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

Perrine Pittet, Jérôme Barthélémy

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

Perrine PITTET, Jérôme Barthélémy

Articque Software, 149 avenue Général de Gaulle, 37230 Fondettes
{ppittet, jbarthelemy}@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): CD7 Online. Specified in OWL DL, the CD7 ontology was conceived for automation of semantic annotation of CD7 Online 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, to automatically transform 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 CD7 Online[1]. This ontology, specified in OWL DL, supports the automatized semantic annotation process of user data. In our case the annotation process generates a graph of RDF 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 3 sections. The second section presents the CD7 Online project background to expose our motivations for developing a formal application ontology and how this ontology can help users better understand their data. The third section describes the CD7 ontology main concepts and justifies their use regarding the task of semantic annotation of user data, which is supported by the ontology. The third section shows the results of the integration of the ontology within the CD7 Online software.

[1] CD7 Online: https://cdonline.articque.com/

2 Project Background

CD7 Online is the SaaS application of the 7th version of Cartes & Données[2] (C & D), which is a French commercial decision statistical mapping and geomarketing software, published by Articque[3]. 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. But 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 CD7 Online, 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 navigation 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 CD7 Online user data consist in statistical and geographical data tables, 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 CD7 Online. Also, because this knowledge strongly depends on the CD7 Online 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. The methodology was actually 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 follow a simple methodology proposed in [5]. The ontology development and evaluation experience was presented in [6]. The following section focuses on the description of the main concepts of the ontology.

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

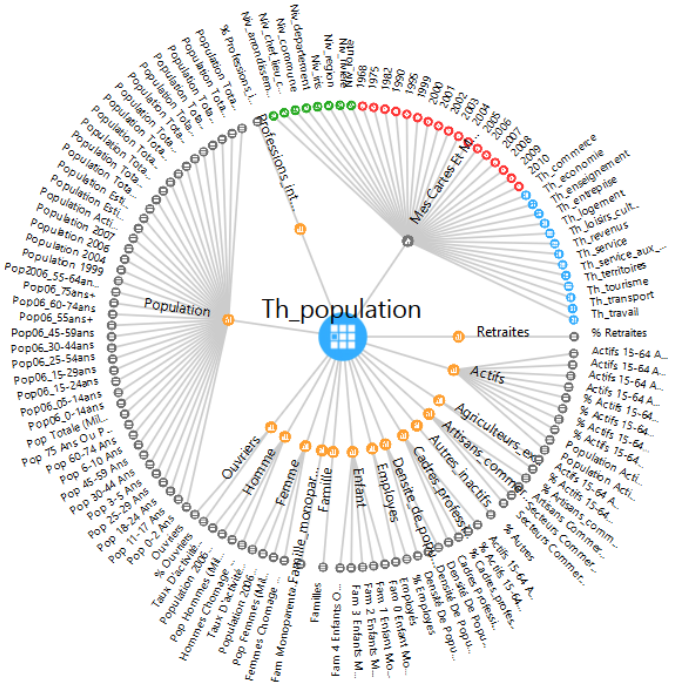

Figure 1: Visualization of user data annotations within the CD7 graph navigation tool.

3 CD7 Application Ontology description

For the purpose of this article, we rely on the ontology definition of [7]. Therefore, we define the application ontology as a formal explicit description of concepts of the CD7 Online application 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 constitutes 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, decidability for 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 logic [8] in the following part.

The CD7 ontology[4] 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 CD7 Online 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

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

DataComponent underlying concepts, which are dedicated to semantic annotation. Three main concepts are considered here: *UserData*, *Metadata and LexicalData*.

UserData is a *DataComponent* englobing the three types of data files a CD7 Online user can have in a group of his workspace and use within CD7 Online, such as statistical data files, basemaps and maps. *UserData* is defined in SHOIN(D) DL axioms as follows:

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

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

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

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

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

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

- *with Map* designating the map project files created by users within CD7 Online, which can import basemaps and statistical data columns, defined by:

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

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:

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

- with *DataType* covering the 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.

$\texttt{DataType} \equiv \texttt{QuantitativeData} \sqcup \texttt{QualitativeData} \sqcup \texttt{DiscreteData} \sqcup \texttt{IdData} \sqcup \texttt{UnknownDataType}$
$\texttt{DataType} \sqsubseteq \texttt{Metadata} \sqcap \geq 0\ \texttt{hasDataType}^{-}\texttt{.DataColumn}$

- 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.).

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

- 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 ⊑ Metadata ⊓ ≥0 hasDataIndicator⁻.DataColumn ⊓ ≥1
hasTheme.Theme ⊓ ≥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.

- with *WeightedTerm* defining a lexicon of all the instances of weighted terms related to statistical indicators that can be compared with data column title lemmas through lexical similarity measures in order to determine the statistical indicator related to the data column. A weighted term is composed of weighted words.

```
WeightedTerm ⊑ DataComponent ⊓ ≥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 ⊑ DataComponent ⊓ ≥0 hasWeightedWord⁻.WeightedTerm
⊓ 1 text.xsd:String ⊓ 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 were 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 WeightedWord, which are composed of instances of WeightedTerm 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 Conclusions and future works

The CD7 application ontology is part of the CD7 Online 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. CD7 Online 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. Adding such amount of new technologies in an industrial project where deadlines strongly matter 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 the CD7 ontology and semantic layer still evolves to stick to the software new features. We are working on adding SWRL rules to provide CD7 Online users suggestions of statistical and geographical analysis processes and map representations within a recommender system.

References

1. Berners-Lee, T. (2000). Semantic Web Stack.
2. Kaladzavi, G., Diallo, P. F., & Lo, M. (2015). OntoSOC: Sociocultural Knowledge Ontology. *arXiv preprint arXiv:1505.04107*.
3. Hignette, G. (2007). *Annotation sémantique floue de tableaux guidée par une ontologie* (Doctoral dissertation, AgroParisTech).
4. Malone, J., & Parkinson, H. (2010). Reference and application ontologies. Ontogenesis.
5. Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.
6. Pittet, P., & Barthélémy, J. (2015). Experience of Formal Application Ontology Development to Enhance User Understanding in a Geo Business Intelligence SaaS Platform. *Formal Ontologies Meet Industry: 7th International Workshop, FOMI 2015, Berlin, Germany, August 5, 2015, Proceedings* (Vol. 225). Springer, 51-63.
7. Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.
8. Baader, F., & Nutt, W. (2003, January). Basic description logics. In Description logic handbook (pp. 43-95).