or the management of persistent identifiers. The first version of the CPM is available online [2], it has not been experimentally tested yet, but this limitation is mitigated by the fact that its development was guided by several use cases. Next steps include the refinement and application of the model. To encourage the adoption in the industrial sector, an instance of the CPM is the core of the *"ISO 23494: Biotechnology - Provenance Information Model for Biological Specimen and Data"* project [5], currently under development. Experts in the life sciences are invited to collaborate, through a national ISO body or by engaging with EOSC-Life online meetings.

Author Contributions and Acknowledgements

R.W. was the primary author of the distributed provenance information model. All the other authors provided continuous feedback and worked on refinements of the initial draft. This poster was prepared by F.F. and C.M. and the rest of the authors revised it. The work described is being co/funded by EOSC-Life Project (EU Horizon 2020, grant agreement no. 824087) and by DIFRA, SVDC and PAM Projects (Sardinian Regional Authority). This publication reflects only the view of the authors, and the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] Wilkinson M, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3(1):1-9.
[2] Wittner R, et al. EOSC-Life Common Provenance Model. Zenodo; 2021
[3] Groth P, Moreau L. PROV-Overview: An Overview of the PROV Family of Documents. (Accessed on 01/19/2022). https://www.w3.org/TR/prov-overview/.
[4] Buneman P, et al. Provenance Composition in PROV; 2017. https://eprints.soton.ac.uk/408513/
[5] Wittner R, et al. Provenance Week 2020 Poster: ISO 23494: Biotechnology - Provenance Information Model for Biological Specimen and Data. Zenodo; 2020.