Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200380

Data Provenance Standards and Recommendations for FAIR Data

Malte-Levin JAUER[a,1] and Thomas M. DESERNO[a]
[a] *Peter L. Reichertz Institute for Medical Informactics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany*

Abstract. This article reviews the main characteristics of five widely used data provenance models and recommendations. We suggest a set of six provenance properties that should be satisfied by any provenance model as a basis for further implementation of provenance mechanisms, supporting the findable, accessible, interoperable and reusable (FAIR) principles for both, research and health data.

Keywords. Data provenance, FAIR data, Metadata, Research data, Health data

1. Introduction

In health research, data capture and data quality varies strongly. Therefore, information on data provenance is needed along the whole processing pipeline [1]. This includes the generation of persistent identifiers (PIDs) to make the data findable and accessible and is crucial to reuse data. Therefore, providing data provenance information is a mandatory step towards findable, accessible, interoperable and reusable (FAIR) data [2].

2. Methods

We consider five provenance standards identified within the FAIR4Health project [3]. A widely used provenance model is the W3C PROV-DM data model [4]: an acyclic directed graph, consisting of nodes “entity”, “activity”, and “agent”. Recommending specific provenance items, the DataCite International Consortium developed a metadata scheme in 2009 [5]. It stresses assignment of digital object identifiers (DOIs) and includes six domain-agnostic mandatory properties. In 2016, a domain-specific extension to the DataCite metadata schema for health was presented: the ECRIN Clinical Research Metadata Schema [6]. It includes information on the source study, associated consent and access details. The Research Data Alliance endorsed 14 recommendations of the Working Group Data Citation (WGDC) [7] targeting reproducibility of data used in experiments and studies. Therefore, persistent identifiers have to be generated in a query-based manner, so that data views can be cited and retrieved by re-executing the query. As a result of the Data Quality Collaborative (DQC), Kahn et al. [8] proposed 20 data quality and provenance recommendations. They especially highlight that each transformation of the source data has to be documented, including data cleansing values.

[1] Corresponding Author: Malte-Levin Jauer, Peter L. Reichertz Institute for Medical Informatics, Mühlenpfordtstr. 23, 38106 Braunschweig, Germany; E-mail: malte-levin.jauer@plri.de.

3. Results

We extracted the following list as minimal “fit for use” requirements for provenance model (Table 1). Check-marks indicate, which recommendation(s) support these items.

Table 1. Comparison of the different provenance recommendation sets.

Criteria	DataCite	ECRIN	WGDC	DQC
Persistent identifier (PID): Each data object is assigned a unique, persistently stored identifier. Ideally, a DOI is assigned.	✓	✓	✓	✗
Data origin: The project or event that generated the data.	✓	✓	✗	✓
Data creator: A person or institution to be credited for.	✓	✓	✗	✓
Data timestamp: The time of dataset creation/modification.	✓	✓	✓	✗
Data versioning: Each transformation result of the data object is stored. Earlier versions are retrievable.	✓	✓	✓	✓
Query PID: If (sub-)sets of data are generated or cited, the query is stored with a persistent ID for reproducibility.	✗	✗	✓	✓

4. Discussion

The present work has identified six minimal criteria from the given provenance overview, implementable using the PROV-DM data model. The feasibility of these items will be investigated in the FAIR4Health project’s demonstrators.

Acknowledgements

This work was performed in the framework of FAIR4Health project. FAIR4Health has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement number 824666.

References

[1] Baum B, Bauer CR, Franke T, Kusch H, Parciak M, Rottmann T, et al. Opinion paper: data provenance challenges in biomedical research. *it - Information Technology*. 2017;59(4):191–196.

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data.* 2016 Mar;3:160018.

[3] FAIR4Health. D2.2. Functional design of the FAIR4Health platform and agents. Internal report. 2019.

[4] Moreau L, Missier P, editors. PROV-DM: the PROV data model [Online]. 2013 Apr 30 [cited 2019 Dec 21]. Available from: https://www.w3.org/TR/2013/REC-prov-dm-20130430/

[5] DataCite Metadata Working Group. DataCite metadata schema documentation for the publication and citation of research data v4.3 [Online]. Version 4.3. 2019 [cited 2019 Dec 21]. Available from: https://schema.datacite.org/meta/kernel-4.3/

[6] Canham S, Ohmann C. A metadata schema for data objects in clinical research. *Trials*. 2016 Nov;17(1):557.

[7] Rauber A, Asmi A, van Uytvanck D, Proell S. Data citation of evolving data: recommendations of the working group on data citation (WGDC) [pdf]. RDA WG Data Citation. 2015 Oct [cited 2019 Dec 21]. Available from: https://www.rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf

[8] Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, et al. Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC).* 2015 Mar;3(1):7.