Representing bioinformatics datatypes using the OntoDT ontology

Panče Panov [1]*, Larisa Soldatova [2] and Sašo Džeroski [2]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2]Brunel University London, Department of Computer Science, Kingston Lane, UB8 3PH, Uxbridge, UK

1 INTRODUCTION

Data processing is at the heart of science. Hence, the problem of data typing is an important problem that has been addressed from different aspects and in different forms. For example, the Research Data Alliance[1] (RDA), whose major goal is to speed up the international data-driven innovation and discovery by facilitating research data sharing and exchange, has identified that the problem of data typing is an important problem that deserves attention. For this purpose, the RDA formed a Data Type Registry (DTR) working group with the goal to: compile a set of use cases for datatype use and management, formulate a data model and expression for datatypes, design a functional specification for type registries, and propose a federation strategy among multiple type registries.

In data mining research it is impossible to efficiently connect parts of workflows (semi-) automatically, such as data pre-processing and data mining, perform analysis of the research results and communicate the research outputs, without machine-processable representation of datatypes and their properties. Hence, there is a need for a standardized semantically-defined and machine amenable representation of scientific datatypes to support cross-domain applications. Unfortunately, the existing representations of datatypes do not fully address such a need. To address this gap, we built an generic ontology for the representation of scientific knowledge about datatypes, named OntoDT (Panov *et al.*, 2015).

2 ONTODT: ONTOLOGY OF DATATYPES

OntoDT defines the meaning of the key entities and represents the knowledge about datatypes in a machine friendly way. The OntoDT ontology is based on the latest revised version of the ISO/IEC 11404[2] standard for datatypes. The design of the OntoDT ontology follows best practices in ontology engineering, such as the OBO Foundry principles. We used the Information Artifact Ontology[3] (IAO) to define the upper level classes and re-used existing ontological resources, such as Open Biomedical Ontologies.

The OntoDT ontology defines the basic entities (see Fig. 1), such as datatype, properties of datatypes, value space, and characterizing operations. We also define a taxonomy of datatypes. The top-level ontology classes include primitive datatypes, generated datatypes, and user defined datatypes. *Primitive datatypes* are defined by explicit specification and are independent of other datatypes. *Generated datatypes* are syntactically and semantically dependent on other datatypes, and are specified implicitly with *datatype generators*. *User defined datatypes* are defined by a datatype declaration and allow defining additional identifiers and refinements to both primitive and generated datatypes. At the lower levels, the datatypes are distinguished with respect to their datatype properties.

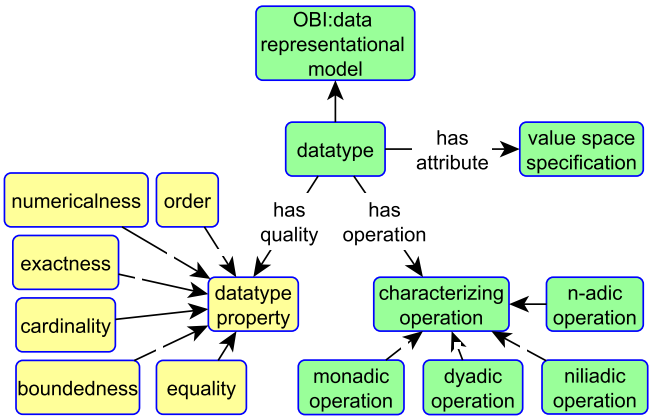

Fig. 1. Representation of datatypes in OntoDT.

OntoDT was used within an Ontology of core data mining entities for constructing taxonomies of datasets, data mining tasks, generalizations and data mining algorithms (Panov *et al.*, 2014). Furthermore, OntoDT can be used for annotation and querying machine learning dataset repositories. OntoDT can also improve the representation of datatypes in the BioXSD exchange format for basic bio-informatics types of data. The generic nature of OntoDT enables it to support a wide range of other applications, especially in combination with other domain specific ontologies: the construction of data mining workflows, annotation of software and algorithms, semantic annotation of scientific articles, etc. OntoDT is open source and is available at `http://www.ontodt.com/` and at BioPortal (`http://bioportal.bioontology.org/`).

3 BIOINFORMATICS DATATYPES

OntoDT is a generic ontology and it allows easy extensions to represent domain specific datatypes. This can be done by directly extending the OntoDT datatype taxonomy and defining the semantic meaning of the domain datatypes by linking them to the corresponding entities in other domain ontologies. For example, we can define an *amino-acid sequence datatype* as a subclass of the *character sequence datatype* class (which is a sequence datatype having characters as its base type). Its semantic meaning can be defined via the IS-ABOUT relation to the *amino acid sequence* entity

*To whom correspondence should be addressed: pance.panov@ijs.si

[1] `http://rd-alliance.org/`

[2] `http://tinyurl.com/qdua9f7`

[3] `http://tinyurl.com/nmjnlw2`

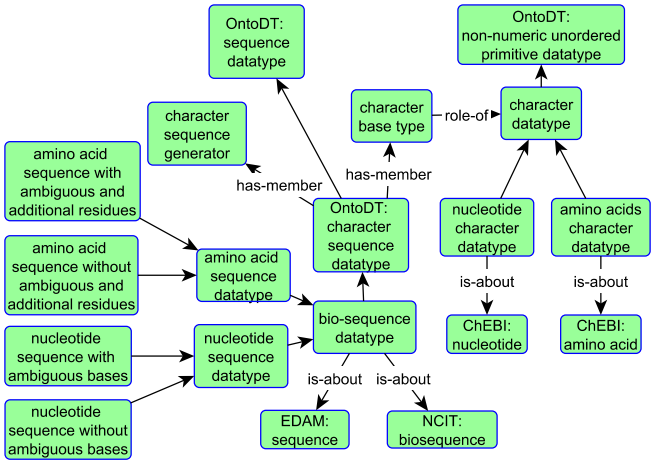

Fig. 2. Representation of bio-sequence datatype from BioXSD in OntoDT.

provided by the National Cancer Institute Thesaurus[4]. In this way, OntoDT can be used for representation of bioinformatics datatypes.

Currently, BioXSD is used to define the basic bio-informatics types of data (Kalaš *et al.*, 2010). BioXSD does not support arbitrary datatypes and it does not provide a clear framework for the representation of the semantic meaning of the data. We propose to enhance the representation of bioinformatics datatypes by exploiting the rigorous taxonomy of datatypes defined in OntoDT and the framework for the representation of semantic meanings adopted by OntoDT. OntoDT is fully interoperable with OBO bio-ontologies because it was developed by following the OBO Foundry recommendations and therefore it fully supports the representation of the semantic meaning of the data by the corresponding entities defined in domain-specific bio-ontologies.

For example, the BioXSD datatype *sequence* represents a string of 1-letter coded nucleotides or amino-acids. A *sequence record* is a datatype containing a sequence, and optionally some metadata about the sequence (for the purpose of identification). The semantic meanings of the terms *sequence* and *nucleotide* are curtail for the capturing of the semantic meaning of the data of the datatype *sequence*. However, the *sequence* datatype is not explicitly linked to the classes *nucleotide* and *amino-acid* defined in the ChEBI ontology[5], recommended by OBO Foundry as a reference ontology.

In Fig. 2, we present the extension of OntoDT to represent the bio-sequence datatype from BioXSD. We represent the *bio-sequence datatype* class as a subclass of the *character sequence datatype* class with the defined semantic meaning in the NCI Thesaurus and the EDAM ontology[6]. In order to define the *nucleotide* and *amino acid sequences datatypes*, we define two subclasses of the *character datatype* class: *nucleotide character datatype* and *amino acid character datatype*. In order to define their semantic meaning, we explicitly link them to the *nucleotide* and *amino acid* classes from the ChEBI ontology. Consequently, the *bio-sequence datatype* class has two subclasses: *nucleotide sequence datatype* and *amino acid sequence datatype*. Furthermore, both datatypes have two subclasses, depending on whether they include ambiguous bases (in the case of nucleotides) or ambiguous and additional residues (in the case of amino acids). For example, the *nucleotide sequence datatype* class has two subclasses: *nucleotide sequence with ambiguous bases* (general nucleotide sequence in BioXSD) and *nucleotide sequence without ambiguous bases* (nucleotide sequence in BioXSD).

In a simmilar way, we represent the *bio-sequence record datatype* class as a subclass of the *record datatype* class. This datatype is defined by a *record generator* and the *bio-sequence-field-list*. As defined in BioXSD, the datatype contains a bio-sequence as a mandatory component and a set of metadata (such as name, note, species, translationalData, reference, inlineBaseQuality) as non-mandatory components. In OntoDT, we model the *bio-sequence field component* class is as a role of the bio-sequence datatype.

BioXSD uses a combined approach of a pure XML Schema annotated by a data-type ontology using Semantic Annotations for Web Services Description Language[7] (WSDL) and XML Schema. SAWSDL defines a set of extension attributes for the WSDL and XML Schema definition languages. Application of attributes allows the description of additional semantics by using references to conceptual semantic models, e.g., ontologies. BioXSD datatypes are annotated with terms from the EDAM ontology Ison *et al.* (2013) using SAWSDL. In the same way, BioXSD datatypes can be annotated with OntoDT terms. For example, by annotating the datatype bio-sequence record from BioXSD with terms from the OntoDT ontology, the web services using this format would have the information that bio-sequence record is in fact a record datatype that is heterogeneous and has components, its values are unordered, it has fixed size, and each component can be accessed by keying.

4 CONCLUSION

The use case presented in this extended abstract demonstrates that OntoDT provides logically consistent representation of bioinformatics datatypes from BioXSD and enables an accurate representation of the semantic meanings of the data of specified datatypes. OntoDT has been designed as a generic and comprehensive ontology of datatypes and consequently any datatype from other resources can also be represented by OntoDT. We suggest that OntoDT can serve as a reference model for the consistent representation of datatypes used within biomedical domains and wider.

REFERENCES

Ison, J., Kalas, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., and Rice, P. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, **29**(10), 1325–1332.

Kalaš, M., Puntervoll, P., Joseph, A., Bartaševičiute, E., Topfer, A., Venkataraman, P., Pettifer, S., Bryne, J. C., Ison, J., Blanchet, C., Rapacki, K., and Jonassen, I. (2010). BioXSD: the common data-exchange format for everyday bioinformatics web services. *Bioinformatics*, **26**(18), i540–i546.

Panov, P., Soldatova, L., and Džeroski, S. (2014). Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, **28**(5–6), 1222–1265.

Panov, P., Soldatova, L., and Džeroski, S. (2015). Generic ontology of datatypes. *Information Sciences*. (accepted for publication).

[4] http://ncit.nci.nih.gov/

[5] http://www.ebi.ac.uk/chebi/

[6] http://edamontology.org/

[7] http://www.w3.org/TR/sawsdl