An Application Ontology to Help Users of a Geo-decision Software Understanding Their Data

Perrine Pittet(✉) and Jérôme Barthélémy

Articque Software, 149 Avenue Général de Gaulle, 37230 Fondettes, France
ppittet@articque.com

Abstract. This paper intends to describe the application ontology of the SaaS version of the decision statistical mapping and geomarketing software Cartes & Données (C & D): CD7Online. Specified in OWL DL, the CD7 ontology was conceived for automation of semantic annotation of CD7Online user data to help users better understand their data and make better selection and representation choices when building maps.

Keywords: OWL DL · Description logics · Application ontology · Ontology development · Semantic annotation · Cartes & données · Geo business software

1 Introduction

Ontologies have been introduced in the Semantic Web research field in the early 2000's to exploit textual documents available on the Web in formalized information [1]. As such, they are sometimes presented as tools for knowledge representation adapted to the Web environment, automatically transforming data into information and information into knowledge [2]. In this paper, we describe an ontology, which was developed to foster users' understanding regarding their data, within a geo business decision SaaS application called CD7Online[1]. This ontology, specified in OWL DL[2], supports the automatized semantic annotation process of user data. In our case the annotation process generates a graph of RDF[3] annotations for each user data workspace, which is stored as a namedgraph in a triplestore. Each namedgraph is automatically queried by an interactive visualization tool, on which users navigate to discover the knowledge behind their data. The rest of the paper is articulated in 4 sections. Section 2 presents the CD7Online project background to expose our motivations for developing a formal application ontology and how this ontology can help users better understand their data. Section 3 describes the CD7 ontology main concepts and justifies their use regarding the task of semantic annotation of user data. Section 4 presents some applications supported by the ontology. Section 5 concludes on feedbacks and future works.

J. Barthélémy—Deceased.

[1] CD7Online: https://cdonline.articque.com/.

[2] OWL reference: http://www.w3.org/TR/owl-ref/.

[3] RDF concepts and abstract syntax: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

V. Tamma et al. (Eds.): OWLED 2015, LNCS 9557, pp. 166–173, 2016.
DOI: 10.1007/978-3-319-33245-1_17

2 Project Background

CD7Online is the SaaS application of the 7th version of Cartes & Données[4] (C & D), which is a French commercial decision statistical mapping and geomarketing software, published by Articque[5]. C & D allows users to obtain effective and interoperable maps built on statistical data, without being mapping specialists. As a business decision tool, it is a data analysis and visualization oriented application, which aims at helping people to take decisions via the maps they build upon geo-visualization. C & D has been designed since the very beginning with the aim of being self-explanatory, simple, and highly intuitive for users - ease-of-use being a major requirement. Nevertheless, C & D still relies on the users good knowledge of their data and their ability to choose the relevant analysis and representation tools to build meaningful maps. Also most of C & D users have a punctual use of the software and often do not have enough available time to study and fully exploit the potential of their data. For solving these issues in CD7Online, we decided to provide users the knowledge they require to quickly understand their data and their potential applications. We chose to use an automated semantic annotation process on these data in order to extract and represent this knowledge. Automatized semantic annotation of data is the process of automatically associating relevant metadata to data, so that each data is described by a set of semantic annotations. The main objective is to exploit these annotations to allow users visualizing, via an interactive graph visualization tool, the concepts related to their data and the semantic relations they share. This tool allows them to intuitively navigate in the annotations, compare and select relevant data to build relevant maps (cf. Fig. 1). As CD7Online user data consist in statistical and geographical data tables specified in xml-based files, we adapted a methodology suited to semantic annotation of tables of data proposed in [3]. In [3], an ontology of the food microbiology domain is adapted to support a semantic annotation process. The concepts of this ontology cover the definitions of microbiological symbolic and numerical types, units, value intervals, relations shared by types and the corresponding lexical data, which are used to name them. We similarly developed an ontology describing the knowledge underlying the geographical and statistical data used by CD7Online. Also, because this knowledge strongly depends on the CD7Online application specific uses and processes, this ontology is not a domain ontology as in [3]'s methodology but an application ontology [4]. This however does not alter the efficiency of the semantic annotation process. In fact the methodology has been designed to accept any ontology, in which semantic relations with lexical data can be added, in order to make possible lexical similarity measures. For the development of the ontology, we have followed a simple methodology proposed in [5]. The ontology development and evaluation experience were presented in [6].

The following section focuses on the description of the main concepts of the ontology.

[4] C & D website: http://www.articque.com/solutions/cartes-et-donnees/.

[5] Articque website: http://www.articque.com/.

3 CD7 Application Ontology Description

For the purpose of this article, we rely on the ontology definition of [7]. Therefore, we define the CD7 application ontology as a formal explicit description of concepts of the CD7Online data domain, properties of each concept describing various features and attributes of the concepts, and restrictions on properties. The ontology together with the set of individual instances constitute the knowledge base designed for the automatized semantic annotation process. A concept can have subconcepts representing concepts that are more specific than this concept. Properties describe properties of concepts and instances. As we needed to keep the maximum expressiveness while retaining computational completeness and decidability for potential inference purposes, we chose to specify the ontology in the OWL DL language. Note that the CD7 ontology terms are originally written in French. Somehow, to facilitate the reading of its description, we translated terms in English and use description logics [8] in the following part. The CD7 ontology[6] defines two main concepts: *DataComponent* and *CDComponent*. *DataComponent* describes all components related to user data, such as metadata, user data. *CDComponent* describes all components related to the CD7Online specific application processes applicable to user data. All the other concepts fall under these two concepts. Due to lack of space we will focus on the main *DataComponent* underlying concepts, which are used in semantic annotation. Three main concepts are considered: *UserData*, *Metadata and LexicalData*.

UserData is a *DataComponent* enclosing the three types of data files a CD7Online user can have in a group of his workspace and use within CD7Online, such as statistical data files, basemaps and maps. *UserData* is defined in $\mathcal{SHOIN}(D)$ DL axioms as follows:

$\mathit{UserData} \sqsubseteq \mathit{DataComponent} \sqcap \mathit{1\ hasFilename.FileName} \sqcap \geq 1\ \mathit{ownedBy.User} \sqcap \mathit{1\ hasGroup.Group}$

$\mathit{UserData} \equiv \mathit{StatisticalDataFile} \sqcup \mathit{Basemap} \sqcup \mathit{Map}$

with *StatisticalDataFile* designating the statistical data files, which contain at least one data table defined by:

$\mathit{StatisticalDataFile} \sqsubseteq \mathit{UserData} \sqcap \geq 1\ \mathit{hasDataTable.DataTable}$

with Basemap designating basemap files used in maps, containing geographical data of a certain geographical space at a certain geographical level, defined by:

$\mathit{Basemap} \sqsubseteq \mathit{UserData} \sqcap \mathit{1\ hasGeographicalSpace.GeographicalSpace} \sqcap \mathit{1\ hasGeographicalLevel.GeographicalLevel}$

with Map designating the map project files created by users within CD7Online, which can import basemaps and statistical data columns, defined by:

$\mathit{Map} \sqsubseteq \mathit{UserData} \sqcap \geq 0\ \mathit{hasBasemap.Basemap} \sqcap \geq 0\ \mathit{hasDataColumn.DataColumn}$

[6] CD7 Ontology url: http://support-articque.com/ressources/CD7Ontology.owl.

Metadata is a *DataComponent* designating all the metadata concepts that can be used to describe the underlying knowledge of components of user data or user data themselves in semantic annotations. *Metadata* is defined by:

$Metadata \equiv DataType \sqcup GeographicalLevel \sqcup GeographicalSpace \sqcup DataIndicator \sqcup Theme \sqcup Date \sqcup Unit \sqcup WeightedTerm \sqcup WeightedWord$

with *DataType* covering three types of data types that can qualify a data column in a data table of a statistical data file: quantitative data, qualitative data and discrete data.

$DataType \equiv QuantitativeData \sqcup QualitativeData \sqcup DiscreteData \sqcup IdData \sqcup UnknownDataType$

$DataType \sqsubseteq Metadata \sqcap \geq 0\ hasDataType^{-}.DataColumn$

with *GeographicalLevel* defining all the geographical division levels that can be considered in a statistical data file or a basemap (ex: regional, national level, etc.).

$GeographicalLevel \sqsubseteq Metadata \sqcap \geq 0\ hasGeographicalLevel^{-}.DataTable$

with *DataIndicator* describing all the statistical indicators a data column can be related to (ex: GDP, mortality rate, etc.). Statistical indicators are categorized by themes. Each statistical indicator is associated with at least one weighted term representing the potential composition of weighted lemmas generally used to designate this indicator.

$DataIndicator \sqsubseteq Metadata \sqcap \geq 0\ hasDataIndicator\text{-}.DataColumn \sqcap \geq 1\ hasTheme.Theme \sqcap \geq 1\ hasWeightedTerm.WeightedTerm$

Another sort of *DataComponent* is used to support the lexical similarity measures used to determine the statistical indicator related to a data column: *LexicalData*. *LexicalData* designates two lexicons instantiated from two concepts: WeightedTerm and WeightedWord.

$WeightedTerm \sqsubseteq DataComponent \sqcap \geq 1\ hasWeightedWord.WeightedWord$

with *WeightedWord* defining a lexicon of all the instances of weighted words that can compose weighted terms. A weighted word is described by a text and a weight, which are respectively typed with string and float values.

$WeightedWord \sqsubseteq DataComponent \sqcap \geq 0\ hasWeightedWord\text{-}.WeightedTerm \sqcap 1\ text.xsd{:}String \sqcap 1\ weight.xsd{:}float$

Additionally, a set of properties, representing the relations between user data components and metadata (as illustrated above) has been defined. Their domains, ranges and facets have also been formalized (c.f. [6]). Finally, in order to set up the knowledge base for the semantic annotation task, a set of individuals was instantiated from the concepts *WeightedWord*, *WeightedTerm*, *DataIndicator*, *Datatype*, *Theme*, *Unit*, *GeographicalLevel*, *GeographicalSpace*. These individuals are required for the automatized semantic annotation process. For example, to identify and annotate a data column with a statistical indicator, the process evaluates the lexical similarity of data cells content with instances of WeightedTerm, which are composed of instances of WeightedWord and associated to instances of DataIndicator. Below is illustrated an example of such instantiation:

WeightedWord(w_mortality0.2)
WeightedWord(w_rate1.0)
weight(w_mortality0.2, 0.2)
weight(w_rate1.0, 1.0)
text(w_mortality0.2, « mortality »)
text(w_rate1.0, « rate »)
WeightedTerm(t_mortality_rate)
hasWeightedWord(t_mortality_rate, w_mortality0.2)
hasWeightedWord(t_mortality_rate, w_rate1.0)
DataIndicator(rate)
associatedWeightedTerm(rate, t_mortality_rate)

4 Applications

One of the main features supported by the CD7Online ontology is the semantic annotation of statistical data tables. It implies the identification of the DataIndicator for each data table column.

4.1 Column DataIndicator Identification

The identification of the data indicator related to a data column involves two steps: first the column data type identification, second the data indicators lexical similarity scores according to the column title content.

The identification of a column data type consists in determining whether its content is quantitative, discrete or qualitative. A set of regular expressions is used to help distinguishing between qualitative numeric values (mostly territorial codes), discrete and true quantitative values. In the ontology, each *DataIndicator* instance is associated to one *Datatype*. Therefore once the data type of a column is identified, the corresponding data indicators lexical similarity scores can be assessed. We adapt here the lexical similarity score definition of [3] (cf. Definition 1). The data indicator, which score is the highest, is then associated to the data column.

Definition 1: Lexical similarity score between column title lemma and weighted terms.

- Let $W = \{w_1 : pw_1; \ldots; w_n : pw_n\}$ and O = $\{o_1 : po_1, \ldots, o_k : po_k\}$, be sets of lemma, with W a set of DataColumn title lemmas w_i and weights pw_i, and O a WeightedTerm instance, with o_i and po_i its respective WeightedWord instances text and weight values.
- Let C be the set of indices pairs (i,j) such as $w_i = o_j$.
- Then the degree of similarity between W and O is:

$$sim_{lex}(w,o) = \frac{\sum_{(i,j)\in C}(p_{w_i} + p_{o_j})}{\sum_{m=1}^{n} p_{w_m} + \sum_{m=1}^{k} p_{o_m}}$$

As recommended in [3], as we do not know which lemma of a column title content W is semantically more important than the others, we automatically associate a default weight of 1.0 to each lemma of the column title considered.

Example: Lexical similarity score of data indicator *gdp* for a column title content *"gdp_per_capita"* composed of two lemma "gdp" and "capita".

DataIndicator(gdp)
hasWeightedTerm(gdp, t_gdp)
WeightedTerm(t_gdp)
hasWeightedWord(t_gdp, w_gdp1.0)
text(w_gdp1.0, "GDP")
weight(w_gdp1.0, 1.0)

If W = ("gdp":1.0; "capita":1.0) and O = ("GDP":1.0) then simlex(W,O) = (1.0 + 1.0)/(1.0 + 1.0 + 1.0) = 2/3.

4.2 Example of Semantic Annotations

A map called "CatchmentArea" created with CD7 is partially described below by a subset of its semantic annotations generated within our system.

```
rdf:type(ns: map298, ns:Map)
ns:nomFichier(ns:map298, "/0/CatchmentArea.cdx")
ns:column(ns:map298, ns:excel55sheet6stores_data_column0)
ns:column(ns:map298, ns:excel55sheet6stores_data_column2)
ns:column(ns:map298, ns:excel55sheet6stores_data_column5)
ns:basemap(ns:map298, ns:basemap15)
```

It uses data columns (c.f. ns:column etc.) from a data table of an excel sheet and a basemap (c.f. ns:basemap). The data column ns:excel55sheet6stores_data_column0 is partially described below.

```
  rdf:type(ns:excel55sheet6stores_data, ns:DataTable)
  ns:column(ns:excel55sheet6stores_data,ns:excel55sheet6stores_data_colu
mn0)
  rdf:type(ns:excel55sheet6stores_data_column0, ns:DataColumn)
  ns:dataType(ns:excel55sheet6stores_data_column0,ns:QualitativeData)
  ns:indicatorType(ns:excel55sheet6stores_data_column0, ns:store_code)
  ns:theme(ns:excel55sheet6stores_data_column0, ns:th_trade)
  ns:columnTitle(ns:excel55sheet6stores_data_column0, "store code)
  ns:geoLevel(ns:excel55sheet6stores_data, ns:level_town)
```

This data column has been annotated with different metadata: title, the data type of its content (c.f.: ns:dataType), the indicator type (c.f.: ns:indicatorType) and the theme (c.f.: ns:theme) it is related to, the excel sheet (c.f.: ns:column) in which it is contained and the geographical level (c.f.: ns:geolevel) on which its content has been processed.

4.3 Visualization

Semantic annotations can be queried with SPARQL to visualize maps related to a specific theme or indicator, to select basemaps or statistical data files sharing the same geographical levels or time in order to select compatible ones, etc. Figure 1 shows a an example of radial visualization built on dynamic SPARQL querying.

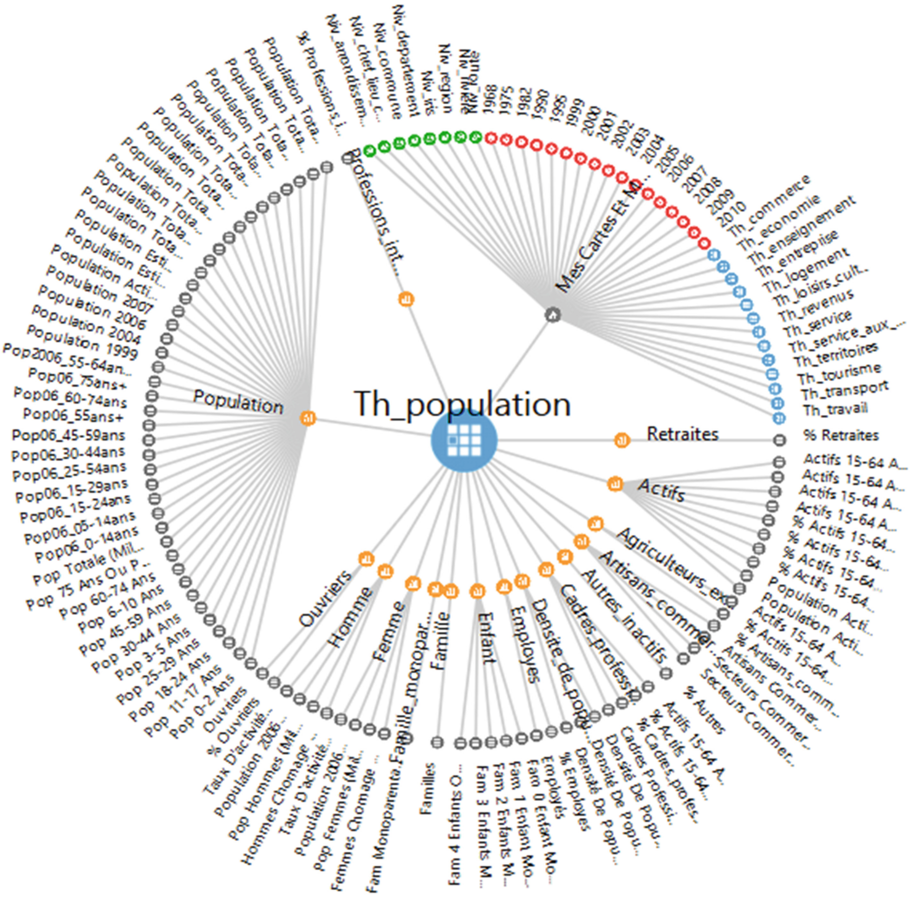

Fig. 1. Visualization of user data annotations within the CD7 graph navigation tool.

5 Conclusions and Future Works

The CD7 application ontology is part of the CD7Online project semantic layer development and supports an automated semantic annotation tool. This tool produces annotations of user data browsable through a graph navigation tool that users can use to better understand their data and build better maps. CD7Online being a commercial software, the development and integration of this layer follows its successive updates. Until now it involved the integration of many semantic tools and technologies. In an industrial project, where deadlines strongly matter, this was a challenge. Hopefully using W3C standards such as OWL clearly helped to reduce development time as many compatible tools for edition, deployment, querying, management and evaluation exist: Protégé, Pellet, Apache Jena-Fuseki, SPARQL, etc. Today, we are working on adding RIF/SPIN rules to provide CD7Online users suggestions of statistical and geographical analysis processes and map representations within a recommender system.

Acknowledgements. This paper is dedicated to the memory of Jérôme Barthélémy, who directed the CD7 project and played a major role in the research work presented here.

References

1. Berners-Lee, T.: Semantic Web Stack (2000)
2. Kaladzavi, G., Diallo, P.F., Lo, M.: OntoSOC: Sociocultural Knowledge Ontology. *arXiv preprint* arXiv:1505.04107 (2015)
3. Hignette, G.: Annotation sémantique floue de tableaux guidée par une ontologie (Doctoral dissertation, AgroParisTech) (2007)
4. Malone, J., Parkinson, H.: Reference and application ontologies. Ontogenesis (2010)
5. Noy, N.F., McGuinness, D.L.: Ontology development 101: a guide to creating your first ontology (2001)
6. Pittet, P., Barthélémy, J.: Experience of formal application ontology development to enhance user understanding in a geo business intelligence saas platform. In: Cuel, R., Young, R. (eds.) FOMI 2015. LNBIP, vol. 225, pp. 51–62. Springer, Heidelberg (2015)
7. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**(2), 199–220 (1993)
8. Baader, F., Nutt, W.: Basic description logics. In: Baader, F., Nutt, W. (eds.) Description Logic Handbook, pp. 43–95. Cambridge University Press, Cambridge (2003)