German Medical Data Sciences: A Learning Healthcare System
U. Hübner et al. (Eds.)
© 2018 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-896-9-28

Practical Extension of Provenance to Healthcare Data Based on the W3C PROV Standard

Ann-Kristin KOCK-SCHOPPENHAUER[a,1], Lina HARTUNG [b], Hannes ULRICH [a,b], Petra DUHM-HARBECK [a] and Josef INGENERF [a,b]
[a] *IT Center for Clinical Research, Lübeck (ITCR-L), University of Lübeck, Germany*
[b] *Institute of Medical Informatics, University of Lübeck, Germany*

Abstract. Secondary use of healthcare data is dependent on the availability of provenance data for assessing its quality, reliability or trustworthiness. Usually, instance-level data that might be communicated by HL7 interfaces entail limited metadata about involved software systems, persons or organizations bearing responsibility for those systems. This paper proposes a strategy for capturing interoperable provenance data needed by data stewards for assessing healthcare data that are reused in a research context. Aimed at a realistic level of granularity even system-level metadata will support a data steward trying to trace the origins or provenance of healthcare data that have been transferred to the research context. Those metadata are extracted from the 3LGM^2-system, used for modelling hospital information systems. Based on the W3C provenance specification interrelated activities, entities and agents can be integrated and stored in RDF triple stores and therefore queried and visualized.

Keywords. Secondary use of EHR data, interoperable provenance data

1. Introduction

In medical research, genotype and phenotype data from multiple sources as well as workflows and log records of corresponding data processing steps are distributed among heterogeneous information systems that are used under the authority of different healthcare actors [1]. Provenance is regarded as the equivalent of a logbook: Capturing all the steps that were involved in the actual derivation of a result, and which could be used to replay the execution that led to that result [2]. In the comparatively young field of bioinformatics the demand of reproducible science is accomplished by provenance-aware workflow systems e.g. Galaxy [3]. These workflow systems take advantage of an ecosystem of almost harmonized services, e.g. from EMBL or NCBI which can be integrated in compact workflows for generating experiment pipelines. Such an approach with executing prospectively defined process models are not suitable for capturing the provenance of healthcare data. The recording of relevant context data for data capturing and processing steps of interest needs to be treated differently.

[1] Corresponding Author, Ann-Kristin Kock-Schoppenhauer, IT Center for Clinical Research, Lübeck, Universität zu Lübeck, Ratzeburgerallee 160, 23562 Lübeck, Germany; E-mail: ann-kristin.kock@uksh.de.

In the following a twofold approach for capturing provenance data is proposed: The data processing activities and entities at instance-level are enriched by linkages to responsible agents at system-level. In research context data stewardship would gain a significant added value, reviewing the quality of medical data, if links to responsible applications, persons or organizations are available, see chapter 5 “Conclusion”.

2. Methods

The de-facto standard W3C PROV for interoperable provenance data [2, 4] is based on a conceptual model with basically three classes *activity*, *entity*, *agent* and nine binary relationships like “*wasGeneratedBy*” or “*startedAtTime*”, see Fig 1. Data processing steps like “*laboratory_data wasGeneratedBy laboratory_ analysis*” are represented as RDF-triples of corresponding instances. Using semantic web formats like OWL for ontologies, RDF for instance data and several data interchange formats like RDF/XML, N3 or Turtle, facilitates the utilization of powerful tools. For processing, querying or visualizing provenance graphs consisting of interconnected triples; e.g. the ontology editor PROTÉGÉ [5], SPARQL-query tools [6] and visualization tools [7] can be used.

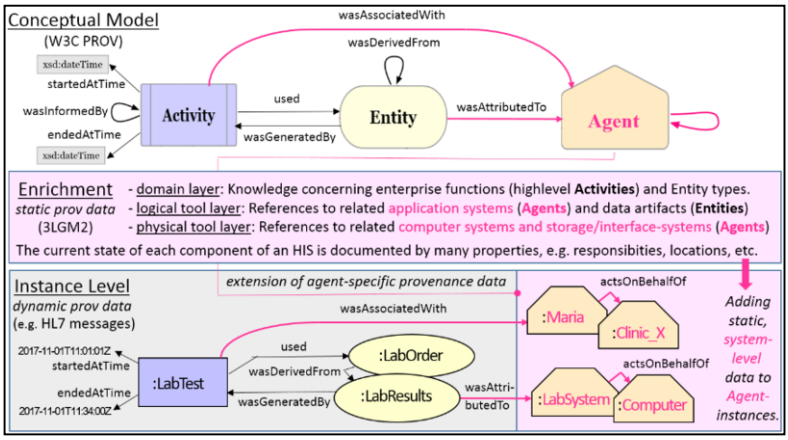

Figure 1. General approach for capturing interoperable provenance data.

In order to capture provenance data of processing steps, analogously to the bioinformatics pipelines, the workflow tool TAVERNA [8] has been evaluated. By launching a defined workflow with suitable service connectors instance-level data from different sources can be accessed. TAVERNA inherently records service invocation, intermediate and final workflow results and exports provenance data as RDF triples formatted in Turtle syntax conformant to the W3C PROV standard. However, the lack of suitable service connectors in routine clinical application systems like a laboratory software, leads to the conclusion that TAVERNA is not suitable for accessing the provenance of healthcare data. For that reason, available clinical data is used directly for deriving a W3C conformant model of provenance data, e.g. by starting from HL7 V2 messages. The main motivation for this paper is the observation that provenance data is not available or is not communicated, e.g. attributes for almost all data elements in databases like “created_by”, created_date”. Therefore, missing information about responsibilities are supplemented by linkages to suitable metadata at the system level, e.g. about agents like administrating persons related to application software systems.

3LGM2 is a tool for modelling hospital information systems at three interrelated layers presented in Fig 1 [9]. The enterprise functions at the domain layer correspond to W3C PROV *activities,* which are supported by software systems at the logical tool layer, known among W3C as *agents*. These systems use data storage or message artefacts correlated to W3C *entities,* that are realizations of Entity types in 3LGM2 models. References to the physical tool layer allow accessing information about responsibilities for hosting application systems opposed to context data concerning responsibilities to operate software systems at the logical tool layer.

The 3LGM^2 tool provides a set of predefined queries, like the availability of enterprise features because of network component failures or generic assessments like functional redundancies within a Hospital Information System (HIS). However, it does not support simulations of enterprise functions by processing real data and adding provenance at all three layers [10]. With respect to missing instance-level data, context data describing W3C agent instances, the static system-level data, structured documentation on the state of a distributed information system is very valuable.

3. Results

A simplified laboratory use case has been chosen to demonstrate the proposed strategy for integrating instance- and system-level provenance data by using the mentioned tools. Aiming at an improved secondary use of healthcare data in the medical research context, consumers like data stewards, members of use-access-committees or researchers should be supported in assessing data´s context and history.

3.1. Phase 1: 3LGM^2 Model Providing System-Level Data to the PROV Agent Node Type

At all three layers the consecutive data processing steps “capturing of laboratory data”, “mapping to LOINC” and “re-used in a research IT platform” are modelled. All components are described in detail via dialog windows, as shown in Fig 2. Of special interest are involved application and computer systems providing provenance data for agents.

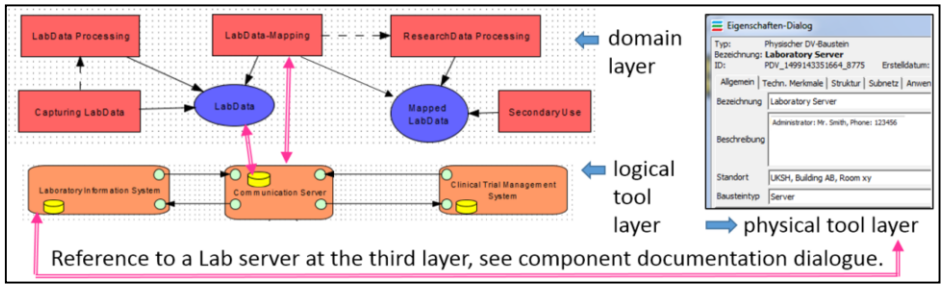

Figure 2. Simplified 3LGM2 model with exemplary inter-layer relationships based on a laboratory use case.

3.2. Phase 2: Instance-Level Provenance Data Extracted from HL7 Messages

Instance-level provenance data can be obtained for example by analyzing and annotating HL7 messages with respect to suitable PROV classes “ACT” (e.g. LabTest “UMIC^ Urinalysis” in the OBR segment) and “ENT” (e.g. LabResult like “UGLB^Glucose” in

OBX segments). The header segment provides the corresponding instance of the PROV class “AGNT”. The laboratory information system presented in the 3LGM² model at the logical tool layer in Fig. 2 are provided by HL7 V2 messages. The LabData-Mapping where proprietary names are mapped to LOINC (e.g. UGLB^Glucose to 25428-4) is not explicitly shown.

3.3. Phase 3: Integrating Instance- and System-Level Data from Phase 1 and 2

RDF (Resource Description Format) is used for representing instance data from HL7 messages [11] as well as system level data from the 3LGM² model complemented by general data about responsible persons and organizations. To avoid name collisions and to refer to defined ontologies namespaces like “f3LGM” [12] and “FOAF” [13] are necessary. For space reasons, the following examples illustrate just the rough idea.

```
<LabTest_2018_9876543...>
  rdfs:label "UMIC^Urinalysis" ;
  rdf:type prov:Activity ;
  hl7:segment “OBR_1 (LabResult)” ;
  hl7:message <http://hl7.org/message/2018/a63d332-17ea-4fa3-.....> ;
  prov:startedAtTime  "2009-05-04T12:13"^xsd:dateTime ;
  prov:wasAttributedTo  <LabSystem#> . “Taken from HL7 message header”
```

```
<LabSystem#>
  rdfs:label "Laboratoryinformation System" ;
  rdf:type prov:SoftwareAgent, f3LGM:Appl.System ;
  f3LGM:basesOn “OPUS::L” ;
  prov:actedOnBehalfOf <Computer_12> .
```

```
<Computer_12>
  rdf:type prov:SoftwareAgent , f3LGM:Comp.System ;
  f3LGM:belongsTo  <subnet_123> ;
  f3LGM:isLocated <UKSH_room_456> ;
  f3LGM:ContactPerson  <Person_789> .
```

```
<Person_789>
  rdf:type prov:Agent, foaf:Person ;
  foaf:givenName “Mr. Smith” ;
  foaf:mbox <mailto:smith@example.org> ;
  prov:actedOnBehalfOf <Organization_159> .
```

```
<Organization_159>
  rdf:type prov:Agent, foaf:Organization ;
  foaf:name “University Hospital ” .
```

Note(f3LGM): Namespaces must start with letters.

3.4. Phase 4: Analyzing the Complete Provenance Graph

Due to the mainstream technology many opportunities of analyzing the integrated RDF dataset occur. Straightforward, the turtle file with provenance data can be visualized [7].

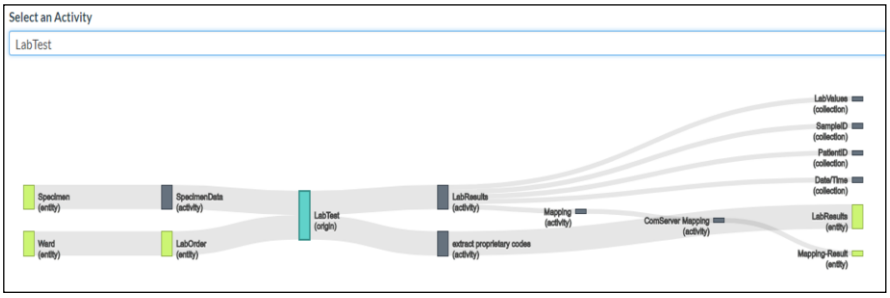

Figure 3. Web-based PROV-O-VIZ tool for visualizing W3C PROV conformant provenance data.

For a deeper analysis, semantic web tools like the ontology editor PROTÉGÉ [5] or triple stores like LUPOSDATE [6] are used for evaluating SPARQL queries. For enabling inferences based on the PROV ontology with suitable axioms like inverse properties, e.g.

"Activity prov:generated Entity" or subclass relationships, e.g. *agent* with subclasses like Person, the exported turtle file and PROV OWL file are merged.

4. Discussion

Extracted healthcare data from HL7 messages and linked metadata provided by a separated HIS modelling tool like 3LGM^2 can be integrated suitably by using RDF triples conformant to the PROV model. This was prototypically implemented to demonstrate the feasibility of this approach for making provenance data available und usable. However, the 3LGM^2-tool needs improved interfaces for accessing agent-related data automatically and a suitable mechanism for uniquely identifying agents.

5. Conclusion

Adding metadata about responsible persons or other relevant context item alone is certainly not enough in the end. Regarding the potential granularity of provenance data targeted to application systems, data elements or values, the proposed approach can be extended by using more ambitioned methods presented by Curcin et al. [14].

6. Conflict of Interest

All authors declare no conflict of interest.

References

[1] V. Curcin, S. Miles, R. Danger, Y. Chen, R. Bache, A. Taweel, Implementing interoperable provenance in biomedical research, *Future Generation Computer Systems* **34** (2014), 1-16.
[2] L. Moreau, P. Groth, J. Cheney, T. Lebo, S. Miles, The rationale of PROV, *Web Semantics: Science, Services and Agents on the World Wide Web* **35** (2015), 235-257.
[3] S. Kanwal, F.Z. Khan, A. Lonie, R.O. Sinnott, Investigating reproducibility and tracking provenance - A genomic workflow case study, *BMC Bioinformatics* **18(1)** (2017), 337.
[4] P. Groth, L. Moreau, PROV-Overview - An Overview of the PROV Family of Documents. 2013, https://www.w3.org/TR/prov-overview/ [2018 May 01.
[5] Protégé, http://protege.stanford.edu/ [2018 May 01].
[6] LUPOSDATE, https://www.ifis.uni-luebeck.de/~groppe/luposdate-js-client/ [cited 2018 May 01].
[7] R. Hoekstra, P. Groth, PROV-O-Viz - Understanding the Role of Activities in Provenance, *Int Provenance and Annotation Workshop), LNCS* **8628** (2014), 215-220, http://provoviz.org/ [cited 2018 May 01].
[8] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, et al., The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud, *Nucleic Acids Res* **41** (2013), W557-61.
[9] T. Wendt, A. Haber, B. Brigl, A. Winter, Modeling Hospital Information Systems (Part 2): using the 3LGM^2 tool for modeling patient record management, *Methods Inf Med* **43(3)** (2004), 256-67.
[10] M. Staemmler, Modeling a health telematics network: does the 3LGM^2 approach assist in its management and operation? *Stud Health Technol Inform.* **124** (2006), 691-6.
[11] F. Prasser, F. Kohlmayer, A. Kemper, K. Kuhn, A Generic Transformation of HL7 Messages into the Resource Description Framework Data Model, *GI-Jahrestagung* (2012), 1559-1564.
[12] A.Winter, B. Brigl, T. Wendt, A UML-based ontology for describing hospital information system architectures, *Stud Health Technol Inform.* **84(Pt 1)** (2001), 778-82.
[13] FOAF Ontology (Friend of a Friend), http://xmlns.com/foaf/spec/.
[14] V. Curcin, E. Fairweather, R. Danger, D. Corrigan, Templates as a method for implementing data provenance in decision support systems, *J Biomed Inform.* **65** (2017), 1-21.