Assessing Schema.org's Coverage of Terms from Key Biomedical Datasets

Kody Moodley[2,3], Josef Hardi[1,3], John Graybeal[1], Mark A. Musen[1], Michel Dumontier[2]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, USA
{johardi, jgraybeal, musen}@stanford.edu
[2]Institute of Data Science, Maastricht University, The Netherlands
{kody.moodley, michel.dumontier}@maastrichtuniversity.nl
[3]These authors contributed equally to this work

ABSTRACT

Motivation: *Schema.org* is an initiative by major Web search engines to define a common vocabulary for structuring Web content from a variety of domains, promoting data interoperability and enabling Web content to benefit from sophisticated search services. Within the wide spectrum of *schema.org* vocabulary, there are specialized data attributes for biomedical objects. Before leveraging these attributes to mark up the actual data, it is valuable for biomedical data publishers to know which of their key data fields can be captured by *schema.org*. There are currently no quantitative evaluations to measure how much of *schema.org* vocabulary aligns with the accepted standards in biomedical domains. In this paper, we provide such an evaluation against selected biomedical standards for drugs, clinical trials and medical datasets.

1 INTRODUCTION

Schema.org [1] is a common vocabulary for Web content from a multitude of domains. Managed by the search engines Google, Bing, Yahoo and Yandex, its goal is to improve the indexing of Web pages in order to facilitate the development of more sophisticated search services. Web developers are incentivised to markup their content using *schema.org* to make it uniformly queryable across the Web, potentially being rewarded with prioritization in search results, appealing presentation to end-users (e.g. through Google's rich snippets [2]) and benefit from future search services to be developed.

The *schema.org* vocabulary has an extension for biomedical data called *health-lifesci*. We have recognized the potential of *schema.org* health-lifesci extension to provide a common terminology for integrating and querying drug and clinical trials data. The discoverability of such data is invaluable for healthcare applications such as drug repositioning [4]. Throughout the paper, we are going to use the word *schema.org* to refer to the health-lifesci extension, unless stated otherwise.

This new set of data attributes in *schema.org* raises a question for biomedical data publishers: which data fields of my current data can already be described by *schema.org*? While there have been attempts from a few data publishers to propose a mapping between *schema.org* and their metadata specification, there are no quantitative studies to assess the degree of alignment between *schema.org* and the respected metadata specifications.

We fill this gap by presenting a quantitative evaluation using schema mapping measures and methodologies we have implemented. We used a peer-reviewed method to cross-check the mapping of *schema.org*'s types and attributes against a selection of well-known metadata standards for drugs, clinical trials and datasets. We then apply two metrics for assessing the overall alignment quality by measuring the *compatibility* and *coverage* of terminology offered by *schema.org*. We believe the results would be of value for demonstrating how *schema.org* can already be used to represent a substantial number of biomedical metadata attributes and help to identify missing important attributes in *schema.org*.

2 RELATED WORK

The Bioschemas [3] community is a prominent research initiative which focuses on improving the interoperability of data in the life sciences. The community has recognized the potential of *schema.org* to provide a common terminology for describing biomedical data and they actively encourage data publishers to markup their content using *schema.org*. They also routinely propose inclusion of missing types and attributes for the *health-lifesci* extension. In this connection, the community develops recommendations for improving descriptions of certain generic types in *schema.org*, such as `Events` and `Datasets`. The community has also initiated specifications for generic biological types such as `Samples` and `Proteins`. However, they have not yet formally investigated the capability of *schema.org* to describe medical data pertaining to drugs and clinical trials.

The *health-lifesci* extension, proposed by the Healthcare Schema Vocabulary Community Group, offers a substantial set of properties pertaining to drugs and clinical trials. However, there has not been a quantitative study to measure

the degree to which the terminology is comparable with the accepted standards, such as the DrugBank metadata or the US Code of Federal Regulations rulebook. Finally, the BIOmedical and HealthCAre Data Discovery Index Ecosystem (BioCADDIE) group has developed another specification for describing medical datasets called the DAta Tag Suite (DATS) model. The group has already proposed a mapping for *schema.org* [6] that is relevant to our work.

There has been a large body of research on generic specification matching methodologies [8, 9, 10]. Many approaches seek to fully automate or semi-automate the process, such as, by adding human judgement in the loop in the latter case to address the sheer size of some specifications. However, these approaches often do not beat human accuracy [9]. For our task in this work, we are not confronted with extremely large specifications to compare and we therefore choose to use mappings made by humans. Strategies for measuring how much information one specification covers of another, have also been studied [7, 11]. However, they have not been applied to biomedical metadata specifications.

3 MAPPING PROPOSAL AND EVALUATION

For our assessment, we focused on three biomedical objects, which are: drugs, clinical trials and medical datasets. Each object type has a variety of published data resources and often uses different metadata specifications.

3.1 Metadata Specifications

Drug: We use the DrugBank entry specification [5] and the National Drug File Reference Terminology (NDF-RT). DrugBank is a comprehensive online resource for detailed information about thousands of medical drugs and NDF-RT is a centrally maintained electronic drug list used by the US Veterans Health Administration (VHA) medical facilities. **Clinical Trial:** we choose a single comprehensive and representative clinical trials specification namely CFR Title 42(11): Clinical Trials Registration and Results Information Submission which is used by ClinicalTrials.gov site. **Dataset:** For biomedical datasets on the Web, we focused on specifications produced by Bioschemas, BioCADDIE and HCLS groups. We have mentioned earlier about Bioschemas and BioCADDIE, and HCLS is a Semantic Web Interest Group that promotes the use of Semantic Web technology in life sciences.

3.2 Mapping Methodology

For each specification, our first task is to find a mapping from each attribute in *schema.org* to a data field belonging to the target metadata specification that can be justified as a potential data conversion, for example the *schema.org* `datePublished` can be mapped to `dct:issued` in the HCLS specification. We then take into account the *requirement levels*[1] imposed by the metadata specification. These requirement levels indicate the necessity of a value to be present in a data field, whether it is *required*, *recommended*, or *optional* and technically they are denoted by the keywords: MUST, SHOULD or MAY, respectively. The second task is to quantitatively measure the degree of alignment of *schema.org* against the metadata specification and its requirement levels, where we will apply two metrics called the *compatibility rating* and *coverage ratio*.

3.2.1 Compatibility Rating

The *compatibility rating* is used to evaluate the degree of alignment of the two schemas by prioritizing the matches of the key attributes by looking at the requirement levels of the data fields in the metadata specification. The idea is to place a high importance on matching the *required* attributes, lesser importance on the *recommended* attributes and even lesser importance on the *optional* ones.

A natural way to apply this is to assign appropriate weights to the required, recommended and optional attribute matches. However, we have to avoid biases in picking the weights and therefore we identified a principled method for calculating the weights, known as the *prioritized aggregation model* proposed by Ronald R. Yager [7]. We use the adopted model by Pereira, et. al. [12] for defining prioritized criteria and computing the weights. In our case, we define our prioritized aggregation model for *compatibility scoring*, $S_{compatibility}$, as follows:

$$S_{compatibility} = \sum_{i}^{n} w_i.C_i(r)$$

where $C = \{C_1, ..., C_n\}$ be a set of n prioritized criteria and w_i is the weight of each criterion C_i given a metadata specification r. For our evaluation, each individual criteria C_i represents the *local* compatibility of *schema.org*'s data attributes against the metadata specification's data fields grouped by their requirement level. We name each group as *compliance, extension* and *accessory*, with respect to the MUST, SHOULD or MAY keyword, and formally define the labelling C_1*: compliance*, C_2*: extension*, C_3*: accessory* and order them according to the importance:

$$compliance \succ extension \succ accessory$$

This ordering lines up with the idea mentioned earlier, that we want to give the most weight to the alignment of the

[1] https://tools.ietf.org/html/rfc2119

required fields, the least weight for the optional fields and the inbetween weight for the recommended fields.

Measuring the *compliance* is done by first finding a match for every **MUST** field in the metadata specification to a *schema.org* attribute, assigning the *matching score* and then dividing the total matching score with the total number of the required fields. We assign a matching score based on a human observation that looks for similarities from reading and comparing the field names, descriptions, prescribed domains and expected data types. When data instances of the metadata specification are accessible, we reinforce the match by judging the field's value [8]. There are three possible matching scores: 1, 0.5, or 0, representing an *exact match, partial match* or *no match*, respectively, where the person doing the evaluation must assign one of the values to all the fields. Formally, we write the compliance measurement as follows:

$$compliance = \frac{\sum_{i}^{|r_{must}|} S_{matching}(a_i)}{|r_{must}|}$$

where $S_{matching}(a_i)$ is the matching score assignment function for a given required field a_i and $|r_{must}|$ is the size of the required fields in the metadata specification. The measurement for the *extension* and *accessory* criteria have an analogous definition as the *compliance*, such that we substitute the evaluation against the **SHOULD** and **MAY** fields, respectively.

Putting the formulae together, we define the compatibility rating, $R_{compatibility}$, as the compatibility score divided by the number of criteria, presented as a percentage. To illustrate the calculation, consider an situation where we have done the mapping and determined 0.8 as the the *compliance* score; 0.5 for the *extension* and 0.6 for the *accessory*. The first step is to compute the weights for each of the criteria, such that:

$$w_1 = 1$$
$$w_2 = w_1 \cdot C_1 = 1 \cdot 0.8 = 0.8$$
$$w_3 = w_2 \cdot C_2 = 0.8 \cdot 0.5 = 0.4$$

Next, we compute the compatibility score:

$$S_{compatibility} = (1 \cdot 0.8) + (0.8 \cdot 0.5) + (0.4 \cdot 0.6) = 1.44$$

And finally, we get the compatibility rating of 48% for our example of situation:

$$R_{compatibility} = \frac{1.44}{3} \cdot 100 = 48$$

3.2.2 Coverage Ratio

The *coverage ratio* is a measure to compute the ratio between the successfully matching attributes (either as an *exact match* or a *partial match*) and the specification size, i.e., the total number of fields in the metadata specification.

$$coverage = \frac{total\ number\ of\ matches}{specification\ size} \cdot 100$$

The coverage ratio gives a good overall impression of the proportion of attributes in *schema.org* that can be successfully mapped to the data fields in the metadata specification. We can also break down the calculation of coverage ratio per requirement level by changing the parameter *size* to include only the fields from a specific requirement level, e.g., MUST, SHOULD or MAY.

3.3 Mapping Process

Two researchers were assigned to produce the mapping. Each researcher independently proposed a *schema.org* mapping for each metadata specification mentioned in Section 3.2. The result is two potentially different mapping proposals per specification where they were recorded in publicly available spreadsheets[2,3,4]. After the task for proposing the mappings are concluded, the two researchers then jointly discussed the discrepancies and arrived at a single mapping for each specification via consensus.

4 RESULTS

Overall, our evaluation revealed that *schema.org* can capture a significant number of drug and dataset metadata fields and there is particularly good coverage overall for required fields. However, it has low coverage of detailed properties in medical trials. A summary is given in Table 1:

Table 1. Overall *schema.org* compatibility ratings and coverage ratios for our mappings from all specifications in this study. Abbreviations: Bioschemas (BIO), BioCADDIE DATS (DTS), HCLS summary level (Hs), HCLS version level (Hv), HCLS distribution level (Hd), DrugBank (DB), VA National Drug File (NDF) and ClinicalTrials.gov Protocol Registration (CFR).

Measure	Dataset					Drug		CT
	BIO	DTS	Hs	Hv	Hd	DB	NDF	CFR
Compatibility	100	22	79	66	60	50	46	14
Coverage	100	56	85	77	60	63	57	21
- MUST coverage	100	43	100	83	100	80	100	30
- SHOULD coverage	100	67	75	80	53	61	30	n/a
- MAY coverage	n/a	52	85	73	53	56	57	14

Schema.org is expressive enough, overall, for representing medical dataset metadata. It matches perfectly with the Bioschemas Dataset Metadata (BIO) due to the close resemblance to the *schema.org* attributes. We found a mixed compatibility between *schema.org* and the HCLS dataset

[2] https://tinyurl.com/ybcrcgbe
[3] https://tinyurl.com/y8a8env9
[4] https://tinyurl.com/ycdgr4n8

specification depending on the choice of the detail level. The scores for the high-level summary presentation (Hs) are higher than the more detailed cataloging specifications (Hv) and (Hd). However, *schema.org* has poor coverage of fields from the BioCADDIE DATS model (DTS) due to the model being very rich and detailed. Our mapping shows that *schema.org* tends to use simple data types (e.g., text, date) rather than complex objects to store values which significantly decreases the compatibility.

When it comes to describing drug data, *schema.org* has moderate compatibility with both DrugBank and NDF-RT with a slightly better total coverage ratio for DrugBank. Lastly, our evaluation results suggest that *schema.org* is not ready for capturing clinical trials metadata. Looking closer at the mappings, it is apparent that *schema.org* misses a fair number of important fields such as: study design, outcome measures, important dates and contact details about the participating groups or sponsor organizations.

Our evaluation provides a quantified scale for *schema.org* developers to look at the current state of the vocabulary in comparison to the accepted standards in the focus area. Our results should also give a hint to the data developers, especially those who are working with the specifications used in this study, to limit their expectation when marking up their data with *schema.org*.

Another contribution to consider in this paper is the scoring method used in the study. We have developed a scoring mechanism that takes into account the field's requirement levels and its matching quality as parameters to create a sensible compatibility grading. For a decision-making application, the rating can be a tool to gauge the completeness of converting data from one specification to another, early, before the whole data transformation process begins.

5 CONCLUSIONS

We have presented a brief assessment of *schema.org*'s coverage against some prominent metadata specifications for drugs, clinical trials and medical datasets. We have also described a method for quantifying the matching assessment.

There are two key takeaways from this study. The first is our analysis has showed that *schema.org* is already good enough to capture the key fields of the respected metadata specifications with the exception of clinical trials. Including more of the recommended fields into *schema.org* would bring the vocabulary closer to the actual application and thus inviting more users to adopt it. The second is our scoring method should be a good proxy for the *schema.org* developers to prioritize the attribute recommendation. We are intrigued to do a comparison with other scoring methods and evaluate which performs better.

Our mappings are open and available to be used and reused for any interested parties. In the future, we plan to use these mappings to automate the marking up of existing drug and clinical trials metadata, and to demonstrate the benefits of using *schema.org* markups by developing a software system that features semantic facets.

6 ACKNOWLEDGEMENTS

This research was supported by the National Institute of Health (NIH) under grants 5U54AI117925-04 and 5U54AI117925.

7 REFERENCES

1. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of structured data on the web. Communications of the ACM 59.2, 44-51, (2016)
2. van der Meer, J., Boon, F., Hogenboom, F., Frasincar, F., Kaymak, U.: A framework for automatic annotation of web pages using the Google rich snippets vocabulary. In Proceedings of the 2011 ACM Symposium on Applied Computing. ACM, New York, USA, 765-772, (2011)
3. Gray, A.J.G., Goble, C., Jimenez, R. C.: Bioschemas: From Potato Salad to Protein Annotation. International Semantic Web Conference Poster Proceedings (2017)
4. Ashburn, T. T., Thor, K. B.: Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug discovery 3, no. 8, 673-683, (2004)
5. Wishart, D. S., Knox, C. Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 36, no. suppl_1, 901-906, (2007)
6. Susanna-Assunta, S., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J. S., Xu, H., Fore, I. M. et al.: DATS, the data tag suite to enable discoverability of datasets. Scientific Data 4 (2017)
7. Yager, R. R.: Prioritized aggregation operators, In: International Journal of Approximate Reasoning, Volume 48, Issue 1, 263-274, (2008)
8. Rahm, E., Bernstein, P. A.: A survey of approaches to automatic schema matching. The VLDB Journal 10, no. 4, 334-350, (2001)
9. Bernstein, P. A., Madhavan, J., & Rahm, E. (2011). Generic schema matching, ten years later. Proceedings of the VLDB Endowment, 4(11), 695-701.
10. Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. the VLDB Journal, 10(4), 334-350.
11. Bellahsene, Z., Bonifati, A., Duchateau, F., & Velegrakis, Y. (2011). On evaluating schema matching and mapping. In Schema matching and mapping (pp. 253-291). Springer Berlin Heidelberg.
12. da Costa Pereira, C., Dragoni, M., Pasi, G.: Multidimensional relevance: Prioritized aggregation in a personalized Information Retrieval setting. Information processing & management, 48(2), 340-357, (2012)