Information Sciences 000 (2015) 1–21

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Generic ontology of datatypes

Panče Panov [a,*], Larisa N. Soldatova [d], Sašo Džeroski [a,b,c]

[a] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Jamova cesta 39, Ljubljana, Slovenia
[c] Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, Jamova cesta 39, Ljubljana, Slovenia
[d] Department of Computer Science, Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, United Kingdom

ARTICLE INFO

Article history:
Received 15 April 2014
Revised 1 July 2015
Accepted 9 August 2015
Available online xxx

Keywords:
Data type
Data mining
Ontology
Knowledge representation

ABSTRACT

We present OntoDT, a generic ontology for the representation of scientific knowledge about datatypes. OntoDT defines basic entities, such as datatype, properties of datatypes, specifications, characterizing operations, and a datatype taxonomy. We demonstrate the utility of OntoDT on several use cases. OntoDT was used within an Ontology of core data mining entities for constructing taxonomies of datasets, data mining tasks, generalizations and data mining algorithms. Furthermore, we show how OntoDT can be used to annotate and query dataset repositories. We also show how OntoDT can 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.

© 2015 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY license
(http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Data processing is at the heart of science. Scientific research workflows rely heavily on datatype representations. Especially in data mining research it is impossible to efficiently (semi-) automatically connect parts of workflows, such as data preprocessing and data mining, perform analysis of the research results and communicate the research outputs, without machine processable representation of datatypes and their properties. 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.

In the literature, there exist different definitions of datatypes. In computer science, a datatype is usually defined as a "classification that identifies various types of data, such as boolean, integer, discrete and others, that determines the possible values for that type, operations on the values of the data, and the way the values of that type can be stored" [56]. Nell and Walker [8] discuss the difference between a data structure and datatype in the sense that "data structure refers to the study of data and how to represent data objects within a program; that is, the implementation of structured relationships" while a datatype defines "the properties of classes of objects in addition to how these objects might be represented in a program". Martin [36] also

* Corresponding author. Tel.: +386 1477 3307.
E-mail addresses: pance.panov@ijs.si (P. Panov), larisa.soldatova@brunel.ac.uk (L.N. Soldatova), saso.dzeroski@ijs.si (S. Džeroski).

http://dx.doi.org/10.1016/j.ins.2015.08.006
0020-0255/© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

discusses the difference between data structures and datatypes and states that "depending on the point of view, a data object is characterized by its type (for the user) or by its structure (for the implementer)".

In this paper, we present OntoDT, a generic ontology of datatypes. OntoDT defines the semantics, i.e., 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 standard for datatypes [23].

This paper is organized as follows. In Section 2, we present the background related to the development of the OntoDT ontology. In Section 3, we review and discuss the related work. Next, in Section 4, we present the ontology design principles and implementation, and in Section 5 we present the key OntoDT classes. In Section 6, we present the OntoDT datatype taxonomy. Finally, we present the ontology evaluation (Section 7), and three use cases of the ontology (Section 8). We conclude the paper with a discussion (Section 9) and a summary of contributions and points for further work (Section 10).

2. Background

The OntoDT development started within the frame of an ontology for data mining (OntoDM) [44]. The main idea of using a formalized description of datatypes for the domain of data mining was to characterize the types of data contained in a dataset, the applicability of a data mining task on data from a given datatype, and the applicability of a data mining algorithm on a dataset. Due to generality and reuse purposes, OntoDT has evolved to become an independent ontology.

The OntoDT ontology aims to address the need for a machine-friendly standard representation of general-purpose datatypes. It is based on the International Standard ISO/IEC 11404 for representing datatypes in computer systems [23]. The standard specifies the terminology and the semantics for a collection of data types commonly occurring in programming languages and software interfaces. The datatypes defined in the standard are general in nature and serve a wide variety of information processing applications. The standard specifies both primitive datatypes, being defined without a reference to other datatypes, and non-primitive datatypes, which are completely or partially defined in terms of other datatypes.

The ISO/IEC 11404 standard includes a list of 62 definitions of datatype related terms. It also specifies the conditions that have to be fulfilled by an information processing entity in order to conform to the standard directly or indirectly. The standard describes fundamental notions such as a definition of a datatype, a value space, datatype properties, a datatype generator, characterizing operations, etc. We extracted the key terms from the standard, organized these terms into a logically consistent IS-A hierarchy of ontological classes, defined their properties and relations to other entities, re-used suitable textual definitions from the standard, where possible, and added new ontological definitions, where necessary.

3. Related work

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 (RDA) [50], 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 registy (DTR) working group [11] with the goal to: compile a set of use cases for datatype use and management, formulate a data model and expression for datatypes (prototype registry available at [10]), design a functional specification for type registries, and propose a federation strategy among multiple type registries.

Meek [37] discussed a proposal for a taxonomy of datatypes using as a base the first version of the ISO 11404 standard [22]. The taxonomy starts with a number of primitive datatypes that are then used to construct others. The proposed taxonomy is given only in the form of an overview and a discussion, without any formal representation.

The W3C XML Schema Definition Language (XSD) [67] is widely used for the recording of data on the semantic web, and it is also based on the ISO 11404 standard [23]. XSD supports simple, complex, and custom-defined datatypes. It is a simple and flexible language, but it is not based on a formal model and consequently many aspects are left to interpretation. XSD terms are not formally defined. For example, a definition of the term *attribute* ('Defines an attribute') is circular and does not explain how an attribute is different from e.g. *an element* [68].

XSD is flexible and does not strictly regulate the custom-defined datatypes; it also does not enforce the separation of data and its semantic meaning. A side effect of those features is an unnecessary proliferation of custom-defined datatypes. The issue is that different users may create different data models for the same data, and it may be hard to reconcile those models. For example, users can define datatypes such as *start of the project, beginning of the project, start date*. All these datatypes are of the same type *date datatype* and the data encoded with these custom datatypes have the same semantic meaning. A formal ontology can serve as a reference model and resolve such an issue.

The RDF data cube vocabulary is focusing on the publication of multidimensional data on the web [51]. It enables the exchange and sharing of statistical data encoded in a tabular form. The adopted cube model has a set of properties for the description of statistical datasets composed of observations. These include *dimensions* (e.g. time, age, sex), *attributes* (e.g. unit measure), and *measures* (values of observations). A dataset can have *reference metadata* (e.g. a SPARQL endpoint where it can be accessed, its publisher). This purpose-specific vocabulary is well defined and sufficient for recoding of statistical information. However, its strict statistic-oriented model prevents the extension of this vocabulary for other non-tabular datatypes.

An attractive feature of this vocabulary is a distinction between the semantic meaning of observations, the measurement units used, and the data structure specification. The *dimension* property links observations to other resources, i.e. Simple Knowledge

Please cite this article as: P. Panov et al., Generic ontology of datatypes, Information Sciences (2015), http://dx.doi.org/10.1016/j.ins.2015.08.006

Download English Version:

https://daneshyari.com/en/article/6857565

Download Persian Version:

https://daneshyari.com/article/6857565

Daneshyari.com