G-PROV: Provenance Management for Clinical Practice Guidelines

Nkechinyere N. Agu[1][0000−0003−1386−8602], Neha Keshan[1][0000−0002−0752−6406], Shruthi Chari[1][0000−0003−2946−7870], Oshani Seneviratne[1][0000−0001−8518−917X], Sabbir M. Rashid[1][0000−0002−4162−8334], Amar K. Das[2][0000−0003−3556−0844], James P. McCusker[1][0000−0003−1085−6059], and Deborah L. McGuinness[1][0000−0001−7037−4567]

[1] Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[2] IBM Research, Cambridge, MA

Abstract. Providing provenance of treatment suggestions made by clinical decision support systems can enhance transparency and trust in these systems by healthcare practitioners. Provenance can aid in resolving ambiguity and conflicts between various guideline sources. We have developed a guideline provenance ontology, G-Prov, by extending existing provenance ontologies, to enable accurate encoding of the source of the reasoning rules that decision support systems rely on to generate diagnosis and treatment suggestions. Our ontology enables provenance representation at different granularity levels within guidelines. For instance, G-Prov can be used to annotate rules with citations found in evidence sentences as well as other sources of knowledge, such as figures and tables. Additionally, we have developed an application to show a range of use cases for our ontology. We demonstrate our work annotating recommendations in a CPG for Type-2 Diabetes and discuss how our approach could be used in various medical settings where CPGs are utilized.

Keywords: Provenance · Clinical Practice Guidelines · Clinical Decision Support Systems · Ontology · Biomedical Knowledge Graphs

Ontology: https://purl.org/heals/gprov

Website: https://tetherless-world.github.io/GProv/

1 Introduction

Guideline-based decision support systems aim to assist healthcare practitioners with patient diagnosis and treatment choices. These systems embody information present in authoritative clinical practice guidelines (CPGs) as computable knowledge (e.g. logical rules). However, most of these systems fail to provide the source of treatment suggestions. Furthermore, the information used to generate these treatment suggestions might become obsolete, or worse yet, invalidated, within a newer guideline. Recording the provenance of treatment suggestions

helps address the challenging task of systematically updating a system when new guidelines or medical literature are published. To concretely tackle the problem of lack of provenance, there needs to be an easy way to extract information within the CPG and associate them as provenance.
We attach provenance to SWRL [5] treatment rules of the Diabetes Mellitus Treatment Ontology (DMTO) [4]. However since DMTO's rules lack provenance, we use DMTO as an example usage of our guideline provenance ontology: G-Prov and annotate these rules with recommendations from the 2018 version of the ADA Standards of Medical Care guideline (ADA CPG)[3]. DMTO is an ontology that provides treatment suggestions for Type-2 Diabetes. They use information from several medical guidelines on Diabetes, including the ADA CPG, Diabetes Canada,[4] and the European Association for the study of Diabetes.[5] As stated earlier, DMTO lacks information on the source of each rule within the ontology, making it difficult to evaluate the accuracy of and the evidence for each rule. Hence, we address this issue, in our paper.
To this end, we develop a guideline provenance ontology, G-Prov, to encode the provenance details in CPGs. For our ontology, we reused several existing ontologies, such as, W3C provenance ontology (PROV-O) [7], Dublin Core Metadata Terms (DCT) [6], and the Bibliographic Ontology (BIBO) [2]. Further, utilzing concepts from G-Prov, we modeled the provenance of DMTO rules with content from the ADA CPGs as Resource Description Framework (RDF) [9] Knowledge Graphs (KGs). We show the diverse use of our ontology-enabled approach by developing an application for three different use cases that cover a variety of users of guideline documents.

2 Related Work

Several general-purpose provenance ontologies currently exist such as the Provenance Ontology (PROV-O) [7], which is an ontology that provides classes and properties to capture generic provenance terms in various domain. The Dublin Core Metadata Terms (DCT) [6], which is a lightweight vocabulary that aids in the description of resources and provides a list of terms that can be used to describe metadata information. DCT can be used to represent articles, figures, tables on a higher level. On the other hand, The Bibliographic Ontology (BIBO) [2], serves as an enhancement of the DCT ontology and contains a more detailed description of referenced articles. However, all these provenance ontologies address the issue of modeling of provenance at a more general level. In G-Prov, we build off these foundational provenance ontologies to associate provenance of clinical decisions with their authoritative guideline evidence.
The clinical domain have plethora of heterogenous data of varying qualities. Hence, there have been several efforts to create ontologies that enable increased traceability, transparency and trust in clinical data. ProvCaRe [12] is an ontology

[3] ADA Standards of Medical Care: http://care.Diabetesjournals.org/
[4] Diabetes Canada: https://guidelines.Diabetes.ca/cpg
[5] EASD: https://www.easd.org/

designed to enhance reproducibility of scientific research by capturing the metadata of published articles. Their ontology focuses on the details of the scientific study including the design description, the data collection methods, analysis of the data and overall research methodology. Provenance Context Entity (PaCE) [11] is a an ontology that enables provenance tracking in scientific data. This ontology has been applied to the Biomedical Knowledge Repository to integrate biomedical data from various biomedical literature, databases, and vocabulary lists. Furthermore, in healthcare, there is an ontology for enabling provenance in healthcare management [3] that enables the efficient integration of patient data from various clinical sources, and helps in understanding patient outcome as a result of a clinical process. While these ontologies exist for the clinical domain, they fail to capture provenance terms necessary for guideline-specific data. The authors of [13] have developed a semantic web system that attempts to capture guideline interactions. Their research has some similarities with our work, however, they focus on identifying causation which is not within the scope of G-Prov ontology.

3 Guideline Provenance Ontology

We aim to provide an ontology that is compatible with best practice, relevant provenance ontologies, and which is focused at directly support clinical, guideline-based decision support systems. Our initial attempts focus mainly on annotating SWRL rules defined within DMTO ontology with the recommendations found within the ADA CPG. Typically, a CPG comprises of recommendations for clinical practice, cited publications backing these recommendations etc. To address the use case of associating provenance with treatment suggestions encoded as rules, we included classes and properties (described shortly in section 3.1) in the G-Prov ontology, to link these rules to different granularities of evidence within CPGs.

3.1 Main Classes and Property Associations

Our ontology is represented in OWL-DL and was developed using Protege [10]. The G-Prov ontology [6] comprises of terms broadly belonging to three sections. A broader provenance section that captures metadata associated with the provenance of a treatment rule in the context of its backing CPG recommendation. These terms were obtained from the general purpose ontologies described earlier in section 2, i.e. PROV-O, DCT, and BIBO. The second section is comprised of more domain specific terms that are used to associate rules to their guideline evidence. The third section discusses the guideline specific classes and properties.

[6] We use the following ontology prefixes: (1) gprov: G-Prov (2) sio: SemanticScience Integrated Ontology (3) bibo: The Bibliographic Ontology (4) prov: PROV-O: The PROV ontology (5) dct: Dublin Core Terms (6) sco: Study Cohort Ontology (7) do: Disease Ontology (8) ncit: National Cancer Institute Thesaurus

3.2 Provenance in Guidelines

Both CPGs (ncit:C28237) and the cited publications within them are specific sub-classes of bibo:Document. Additionally, citations within CPGs are represented by gprov:Citation, and are linked to their reference information using prov:hasPrimarySource. Further, every published document has at least one author and, to capture the list of all the authors in a publication, we use the class bibo:Author, which contains one or more authors. We also capture other metadata information associated with a document using terms from both BIBO (bibo:volume, bibo:issue, bibo:pageStart, bibo:pageEnd, bibo:uri) and DCT (dct:creator, dct:title) ontologies.

3.3 Associating provenance with treatment rules

Since we are annotating treatment rules, we need a class to model them and we introduce the gprov:FormalRule class. Additionally, these rules (gprov:FormalRule) were obtained from guideline recommendations (gprov:Recommendation) and we use the prov:wasDerivedFrom property to link the two classes. At a more generic level, we can also link the rules directly to the guideline (ncit:C28237). We use dct:partOf association to represent the relationship between a recommendation and its host CPG.

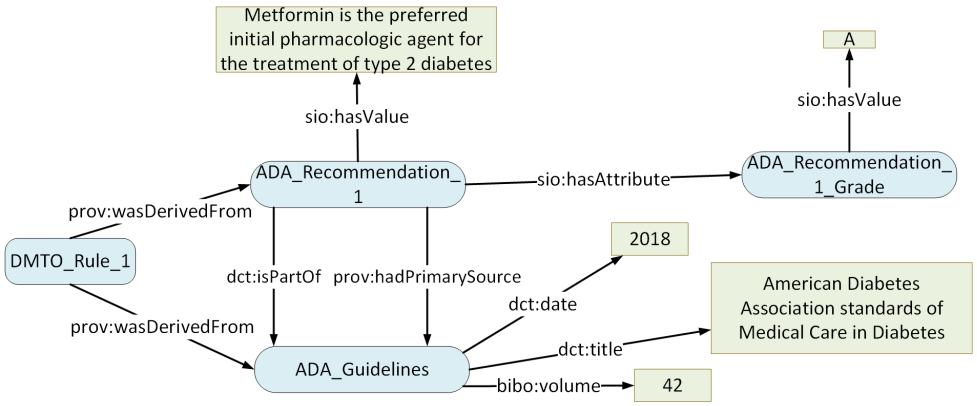

Fig. 1. A simple instance diagram showing the modeling of a recommendation within the ADA CPG using G-Prov Ontology

3.4 Guideline Specific Classes and Properties

CPGs are developed to manage a specific health condition. In G-Prov, we link the guideline class (ncit:C28237) to the health condition (gprov:DiseaseManagement)

it addresses via prov:used property. The gprov:DiseaseManagement class links to (via prov:used) at least one specific disease type (do:Disease). Additionally, some guidelines carry a measure of the quality of each recommendation, and we capture this information using gprov:Grade. Furthermore, guideline recommendations are backed by one or more evidence sentences (gprov:EvidenceSentence) within the guideline. These evidence sentences contain citations, whose modeling using the existing provenance ontologies is explained in section 3.2.

4 ADA Standards of Care Guideline Knowledge Graph

There are numerous CPGs for effective diagnosis and treatment of diseases. Through G-Prov, we support the modeling of recommendations in the ADA CPG. The ADA CPG is aimed at providing treatment recommendations for healthcare practitioners to help in the management of Type-1 and Type-2 Diabetes. The ADA CPG is released annually, with updates to recommendations, recommendation grade, evidence supporting each recommendation, and the medical literature that backs each piece of evidence.

To annotate a treatment rule with its ADA CPG recommendation, we first did a manual pass on the CPG to understand its structure. Subsequently, we wrote a web scraping script using BeautifulSoup[7] to extract all the contents of the guideline into a JSON file. The contents of the JSON file included all the recommendations, their grades, possible evidence sentences, and the citations within each sentence. Our initial efforts focused on two chapters from the ADA CPG, namely the "Pharmacologic Approaches to Glycemic Treatment" and "Cardiovascular Disease and Risk Management" chapters.

To construct the KG, we created a spreadsheet to organize the extracted guideline data. Further, we manually identified applicable guideline recommendations for DMTO treatment rules and mapped them within the spreadsheet. We had our medical domain expert (also a co-author) go over these annotations to confirm both completeness and accuracy. Finally, we wrote a SETLr [8] script to automatically convert the contents of the spreadsheet to an RDF KG. Fig. 1 highlights the resulting KG from modeling a recommendation within the ADA CPG using G-Prov ontology.

5 Applications of G-Prov

The G-Prov ontology and the ADA KG were used to develop an application and showcase its various uses for a variety of users. The application evolved out of the need for healthcare practitioners to identify the sources of treatment suggestions made by CDS. The various pieces of information that a healthcare practitioner needed to make clinical decisions were identified with the help of our medical domain expert and we began iteratively building an ontology-enabled system for

[7] https://www.crummy.com/software/BeautifulSoup

our use case. We used a standard ontology engineering process to ensure both completeness and adaptability into a wide range of applications. Our application captures three main use cases, described in the next three paragraphs. Screenshot of the various views of the applications can be viewed on our resources website [8].

The first use case focuses on providing provenance information of treatment

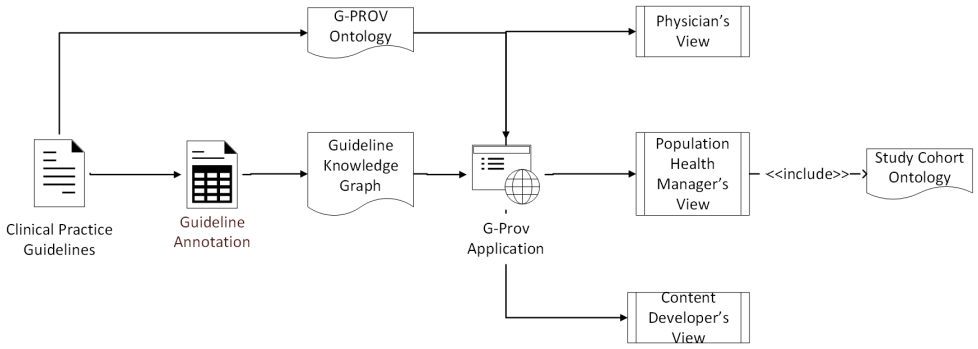

Fig. 2. A high-level work flow architecture diagram for the guideline provenance application

suggestions made by CDS, including the source of the information, other medical literature that support the article and the date of publication. Thus, when a treatment suggestion is made by a CDS system, the healthcare practitioner's view enables the healthcare practitioner to query the system for more information about its suggestion.

The second use case focuses on a Population Health Manager (PHM). To create a view for this user, we integrate the Study Cohort Ontology [1], to assist in visualizing information about citations. Through this view, we aim to help the PHM identify subgroups (cohorts) of patients who may benefit from the recommendations. Additionally, the visualizations also help them analyze the research publications that serve as evidence for recommendations.

The final use case is focused on a content developer. This view would allow for developers of CDS systems to enter provenance information while creating/editing the decision rules. The added advantage of creating a content developers view in our application, over other editing tools such as Protege, is that this view does not assume the user has any semantic knowledge whatsoever. This view includes a form to collect information from the user in a friendly, easy to read manner and creates a KG of the information entered using the G-Prov ontology.

[8] Visit https://tetherless-world.github.io/GProv/

6 Results and Discussion

We have developed a guideline provenance ontology, G-Prov, and used it to enrich DMTO ontology, by associating provenance information with treatment rules defined by their ontology. We have reused terminology from relevant ontologies such as DCT, Prov-O, BIBO. Specifically, DCT provides concepts that were relevant to model parts of the reference information, the remaining information was modeled using BIBO. Further, we represent the unstructured guideline data as a structured RDF knowledge graph.
We build an application that comprises of different views rendered by results of SPARQL queries triggered to the ontology and the KG. The different views of the application represent the various use cases that our ontology-enabled system can be used in. Our application serves as a proof of concept for the development of systems that can benefit from identifying provenance of CDS treatment suggestions.
Through, G-Prov we are taking a step towards enhancing transparency in healthcare by providing more information to healthcare practitioners regarding the granularities of guideline evidence behind treatment suggestions. We are focusing on improving the scalability and automation aspects of our system. Specifically, we are exploring the use of NLP techniques to automatically/semi-automatically identify evidence sentences that support a given guideline recommendation. Although the G-Prov ontology has currently been used to model the ADA CPG, we envisage that we can extend upon the ontology and introduce classes to model guideline provenance from other CPGs.

7 Conclusion

In this paper, we have described the guideline provenance ontology: G-Prov, which is a simple ontology for modeling provenance information of treatment rules generated by CDS systems. In the ontology design process, we have reused terms from several widely accepted ontologies to increase interoperability. We have built a prototype application supported by the G-Prov ontology to show its various use cases. Currently, we only map provenance of treatment rules to the ADA 2018 CPG. However, G-Prov can be used to track the provenance of information in other guidelines. We believe G-Prov is a good complement to existing provenance ontologies, and thus it is our belief that the our paper will be a useful contribution to both healthcare practitioners and CDS system designers who are looking to develop treatment suggestion tools.

8 Acknowledgements

This work is partially supported by IBM Research AI through the AI Horizons Network. We thank our colleagues from IBM (Ching-Hua Chen) and RPI (James Hendler, John Erickson, Kristen Bennett, Rebecca Cowan) who provided insight and expertise that greatly assisted the research.

References

1. Chari, S., Qi, M., Agu, N.N., Seneviratne, O., McCusker, J.P., Bennett, K.P., Das, A.K., McGuinness, D.L.: Making study populations visible through knowledge graphs. In: International Semantic Web Conference (ISWC). p. to appear. Auckland, New Zealand (2019)
2. Dabrowski, M., Synak, M., Kruk, S.R.: Bibliographic ontology. In: Semantic digital libraries, pp. 103–122. Springer (2009)
3. Deora, V., Contes, A., Rana, O.F., Rajbhandari, S., Wootten, I., Tamas, K., Varga, L.Z.: Navigating provenance information for distributed healthcare management. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. pp. 859–865. IEEE Computer Society (2006)
4. El-Sappagh, S., Kwak, D., Ali, F., Kwak, K.S.: Dmto: a realistic ontology for standard diabetes mellitus treatment. Journal of biomedical semantics **9**(1), 8 (2018)
5. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al.: Swrl: A semantic web rule language combining owl and ruleml. W3C Member submission **21**(79) (2004)
6. Kunze, J.A., Baker, T.: The dublin core metadata element set (2007)
7. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. W3C recommendation **30** (2013)
8. McCusker, J.P., Chastain, K., Rashid, S., Norris, S., McGuinness, D.L.: Setlr: the semantic extract, transform, and load-r. PeerJ Preprints **6**, e26476v1 (2018)
9. Miller, E.: An introduction to the resource description framework. Bulletin of the American Society for Information Science and Technology **25**(1), 15–19 (1998)
10. Noy, N.F., Fergerson, R.W., Musen, M.A.: The knowledge model of protege-2000: Combining interoperability and flexibility. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 17–32. Springer (2000)
11. Sahoo, S.S., Bodenreider, O., Hitzler, P., Sheth, A., Thirunarayan, K.: Provenance context entity (pace): Scalable provenance tracking for scientific rdf data. In: International Conference on Scientific and Statistical Database Management. pp. 461–470. Springer (2010)
12. Valdez, J., Kim, M., Rueschman, M., Socrates, V., Redline, S., Sahoo, S.S.: Provcare semantic provenance knowledgebase: evaluating scientific reproducibility of research studies. In: AMIA Annual Symposium Proceedings. vol. 2017, p. 1705. American Medical Informatics Association (2017)
13. Zamborlini, V., Hoekstra, R., Da Silveira, M., Pruski, C., ten Teije, A., van Harmelen, F., et al.: Generalizing the detection of internal and external interactions in clinical guidelines. In: HEALTHINF. pp. 105–116 (2016)