On the Road to Bring Government Legacy Systems Data Schemas to Public Access

Raquel Lima Façanha[1,2] and Maria Cláudia Cavalcanti[1]

[1] Department of Computer Engineering, Instituto Militar de Engenharia, Brazil,
yoko@ime.eb.br

[2] Justiça Federal do Rio de Janeiro (JFRJ), Brazil,
rlfacanha@jfrj.jus.br

Abstract. Government organizations produce and disseminate a large quantity of information every day. Open government data movement have made these data available for reuse and accessibility. However, to merely publish data retrieved from legacy systems is not enough for reuse and integration. Despite the approaches proposed to publish government data, using technologies like XML, RDF e OWL, these are not suitable for representing real intended meaning of database conceptual schemas belonging to legacy systems. Some approaches propose using top-level ontologies. This paper aims to raise the difficulties to establish the ontological commitment of legacy schemas towards top-level ontologies. It illustrates the difficulties by describing a small case study on the Legal domain.

Keywords: Open government data, top-level ontologies, legacy systems.

1 Introduction

E-Government pursues public services modernization by reducing bureaucracy and by increasing citizen and civil society engagement. The term e-Government emerged only at the end of the 1990s and the first Brazilian government's initiative[3] took place in 2000. Nowadays, public organizations produce, keep and disseminate a large quantity of information every day in many different ways and formats. There are innumerous challenges in providing an efficient and effective access to such information. More recently, in 2011, the Brazilian Open Data Group[4], associated with Brazilian W3C, was formed with the aim to provide directions and guidelines for using and publishing government data. The idea is that data produced by government organizations are made available to people, not only for reading and monitoring, but also for reuse in new projects, sites and applications; for crossing with data from different sources; and for offering interesting and clarifying visualizations.

The Brazilian Open Data Group recommends publishing annotated datasets, i.e., to structure data into knowledge through the use of a domain specific vocabulary and/or metadata. It also recommends following the open data principles[5] and using Semantic Web technologies, such as RDF and OWL. The LEXML project[6] is an example of the use of Semantic Web technologies within Brazilian government. The solution proposed by LEXML to deal with problems related to interoperability, data publishing and accessibility reinforces (and recommends) the following good practices: to structure full

[3] Available in: http://www.governoeletronico.gov.br/o-gov.br.

[4] Available in: http://www.w3c.br/GT/GrupoDadosAbertos.

[5] Available in: http://opengovdata.org/.

[6] Available in: http://projeto.lexml.gov.br.

documents content in XML format; to use controlled terminology or vocabulary; and to use a conceptual reference model named $FRBR_{OO}$. This reference model is a formal ontology intended to capture and represent semantics of bibliography information and to facilitate integration and interchange bibliographic and museum information.

Despite the importance and merits of Brazilian initiatives, government organizations produce and retrieve a large quantity of information from legacy systems, often based on poorly designed database conceptual schemas. It is not enough to publish data and schemas retrieved from these systems. Usually, these schemas provide only the implicit meaning of data and sometimes even miss their meaning, which complicates data reuse and integration. If the real intended meaning of the elements in legacy schemas is not explicit, technologies like XML, RDF and OWL are not sufficient to provide reuse and integration.

Interoperability is essential to materialize the main practical benefits of open government data. A major cause of interoperability problems is the False Agreement Problem. According to [1], although systems can adopt the same vocabulary for data description, there is no guarantee that they can agree on a certain information unless they commit to the same conceptualization, which leads to the False Agreement Problem. From this perspective, data published without their real intended meaning will not effectively reach interoperability since there is no formalization of the ontological commitment [2]. Consequently, in order to publish real meaningful data for reuse and integration, it is necessary to make explicit the ontological commitment of data conceptual schemas from legacy systems. However, since these systems provide little or no description of their data, this is a hard task.

This work discusses the existing approaches to establish the ontological commitment of a database conceptual schema applied to legacy systems and raises some of the difficulties while performing this task using a small case study on the Legal domain.

2 Existing Initiatives and Approaches

Recent works [4] [3] [5] propose the use of top-level ontologies to support conceptual modelling and ontology reengineering. Top-level or foundational ontologies describe very general concepts like space, time, matter, object, event, action, etc., which are independent of a particular problem or domain communities of users [2]. Top-level ontologies are based in philosophical formalisms, which means a philosophically well-founded domain-independent system of formal categories that can be used to articulate domain-specific models of reality [3].

Guizzardi [4] proposes the OntoUML, an extension of UML that incorporates ontological distinctions and axioms. A foundational (top-level) ontology named UFO (Unified Foundational Ontology), supports it and enables the modelling of ontological well-founded schemas. In Guizzardi's work [4], the UFO concepts are described in terms of the metaproperties proposed by [6] and their constraints. This approach provides a guidance on the analysis of each domain category, according to those metaproperties, as way to map it into OntoUML [7]. However, it is oriented to the analysis of each single domain category, while the conceptual model is under development. It does not address the situation where there is an existing conceptual schema from a legacy system, and it is necessary to map all of its concepts to each of the UFO concepts in order to fundament them. These existing schemas are not ready for that, as they were not originally conceived with UFO concepts in mind.

In [8], Guizzardi and Wagner state that the intended meaning embedded in the entities of either a conceptual schema or an ontology representation should be made explicit through the association to a system of meta-level categories, or a top-level ontology.

Some other works [3] [5] describe an ontology reengineering using UFO as the foundational support. Both papers align concepts and relations from a software engineering domain ontology to concepts and relations from UFO-B and UFO-C fragments through an analytical point of view. These works emphasize the importance of foundational ontologies in the development of domain ontologies. However, neither of them raise the difficulties faced throughout the analytical process, nor propose a systematic way to analyze the existing domain ontology.

As stated before, in the context of legacy systems, the schema documentation, if available, is often not complete, may embed misconceived concepts, or simply is not sufficient to provide the real intended meaning of its elements. Nevertheless, even if the system provides a conceptual schema, according to [9], conceptual schemas and ontologies belong to different epistemic levels, have different objects and are created with different objectives, and thus cannot be taken as equivalents. Hence, a gap needs to be filled in. Some works [10] [11] suggest the discovery of knowledge and mapping of business processes for ultimately create a well-founded ontology. Differently, our goal is not to create a domain ontology, but to start with an existing conceptual schema, which is not well-founded. The idea is to identify some of the main difficulties on the review of a legacy schema in the light of a top-level ontology as a way to make explicit its ontological commitment. In the next sections, we describe the SIAPRO's database schema and illustrate some of these difficulties.

3 SIAPRO

SJRJ (Seção Judiciária do Rio de Janeiro) is a trial court for cases involving a federal organ, a governmental agency or a public corporation as interested parties. SJRJ currently uses a database named SIAPRO to store information about those legal cases. In the beginning, SIAPRO's database schema was designed to reflect the requirements of a MUMPS legacy system conceived to register, distribute, classify and record all procedures related to lawsuits at SJRJ. Nowadays, SIAPRO's database schema evolved and not only court servants and judges use SIAPRO; its data is also available on the Web for citizens, legal professionals and other governmental agencies. There is no official documentation about SIAPRO, except for the DBMS metadata, which was automatically extracted for our study.

SIAPRO's physical data schema has about 600 tables. Provided domain's complexity, the schema could not be analyzed as a whole, therefore which criteria should we use to cut-off the database schema? Tables and attributes were selected based on the following criteria: relevance to the domain in focus, data exchange with other systems (interoperability), system specialists and domain experts' knowledge. Figure 1 shows a very small fragment of SIAPRO's database logical schema.

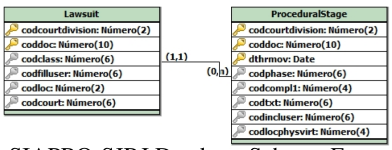

Fig. 1. SIAPRO-SJRJ Database Schema Fragment

The scope of the Legal domain incorporates the legal knowledge obtained by a set of practices used by judges and legal professionals and the interactions of the citizens and federal judicial organs, besides the legal knowledge codified in legal norms. As a consequence, some considerations about Procedural law are required. The key concepts of Procedural law are: lawsuit (*processo*), action (*ação*) and jurisdiction (*jurisdição*).

According to Liebman [12], action is the subjective right to judicial assistance. The existence of an action depends on some essential requirements termed action conditions. Jurisdiction is the right and the duty of State to resolve actual conflicts. Lawsuit is the instrument through which jurisdictional organs act to enforce the law in practical litigious cases. A lawsuit is formed by procedural stages. A procedural stage is a set of procedural acts that occurs during the lawsuit. A procedural act is the smallest unit of a lawsuit and itself is a manifestation of intention from the subjects (parties) in the lawsuit. A procedural act is every human action that produces juridical effects related to the lawsuit [14].

Here is a brief narrative about how the SIAPRO's system works. When a plaintiff files a lawsuit, the system registers a new instance of lawsuit. After that, the court appreciates the complaint. Then, the court formally notifies the defendant of the lawsuit, and orders the defendant to answer the complaint or to make a motion within a specific time interval. Until judgment, parties make allegations and denials, and court rules and make decisions. Each act performed by the judge or the parties or the clerks of justice is a procedural act. Nevertheless, note that the SIAPRO system does not register each procedural act, grouping some of them as procedural stages.

4 Metacategorization of SIAPRO schema using UFO-B/C

Assuming the top-level ontology UFO, which fragment(s) of it would better represent the Legal domain? An initial analysis of the SIAPRO's schema suggests concepts like events, temporal relations and actions, which led us to focus on UFO-B and UFO-C fragments. Still, which concepts from these fragments should we begin with? It seemed to be a good choice to start with the UFO-B main category, which is *Event*, also known as *Perdurant* or *Ocurrent*. Events are individuals composed of temporal parts, such as, a conversation or a business process. Events can be broken down into other events, allowing them to be Atomic or Complex Events. Events are ontologically dependent entities, meaning that they depend on their participants in order to exist. Events happen in time, thus they are framed by a time interval. Events change reality by changing the state of affairs from a previous (pre-state) to a posterior situation (post-state). Finally, events can cause events to happen.

UFO-C is built on the top of fragments UFO-A and UFO-B and its main distinction is between *Agents* and *Objects*. Agents are substantials that can bear intentional moments. *Intentional Moments* are a special kind of moments that have a type: Belief, Desire or Intention; and a propositional content. Intentional Moments can be social or mental moments. *Intentions* are mental moments that represent an internal commitment of an agent and cause the agent to perform actions.

Now, which of the concepts of the SIAPRO schema should we begin with? We chose to start with the *ProceduralStage* entity. Every procedural stage has a one to many relationship with entity *Lawsuit*. The attribute *dthrmov* establishes beginning dates for procedural stages and allow system users to control terms, expressing the idea of temporal extension. As stated before, a procedural stage involves a set of procedural acts. The occurrence of a procedural act leads to a change in reality and can cause other procedural acts to happen. Procedural acts also depend on their participants to exist. For instance, after a service of process occurs, the defendant, whom initially is not aware of the lawsuit, is notified. After notification, the defendant has a certain time to answer the complaint.

Back to the UFO-B fragment, which metacategory represents the *ProceduralStage* entity? Which one best provides real-world semantics to the domain concept? What is the concept's real nature? Based on their definitions, the *ProceduralStage* entity was metacategorized as an *Event*. However, is *Event* the best metacategory for *ProceduralStage*? Is there some specific metacategory that would suit better? As stated in Section 3, a procedural act is every human action that produces juridical effects related to the lawsuit [14]. Thus, it depends on an agent, which is the only one capable to perform actions and the only one capable of bearing intentional moments. There should be an intention in order to a procedural act occur. If we extend the analysis to a model level, the *ProceduralStage* entity could be seen as an intentional event, that is, an *Action*.

In summary, to step towards making explicit ontological commitment of database conceptual schemas from legacy systems, some questions arise. Which element of the domain schema concept should be tackled first? Which is the most promising top-ontology fragment? Which is the most suitable (meta) category to each of the domain schema concepts? Is it enough to (meta) categorize studying only the database conceptual schema? Which characteristics must be present in both domain and top-ontology concepts to find a one–to–one correspondence between them? How to identify if a subclass/superclass of a top-ontology concept would better (meta) categorize the domain concept in hands? It is not in the scope of this work to provide an answer to all these questions. The small case study presented here serves as a scenario to reveal some of the difficulties on this task.

5 Conclusion

This work presented a real, but small case study on the Legal domain, and identifies some of the main difficulties on the task of expliciting the ontological commitment of the SIAPRO legacy schema. A previous work [15] establishes guidelines to obtain a well-founded conceptual representation from a database conceptual schema with little or none documentation, using UFO-A fragment. New efforts should be done on facilitating the use of UFO-B and UFO-C. To make explicit the ontological commitment of a database schema, besides the need to involve domain experts and system specialists, it is necessary the use of some guide or systematic that could help the modelers on the many decision points throughout this task. In addition, ontological distinctions require that modelers have a deep knowledge on the theoretical fundaments of the top-ontology. An ongoing work aims to provide a systematic approach to facilitate the process of obtaining ontologically well-founded schemas from legacy data.

References

1. Guarino, N.: Formal Ontology in Information Systems. In: Proc. of FOIS 98, pp. 3–15. IOS Press, Itália (1998)
2. Guarino, N., Carrara, M., Giaretta P.: Formalizing Ontological Commitment. In: Proc of the 12th Nat. Conf. of Artificial Intelligence, pp. 560–567. AAAI, California (1994).
3. Guizzardi, G., Guizzardi, R.S.S., Falbo, R.A.: Grounding Software Domain Ontologies in the Unified Foundational Ontology (UFO): The case of the ODE Software Process Ontology. In: CIbSE, pp. 127–140. Recife (2008)
4. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. CTIT PhD, Univ. of Twente, Enschede (2005)
5. Bringuente, A.C.O., Falbo, R.A., Guizzardi, G.: Using a Foundational Ontology for Reengineering a Software Process Ontology. J. of Inf. and Data Management, pp. 511–526 (2011)
6. Guarino, N., Welty, C.: A formal ontology of properties. In: Proc of 12th Int. Conf. On Knowledge Engineering and Knowledge Management, pp.97–112. Springer, Heidelberg (2000)
7. Guizzardi, G., das Graças, A.P., Guizzardi, R.S.S.: Design Patterns and Inductive Modeling Rules to Support the Construction of Ontologically Well-Founded Conceptual Models in OntoUML. Advanced Information Systems Engineering Workshops, pp. 402–413. Springer, Heidelberg (2011)
8. Guizzardi, G., Wagner, G.: Towards Ontological Foundations for Agent Modeling Concepts using UFO. In: Proc of Agent-Oriented Inf. Systems, pp.110–124. Springer, Heidelberg (2005)
9. Fonseca, F., Martin, J.: Learning the Differences Between Ontologies and Conceptual Schemas Through Ontology-Driven Information Systems. In: Proc of Journal of the Association for Information Systems(JAIS),129–142 (2007)
10. Bouamrane, M. M., Rector, A. , Hurrell, M.: Using OWL tologies for adaptive patient information modelling and preoperative clinical decision support, Knowledge and Information Systems, pp. 405–418. Springer, Heidelberg (2011)
11. Dang, J., Hedayati, A., Toklu, C.: An ontological knowledge framework for adaptive medical workflow, Journal of Biomedical Informatics, pp. 829–836. (2008)
12. Liebman, E. T.: Manual de Direito Processual Civil. Malheiros, Rio de Janeiro (2005)
13. Código de Processo Civil. (1973)
14. Pacheco, J.S.: Direito Processual Civil. Saraiva, São Paulo (1976)
15. Silva, A.M.F.R.: Diretrizes para o Resgate do Esquema Conceitual e Seu Compromisso Ontológico A partir de um Banco de Dados. M.Sc. Dissertation, Instituto Militar de Engenharia, Rio de Janeiro (2012)