Metadata standards for the FAIR sharing of vector embeddings in Biomedicine

Şenay Kafkas [1,*], Remzi Çelebi [2,*], Mehdi Ali [3,4], Hajira Jabeen [3], Michel Dumontier [2],Robert Hoehndorf [1]

[1]Computer, Electrical and Mathematical Sciences & Engineering with Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955, Kingdom of Saudi Arabia
[2]Institute of Data Science, Maastricht University Maastricht, The Netherlands
[3]Smart Data Analytics Group, University of Bonn, Germany
[4]Department of Enterprise Information Systems, Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Sankt Augustin and Dresden, Germany

ABSTRACT

Motivation: Today, we have an enormous amount of biomedical data and its size, as well as complexity, have been increasing over time. Implementation of standards represents one of the key drivers in the life sciences research as well as the technology transfer. More specifically, standards enable data accessibility, sharing, integration and therefore facilitates data harnessing and accelerates research and innovation transfer. The life sciences community has widely developed and used Semantic web technology standards for data representation and sharing. However, given the success of unsupervised machine learning methods such as Word2Vec and BERT, there is a need to develop new standards for sharing the (pre-trained) vector space embeddings of the entities to facilitate reusability of data and method development. Motivated by this, we propose data and metadata standards for the FAIR distribution of vector embeddings and demonstrate utilization of these standards in Bio2Vec, a platform providing a flexible, reliable and standard-compliant data representation, sharing, integration and analysis.

Availability: The proposed metadata standard and an example are available in the ShEx format at Zenodo.

Contact: senay.kafkas@kaust.edu.sa, remzi.celebi@maastrichtuniversity.nl

1 INTRODUCTION

Recent technological advancements in computing and availability of big datasets have resulted in a significant number of studies which focus on developing and using machine and deep-learning methods. Particularly, unsupervised techniques like learning vector embeddings, structure preserving maps into vector space usually from high dimensional data in a lower-dimensional vector, have gained a lot of attention. These embeddings can be generated from structured data (e.g., relational data, knowledge graphs, etc.) (Bordes *et al.* (2013); Ali *et al.* (2019)) and unstructured data (e.g., text, image etc.) (Devlin *et al.* (2018); Mikolov *et al.* (2013)). A large number of vector space embeddings generated from text or structured data is now being generated and used as part of machine learning models. Generating embeddings for a dataset can be time-consuming, and it is beneficial to make these embeddings available to be reused. Reuse requires appropriate meta-data and standards to be applied so that all necessary information that determines the embedding is available. Therefore, there is a need for the development of metadata standards for the FAIR (Findable, Accessible, Interoperable and Reusable) (Wilkinson *et al.* (2016)) sharing of these (pre-)trained vector embeddings.

To improve FAIRness of any data in the life science domain, particularly to make it more findable and interoperable, Bioschemas (Garcia *et al.* (2017)) extends schema.org to reuse some existing types in schema.org and to suggest new types and properties that are applicable in the life sciences. If data providers add such types and properties to their web pages as semantic markup, their content will be discoverable more easily through search engines and through semantic search. Here, we adopt the Bioschemas approach to describe the data and metadata for generated and sharing embeddings. We propose metadata standards for FAIR sharing of vector embeddings as well as demonstrate the utilization of these standards in Bio2Vec, which is a knowledge discovery platform providing a flexible, reliable and standard-compliant data representation for vector sharing, integration and analysis (Bio2Vec (2020)).

The Life sciences community has developed and widely used Semantic Web standards (Berners-Lee *et al.* (2001)) to represent and share data and knowledge. Data in the life sciences is increasingly being represented using the Resource Description Framework (RDF) format (Lassila *et al.* (1998)) and is made available through Semantic Web technologies as part of the Linked Data (Bizer *et al.* (2011)). On the other hand, biomedical literature is distributed in several formats. Our standards rely on the existing Semantic Web technologies and we utilized schema.org (schema.org (2020)) and ML-Schema (Correa Publio *et al.* (2018)) for the representation of data fields. The proposed metadata standard is available at Zenodo[1] in the form of the Shape Expressions (ShEx) format which describes RDF graph structures.

2 MATERIALS AND METHOD

We utilized *schema.org* to standardize the properties where the descriptions were available. We utilized ML-Schema (Correa Publio *et al.* (2018)) to describe the properties specific to the machine

[1] https://zenodo.org/record/3708487

learning method used to generate the vector embeddings. The ML-schema ontology was proposed by the W3C Machine Learning Schema Community Group. The ontology consists of classes, properties and restrictions for representing and interchanging machine learning models, datasets and experiments. Within the ML-schema ontology, there is an Run (<ML-schema:Run>) class that defines the machine learning algorithm execution. The Run class is associated with a task (achieves <Task>), an implementation (executes Implementation), parameters of implementation (hasInput HyperParameterSetting), representations of Model (hasOutput Model) and model evaluation (hasOutput ModelEvaluation). For each <ML-schema:Run>, we have ML-schema:hasInput that contains <schema:Dataset> having name, distribution, version and Url. An example can be seen in Listing 1. This example describes the embedding dataset trained on a DrugBank (Wishart *et al.* (2008)) knowledge graph using the TransE model Bordes *et al.* (2013). We gave the details of the machine learning experiment on how these embeddings were generated including the performed task (Task), hyper-parameter settings (HyperParameterSetting), the model evaluation (ModelEvaluation) and the provenance of input dataset (Dataset).

3 PROPOSED METADATA STANDARDS FOR FAIR SHARING OF VECTOR EMBEDDINGS

Table 1 describes the proposed metadata for vector embedding representation and Listing 1 gives an example of embedding metadata represented using the JSON-LD syntax. We identified minimum required information and recommended properties for the submitted embedding dataset. We reused existing schema.org properties and types whenever possible and included two new properties: embeddingSize, generatedBy. embeddingSize is used to provide the information about the embedding dimension. This property will be generated automatically based on the actual vector embedding during the data submission. The generatedBy property is used to link our embedding dataset to the machine learning experiment conducted to generate these embeddings. As a dataset grows, using whole embedding vectors would be challenging due to the size and computational limitation. We also propose a profile for sharing embedding of a single entity. Although an individual vector might not be useful alone, to be able to query a pair or group of entities from a repository may provide some insights about how similar the entities are and how they are clustered.

Table 2 lists our proposed properties for a single embedding vector. We propose the following new properties to schema.org for inclusion: embeddingEntity, xCoord and yCoord. These properties will be generated automatically based on the actual vector embedding during the data submission. While the embeddingEntity property is used to link the vector to the identifier of an entity such as the drug *Leucovorin*, the xCoord and yCoord properties can be used for visualization or clustering purposes via dimension reduction techniques such as TSNE (Maaten and Hinton (2008)).

4 CONCLUSION

We propose metadata standards for the FAIR sharing of vector embeddings in the biomedical domain. We have used the existing standards schema.org and ML-schema in addition to proposing new properties to be added to schema.org. The presented profile can bridge the gap between the data, a variety of machine learning embedding models and the embeddings created using these models. The standard assists FAIR publication of embeddings together with the metadata describing the model used to generate the embeddings alongside its parameters.

Funding: This work is supported by King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3454-01.

REFERENCES

Ali, M., Hoyt, C. T., Domingo-Fernández, D., Lehmann, J., and Jabeen, H. (2019). Biokeen: a library for learning and evaluating biological knowledge graph embeddings. *Bioinformatics*, **35**(18), 3538–3540.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific american*, **284**(5), 34–43.

Bio2Vec (Last accessed 2 Mar 2020). Bio2vec. https://bio2vec.net/.

Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Correa Publio, G., Esteves, D., Lawrynowicz, A., Panov, P., Soldatova, L., Soru, T., Vanschoren, J., and Zafar, H. (2018). Ml-schema: Exposing the semantics of machine learning with schemas and ontologies. In *Reproducibility in Machine Learning Workshop, ICML*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garcia, L., Giraldo, O., Garcia, A., and Dumontier, M. (2017). Bioschemas: schema.org for the life sciences. *Proceedings of SWAT4LS*.

Lassila, O., Swick, R. R., *et al.* (1998). Resource description framework (rdf) model and syntax specification.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(Nov), 2579–2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

schema.org (Last accessed 2 Mar 2020). schema.org. https://schema.org/.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., *et al.* (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, **3**.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, **36**(suppl_1), D901–D906.

Table 1. Metadata of data submission. Properties with * are required information from the data creator.

Property	Description	Type
schema:identifier*	ID of a dataset (that can be automatically generated)	schema:URL schema:Text
schema:name*	name of the generated embedding dataset	schema:Text
schema:description*	description of a given dataset	schema:Text
schema:creator*	names of the embedding creator	schema:Organization schema:Person
schema:contributor	List of all the contributors	schema:Person schema:Organization
schema:publisher	institute name as the publisher of a given dataset	schema:Organization
schema:keywords*	keywords used to describe the content such as the types of entities, relations used in the embedding	schema:Text
schema:citation	A citation or reference to publication of the embedding study	schema:URL
schema:license	A license under which the dataset is distributed.	schema:URL
schema:distribution	A downloadable form of the dataset	schema:DataDownload
bio2vec:embeddingSize*	vector size	schema:integer

bio2vec:generatedBy	details on the experimental design used to generate vectors	ML-Schema:Run {
		achieves<Task>
		executes<Implementation>
		hasInput <Dataset>
		hasOutput <Model >
		hasInput <HyperParameterSetting>
		{ specifiedBy <HyperParameter>}
		hasOutput <ModelEvaluation>
		{ specifiedBy <ModelMeasure>}
		}

Table 2. Proposed additional properties for single vector-embedding.

Property	Description	Type	Sample Value

embeddingEntity	biomedical entity which is represented as vector embedding	schema:MedicalEntity {
		name* <Text>
		identifier* <URL>
		additionalType <URL>
		alternateName <Text>
		sameAs <URL>
		}
			"embeddingEntity":{ "@type": "MedicalEntity", "name": "Leucovorin", "identifier": `https://bio2rdf.org/drugbank:DB00650` }

xCoord	X coordinate of the vector embedding obtained through dimension reduction by TSNE	schema:Float	"xCoord": 2.0
yCoord	Y coordinate of the vector embedding obtained through dimension reduction by TSNE	schema:Float	"yCoord": 4.0

Listing 1. An example RDF representation of the embedding metadata in JSON-LD syntax.

```
{
"@context":"http://schema.org",
"@id": "transe:0000",
"@type": "Dataset",
"name": "TE-Drugbank",
"description": "Graph embeddings for Drugbank computed
     using TransE.",
"creator": {
    "@type": "Person",
    "name":"Remzi Celebi",
    "email": "remzi.celebi@maastrichtuniversity.nl"
  },
"keywords": [
    "Protein",
    "Drugs",
    "Diseases"
  ],
"citation": "https://doi.org/10.1186/s12859-019-3284-5",
"url":"https://github.com/rcelebi/GraphEmbedding4DDI",
    "contributor": {
        "@type": "Person",
        "name":"Ali Mehdi",
        "email": "mehdi.ali@cs.uni-bonn.de"
  },
"publisher":{
    "@type": "Organization",
    "name":"DrugBank",
    "url":"https://www.drugbank.ca"
  },
"distribution": {
    "@type": "DataDownload",
    "name": "Drugbank embeding",
    "fileFormat": "text",
    "contentURL": "https://raw.githubusercontent.com/
         rcelebi/GraphEmbedding4DDI/master/vectors/
         DB_KEGG_PGK/
         Entity2Vec_cbow_200_5_5_2_500_d5_uniform.txt"
    },
"license":"https://creativecommons.org/licenses/by/4.0",
"embeddingSize": "200",
"generatedBy": {
    "@context": "http://www.w3.org/2016/10/mls#",
    "id": "run_transe_23434245435",
    "@type": "Run",
    "executes":{
        "@type": "Implementation",
        "name": "TransE"
    },
    "achieves":{
        "@type": "Task",
        "name": "Link prediction",
        "specifiedBy":{
            "@type": "EvaluationSpecification",
            "label": "Measuring AUC for DDI prediction
                 using standard ten-fold cross validation"
                 ,
             "hasPart":{
               "@type": "EvaluationProcedure",
               "label": "Ten-Fold CrossValidation"
             }
          }
    },
    "hasInput":{
        "@type": "Dataset",
        "name": "Drugbank",
        "url":"https://www.drugbank.ca",
        "version": "4.1",
        "distribution":{
            "@type": "DataDownload",
            "name": "Drugbank- Bio2RDF version",
            "version":"R4",
            "fileFormat": "nquad/gzip",
            "contentURL": "https://download.bio2rdf.org/
                 files/release/4/drugbank/drugbank.nq.gz"
          }
     },
    "hasOutput":{
        "@type": "Model",
        "name": "Trans E"
    },
    "hasOutput":{
        "@type": "ModelEvaluation",
        "hasValue": "81.3",

        "specifiedBy":{
            "@type": "EvaluationMeasure",
            "label": "AUC"
          }
    },
    "hasInput":{
        "@type": "HyperParameterSetting",
        "hasValue": "true",
        "specifiedBy":{
            "@type": "HyperParameter",
            "label": "automatic_memory_optimization"
          }
    },
    "hasInput":{
        "@type": "HyperParameterSetting",
        "hasValue": "Adam",
        "specifiedBy":{
            "@type": "HyperParameter",
            "label": "optimizer"
        }
    }
  }
}
```