Towards a methodology for the semi-automatic generation of scientific knowledge graphs from XML documents

George Hannah[1], Terry R. Payne[1], Valentina Tamma[1], Andrew Mitchell[2], Ellen Piercy[2] and Boris Konev[1]

[1]University of Liverpool, Foundation Building, Brownlow Hill, Liverpool, L69 7ZX, United Kingdom
[2]Materials Innovation Factory, University of Liverpool, 51 Oxford Street, Liverpool, L7 3NY, United Kingdom

Abstract

Robots used in analytical laboratories, such as those at Unilever, generate vast amounts of log data. This log data is typically stored in semi-structured formats (e.g. XML) according to some standard schema, e.g. the Analytical Information Markup Language (AnIML). Representing this data in a structured format such as a knowledge graph would allow for a more consistent data interpretation, as the relationships between concepts would be formalised in an ontology; consequently making the process of complex data analysis simpler for the scientists involved. We propose a semi-automatic pipeline that exploits the inherent structure of XML schemata, as well as previously represented domain knowledge, to create a knowledge graph that represents log data with its relevant metadata. We utilise ontology alignment techniques to identify related concepts in different ontologies, and therefore provide additional context when predicting the property linking two classes while building the graph.

Keywords

Ontology alignment, knowledge graph generation, XML,

1. Introduction

As technology progresses, we are producing rapidly increasing amounts of data. This is particularly true in science, with the introduction of laboratory robots. These robots consistently carry out repeated actions faster, and with more accuracy than a human scientist; resulting in a large volume of log data. This data is often stored in semi-structured formats such as XML, due to the need to store this data in a machine readable and actionable format, thus granting scientists access to advanced data analysis techniques [1]. Knowledge graphs (KGs) are data structures that can support advanced data analytics by semantically representing entities (as vertices) and the relationships between them (as directed edges) in a graph or semantic network.

This research is motivated by the work carried out by Unilever, where several robots are used in their laboratories to carry out scientific experiments. These experiments involve the

OM 2023: The 18th International Workshop on Ontology Matching collocated with the 22nd International Semantic Web Conference ISWC-2023 November 7th, 2023, Athens, Greece

g.t.hannah@liverpool.ac.uk (G. Hannah); trp@liverpool.ac.uk (T. R. Payne); valli@liverpool.ac.uk (V. Tamma); Andrew.Mitchell@unilever.com (A. Mitchell); ellen.piercy@unilever.com (E. Piercy); konev@liverpool.ac.uk (B. Konev)
0000-0002-3218-4559 (G. Hannah); 0000-0002-0106-8731 (T. R. Payne); 0000-0002-1320-610X (V. Tamma)

testing and formulation of many of their different products. During these experiments, log data is generated by the robots and stored in AnIML (Analytical Information Markup Language)[1] files. AnIML is an XML standard that has been designed for analytical chemical and biological data and processes. In addition to the log data, this representation can also define metadata surrounding experiments such as: the person that started the experiment, or the experiment start time. Such metadata provides additional context, and if exploited correctly can improve the semantic representation of the data. The structure of any given XML file is determined by its schema, and the same is true in AnIML. These schemata state what attributes and children a given element can have and the required cardinality of those attributes and child elements. We hypothesise that the structure provided by the AnIML schema can be leveraged to bootstrap an ontology that correctly models the knowledge in AnIML files. In a KG, the relationships between concepts are formalised by an ontology. However, creating an ontology to describe a domain can be both labour and time intensive, as the knowledge that is held by a domain expert has to be correctly represented by an ontology engineer. Thus, we propose a semi-automatic pipeline for creating a KG and corresponding ontology from semi-structured data in XML, with the aim of reducing the time required to create and maintain a KG. We exploit the implicit semantics found in XML documents, which follow a hierarchical structure consisting of two types of relationship, parent-child, and element-attribute. These relationships identify a link between two entities; however the specific nature of these links are unknown. Our aim is to align concepts in our ontology with equivalent concepts in other ontologies, and then use this additional context to assist in the prediction of the nature of these unknown relationships.

2. Related work

The use of KGs across many domains has increased over the past few years with the rise in popularity of different machine learning (ML) models. As discussed in [1], there is a desire to explain the decision process followed by an AI model. Since KGs are structured, they can be used to explain the data they model in a consistent way, that is readable by humans as well as being machine processable. Businesses such as Unilever are now investing in solutions that support the representation of their data in KGs to enhance AI powered data analysis [2].

The reuse of concepts from other ontologies is both a key step in many ontology development methodologies [3, 4] and in ontology alignment [5]. One possible approach to support reuse consists of collecting a set of terms and utilising the NCBO recommender [6] to recommend a set of candidate ontologies. Some examples of ontologies in the relevant domains are: Nanomine [7], SIO (Semanticscience Integrated Ontology) [8], and ChEBI (Chemical Entities of Biological Interest) [9]. However, often ontology recommendation engines rely mainly on the syntactic similarity between class and property labels, thus potentially missing good matching candidates. Furthermore, some of the ontological entities can be modelled differently; for example, all of the compounds in ChEBI are represented as classes, whereas, in other cases these compounds may be represented as instances of a generic class like “Chemical Compound”.

Whilst these ontologies have differing target domains, there is still some overlap between them as they describe general concepts. This raises the issue of the terminological heterogeneity in

[1]Analytical Information Markup Language (AnIML): https://www.animl.org/

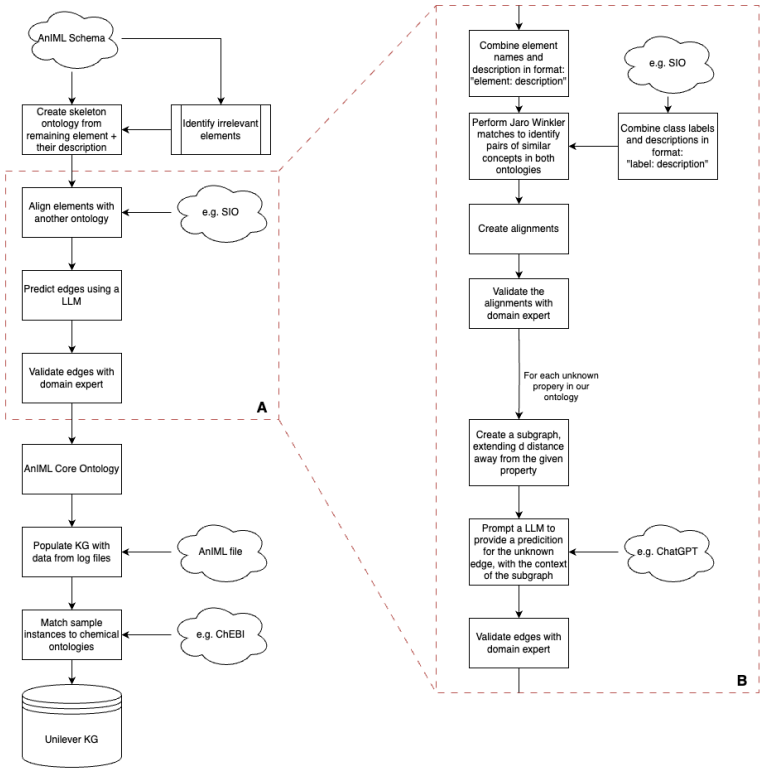

Figure 1: The AnIML2KG pipeline: the overall pipeline is shown on the left; Section A is expanded on the right (Section B) to illustrate the use of ontology alignment in the knowledge capture stage.

ontologies, as the same concepts may be described in different ontologies in different ways [10]. This can be particularly problematic when aligning ontologies, as alignment algorithms may struggle to identify that two concepts are related when they represented in different ways, leading to a low precision and recall [11].

A key aspect of this work is the extraction of knowledge from semi-structured data sources. We focus on semi-structured data in XML, utilising mapping languages to represent the knowledge in the XML files [12]; however, this is not the only semi-structured data format. The SemTab challenge [13] provides a framework to evaluate the effectiveness of different methodologies on the task of matching tabular data to a KG. Similar to this work, matching tabular data to a KG allows the implicit semantics of the relationships between columns in a table to be formalised. SemTab utilises a pre-existing KG to match the data, therefore deriving its semantics [13]. Although our approach focuses on the generation of a KG directly from the XML data, it is related to the SemTab challenge as use the knowledge captured in other ontologies to validate the relationships extracted from the XML schema.

3. Knowledge graph generation pipeline

We propose the pipeline shown in Figure 1 to transform the log data contained within an AnIML file to a KG. We begin by generating a *skeleton ontology*; i.e. an ontology where the true nature of the edges is not known. This involves extracting elements from the AnIML core schema[2] and the AnIML technique definition schema,[3] discarding the elements in the schema that are unnecessary for our use case. In AnIML, there are several container elements such as `SampleSet` and `ExperimentStepSet` that provide no experiment specific information. By ignoring these elements, we can extract the relevant information from the AnIML files for a simpler representation within the KG. The remaining concepts are then represented in the graph, with edges connecting concepts being taken from the structure of the AnIML schema. The names of these properties are hypothesised by the adding the "has" prefix to the name of the child concept, or attribute, whilst the true nature of the property is unknown. For example the property connecting `ani:Sample` and `ani:SampleID` would be `ani:hasSampleID`.

The following steps are shown in section A of Figure 1, and are expanded in more detail in section B. We combine the names of concepts and their descriptions into a single string from both our skeleton ontology and the ontology we wish to align to, in this case SIO [8]. We then identify conceptual matches between the two ontologies. For example, comparing `ani:role` with `sio:SIO_000016` (which has the label "role") would result in a similarity score of 77.6%. Through experimentation we can fine tune a threshold of acceptance for a correct match. In Figure 1, the Jaro-Winkler string distance metric is used to find matches as it places value on a shared prefix [14]. By combining the concept name and description into a single string with the format "`name, description`", concepts that share a name are more likely to be identified, whilst not disregarding concepts that have different names but similar descriptions. Once a set of alignments have been established, they are validated by hand as automatic alignment algorithms may struggle to achieve high precision and recall [11].

Using the context provided by the alignments, we predict the semantics of the unknown properties in our ontology. This is done by selecting a triple with an unknown property, and creating a subgraph from our ontology containing all triples with a distance $d \in [0, \dots, 1]$ away from the subject of our selected triple, thus capturing the context surrounding this triple, in the form of related concepts from both our skeleton ontology and the aligned ontology. For example, to predict the true nature of `ani:hasRole`, the subgraph would include triples such as "`ani:Sample ani:hasRole ani:role .`", "`ani:Sample ani:hasSampleID ani:sampleID .`", and "`ani:role skos:broadMatch sio:SIO_000016 .`" The subgraph can be used to form a prompt for an LLM such as ChatGPT,[4] requesting a prediction of the predicate in our selected triple, which is validated by a domain expert to finalise the AnIML Core Ontology.

We can now consider the AnIML log files. We extract the data from the files and populate the KG, assigning the data as an instance of the class related to the element or attribute that the data came from [12]. At this point we can enrich the knowledge in the graph by aligning to other ontologies. In our case we initially consider chemical ontologies such as ChEBI [9], as it provides access to a large dataset that can be used to train machine learning models for

[2] https://github.com/AnIML/schemas/blob/master/animl-core.xsd
[3] https://github.com/AnIML/schemas/blob/master/animl-technique.xsd
[4] https://chat.openai.com/

tasks such as chemical to chemical reaction prediction [15]. This enriched KG will support Unilever's scientists in their data analysis. To compute these alignments we will match the chemical formula of compounds found in AnIML files to those found in ChEBI.

4. Discussion and conclusions

The pipeline illustrated in Figure 1 provides a novel method for semi-automatically creating an ontology from an XML schema. This contribution has the potential to increase the rate in which data is translated into semantic formats. There is vast amount of data across the web stored in XML based formats like AnIML, so processes for consistently converting the schemata describing this data into ontologies can eliminate one of the issues preventing the mass-adoption of semantic web technologies in the analytical domain [16]. The pipeline consists of two stages that involves the human validation of alignments and predictions. These stages can be used to quantitatively evaluate the effectiveness of the alignment and prediction methods respectively, by having the evaluator record the results and calculate the precision, recall, and F-score of the approach, which can be used to compare with other methods. Another possible evaluation metric is the KG's ability to answer a set of competency questions defined by the domain experts in our industrial use case (although this qualitative evaluation method is use-case specific).

Although our pipeline was designed with a very specific use case in mind, it can be generalised with limited effort to other domains, provided that the data is expressed in a semi-structured format with an explicit schema. An additional requirement is the existence of some ontological coverage or overlap, so that the knowledge extracted from the schema can be validated.

When designing this pipeline, we decided not to re-use any specific ontology or extend an ontology to cover the AnIML schema. The motivation for this decision was our use case. As all of the data generated in the experiments is the intellectual property of Unilever, some data may be expressed in a non-standard way. In these cases, extending pre-existing ontologies to cover these proprietary concepts may become difficult, as the way they relate to other concepts may be overly complex. Instead, the creation of a specific ontology for Unilever that is supported by other ontologies appears to be a better alternative, as the relationship between proprietary concepts and the standard concepts will be simpler and the ontology will be smaller. However, evaluating the differences between these two approaches is an avenue for future work.

Acknowledgements

This work has been funded by an EPSRC ICASE studentship, 201146 with Unilevel PLC.

References

[1] I. Tiddi, S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey, Artificial Intelligence 302 (2022) 103627.

[2] D. Zhou, B. Zhou, Z. Zheng, E. V. Kostylev, G. Cheng, E. Jimenez-Ruiz, A. Soylu, E. Kharlamov, Enhancing knowledge graph generation with ontology reshaping - bosch case, in: European Semantic Web Conference, Springer, 2022, pp. 299–302.

[3] N. F. Noy, D. L. McGuinness, Ontology development 101: A guide to creating your first ontology, Technical Report, Stanford knowledge systems laboratory technical report KSL-01-05, 2001.
[4] M. Uschold, M. King, Towards a methodology for building ontologies, in: Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence, 1995.
[5] M. Granitzer, V. Sabol, K. W. Onn, D. Lukose, K. Tochtermann, Ontology alignment—a survey with focus on visually supported semi-automatic techniques, Future Internet 2 (2010) 238–258.
[6] M. Martínez-Romero, C. Jonquet, M. J. O'connor, J. Graybeal, A. Pazos, M. A. Musen, Ncbo ontology recommender 2.0: an enhanced approach for biomedical ontology recommendation, Journal of biomedical semantics 8 (2017) 1–22.
[7] J. P. McCusker, N. Keshan, S. Rashid, M. Deagen, C. Brinson, D. L. McGuinness, Nanomine: A knowledge graph for nanocomposite materials science, in: The Semantic Web-ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II 19, Springer, 2020, pp. 144–159.
[8] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, et al., The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery, J. of biomedical semantics 5 (2014) 1–11.
[9] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, Chebi: a database and ontology for chemical entities of biological interest, Nucleic acids research 36 (2007) D344–D350.
[10] P. Mitra, G. Wiederhold, Resolving terminological heterogeneity in ontologies, in: Proceedings of the ECAI workshop on Ontologies and Semantic Interoperability, 2002, pp. 45–50.
[11] Z. Dragisic, V. Ivanova, P. Lambrix, D. Faria, E. Jiménez-Ruiz, C. Pesquita, User validation in ontology alignment, in: 15th International Semantic Web Conference , Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15, Springer, 2016, pp. 200–217.
[12] A. C. Junior, F. Orlandi, D. O'Sullivan, C. Dirschl, Q. Reul, Using mapping languages for building legal knowledge graphs from xml files, in: 2nd International Contextualized Knowledge Graphs Workshop (CKG'19) at the 18th International Semantic Web Conference, 2019. ArXiv preprint arXiv:1911.07673.
[13] N. Abdelmageed, J. Chen, V. Cutrona, V. Efthymiou, O. Hassanzadeh, M. Hulsebos, E. Jiménez-Ruiz, J. Sequeda, K. Srinivas, Results of semtab 2022, Semantic Web Challenge on Tabular Data to Knowledge Graph Matching 3320 (2022).
[14] M. Cheatham, P. Hitzler, String similarity metrics for ontology alignment, in: The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12, Springer, 2013, pp. 294–309.
[15] R. T. Sousa, S. Silva, C. Pesquita, The supervised semantic similarity toolkit, in: European Semantic Web Conference, Springer, 2022, pp. 42–46.
[16] P. Hitzler, A review of the semantic web field, Communications of the ACM 64 (2021) 76–83.