Ontology Based Data Management: A Study in a Brazilian Federal Agency

Márcia Myuki Takenaka Fujimoto, Edna Dias Canedo

Ontology Based Data Management

A Study in a Brazilian Federal Agency

Márcia Myuki Takenaka Fujimoto [1][0000-1111-2222-3333] and Edna Dias Canedo[2][1111-2222-3333-4444]

[1] Ministry of Transparency and Comptroller General, Brasília, Brazil
[2] Computer Science Department - Professional Masters in Applied Computing - University of Brasília - Unb, Brasília, Brazil
ednacanedo@unb.br

Abstract. The Ministry of Transparency and Comptroller General is the agency of the Federal Government in charge of assisting the President regarding the treasury and public assets and the government's transparency policies. The Agency takes care of active transparency mechanisms on federal public resources, that is, improves actions related to information the State must disclose, without being demanded for. It establishes ways of assuring information will be appropriated and effectively used by society, including through web applications for citizens. Thus, the Agency is part of a governmental environment in which the complexity of data management involves improvement of information quality and interoperability between systems, as a consequence the need of a capable model to manage all of its data assets rises up. It's necessary to organize and implement a data management capability that allows understanding (semantic), finding and sharing data. This article describes a research study directed to data management under ontological approach, which proposes an Enterprise Information Architecture Model in the form of a conceptual layer-based data prototype, taking into account both academic and industry-driven studies.

Keywords: Information and Communications Technology (ICT), Data Management, Data Integration, Ontology, Public Company.

1 Introduction

The Ministry of Transparency and Comptroller General (CGU) is the agency of the Federal Government in charge of assisting the President of the Republic regarding the treasury and public assets and the government's transparency policies. These tasks are out by way of public audits, fraud deterrence procedures, and other sort of internal control, corruption prevention and ombudsman activities. The performance of its various activities is often done through the consolidation of information, cross-referencing data, data mining and open data availability. Such activities require a holistic and integrated data assets view. The existing data management structure is elementary and has not shown to be effective in meeting CGU requirements, resulting in several challenges.

The main input of the many tasks of the CGU is the data from three main sources: self-owned systems to support their final activities, databases of structuring systems of the Federal Government and external data provided by systems of institutions connected to the Government. The data unchecked proliferation in the Agency leads to problems in data governance and management. Some of that is due to the system's architecture being organized in "silos", meaning it's made up of a myriad of data sources, independent and distributed, each serving a specific application [1]. Thus, even though there is a significant amount and variety of data sources, there are basic questions which are hard to answer, such as:

1. Q1. How many and which are the data sources that contain information regarding the organizational structure of government, regardless of rank or power?
2. Q2. What are the sources from the CGU systems that contain citizens' data and can be integrated?
3. Q3. What are the data sources to be searched in order to obtain information about all CGU control actions in a particular Agency ?

There are challenges in finding, integrating and improving organizational data quality. As a consequence, interoperability among systems is jeopardized. Any solutions to get over such difficulties need management capabilities focused on semantic integration, both intra- and inter-organizationally. Such solutions demand data sources labeling with business and technical metadata, logically ordered for these. Aware of this situation, CGU launched in 2014 a working group to study the Cobit 4.1 PO2 Process - Define the Information Architecture [10] for a coming implementation. Considering that the implementation has not occurred so far, there is an opportunity to a new one PO2 process implementation proposal. Cobit 4.1 presents two control objectives directly related to the data understanding and sharing: PO2.1 - Enterprise Information Architecture Model; PO2.2 - Enterprise Data Dictionary and Syntax Data Rules.

The data integration and sharing issue is core to the PO2 process. Because of this, this study considered holistic approaches related to ontology based data management - OBDM to ensure the true data integration. This study intends to, more specifically, present an information architecture proposal, in the form of a conceptual ontology based data model, supported by a data dictionary (DD) solution architecture coherent with the recommended approach. The information architecture will generate a prototype for evaluation purposes.

The remainder of the paper is organized as follows. Section 2 presents the background in which the theoretical bases for conducting this research is raised. Section 3 summarizes the research and architecture development methodologies. Section 4 presents the Case Study and the preliminary results of this work. We conclude the paper presenting some final remarks in Section 5.

2 Background

The systematic review carried out, described below, resulted in the choice of the Ontology based data management - OBDM proposal [2] and the holistic multi-domain

architectural structure by [3]. They have a great similarity of purpose with this work with regard to an overview of the enterprise as a whole and to the data integration and sharing. Other sources for the study include some Federal Government initiatives related to the data management field.

2.1 Ontology Based Data Management

The OBDM approach can be seen as a way of integrating information in which the global schema of data is substituted by a conceptual model from the domain of interest to a given organization formally specified in an ontology. The architecture to which the main idea of OBDM relates is divided in three levels: ontology, data sources and mapping between these two. More specifically for OBDM, ontology is the formal description of a domain of interest to a given organization, expressed through its relevant concepts, such as concepts' attributes and the logical affirmations that characterize knowledge on the domain. The data sources, in turn, are repositories accessible by the organization where the data domain is stored. Frequently, these repositories are numerous and heterogeneous, with each one being managed and kept independently from the others. The level of mapping is a precise specification of the correspondence between data kept in the sources and elements of the ontology [2].

This division in layers has three main advantages. Firstly, the ontological layer which, by making the representation of the domain explicit, allows the re-usability of the acquired knowledge and a unified description of underlying data sources. Secondly, the mapping layer explicitly specifies the relationships between the concepts of domain (ontology) and the data sources. The ontology and mapping corresponding to the data sources provide a common element for the documentation of all the data in the organization, with obvious advantages for the governance and the management of the information systems. The third advantage relates to the extensibility of the system, which doesn't require to fully integrate the data sources at once. Instead, after building a skeleton of the domain model, new sources or elements therein can be incrementally added [2].

The study presented by [13] regarding semantic databases details how the mapping between sources and ontology occurs. The main idea is to map tables and attributes of a Database (DB) for a determined ontology. This ontology must be formal in regards to the implementation of a transitive hierarchy "is-a", which connects all concepts.

2.2 Multi-Domain Reference Architecture

Fitzpatrick [3] explains that the ontology must distinguish between domain knowledge, that may be extra organizational, versus localized application level knowledge. Besides, he explains the idea of the criterion of orthogonality that is applied to the structure proposed. The criterion of orthogonality is defined as the requirement of basing a newly created ontology on one or more existing ontologies. This practice, if generalized, would help reduce the "silo" effect in the development of ontologies [19].

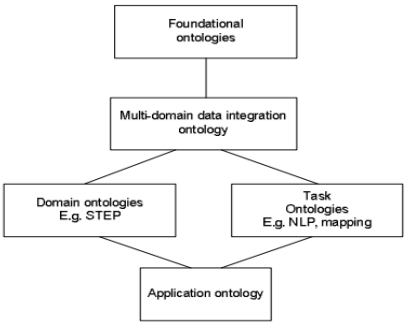

Fig. 1. Reference Architecture Ontology Structure

Fitzpatrick [3] proposed reference architecture ontology structure is composed of the top-to-bottom ontologies (see Figure 1). In light of the criterion of orthogonality, the proposed multi-domain data integration ontology subsumes in respect with the foundational ontologies. Domain specific ontologies are subsumed to the multi-domain ontology proposed. Then, the ontology structure comprises generic task ontologies and the structure is completed with application ontology to support domain specific tasks.

The structure proposed by [3] fits into the ontology layer of the OBDM architecture and allows the implementation of master data management (MDM). The multi-domain data integration layer comprises master data ontologies, that is basically fundamental data for all business transactions, essential cross-enterprise assets that contribute to many management paradigms [3]. This reference architecture by allowing MDM implementation ensures a collaborative environment between the many business paradigms or processes.

2.3 Federal Public Administration Initiatives

The Framework of Enterprise Architecture for Interoperability in Governance Support (FACIN) seeks to support the Brazilian Digital Governance Strategy. Through the establishment of the Enterprise Architecture and of interoperability standards, the FACIN will act as a reference for the many agencies of the Federal Public Administration FPA[1] [6].

The FACIN is made up of nine visions (domains), in which are the Data vision and the Applications vision. Structurally, it's composed of four main parts, among them:

- Content Model (CM) - Describes the structure of related elements that describe generic models for representation of Federal Public Administration organizations. CM is specified through a conceptual model presented by a diagram that joins metadata according to two perspectives [4]: Data Description and Data Sharing. The FACIN conceptual Data Description Model is supported by the Entity Rela-

[1] FPA corresponds to the set of agencies of the direct, autarchic and foundational administration.

tionship (ER) modelling and it encompasses semantic, syntactic and technical metadata.

- Reference Model (RM) - Describes standards, guides and best practices for the development of the FACIN from the strategic to the operational levels, focusing in the integration and construction of the Government vision as a whole [6]. RM points to the observation and care with the de definition of master data as one of the critical factors of success in implementation of the data vision. Besides that, it stipulates the utilization of the Global Data Model (GDM) and the Controlled Vocabulary of Electronic Government (CVEG) as standards.

Global Data Model - GDM is a model created with the idea of enabling and ensuring the integrability of information generated for the decision process. Essential data and process modeling is used to create GDM model. It is called essential because it considers only the relevant information to the understanding of the business domain, discarding operational or technological details. Essential modeling acts similarly to a reverse engineering process of description of the database of the legacy systems [18]. The GDM aims to schematically represent the data treated by the information systems, seeking to create an integrated metadata platform. Therefore, the GDM produces some artifacts related to the data semantic and process understanding, among them, a Complete Integration Diagram. It's a vision that puts all ER entities together, a global schema of data that is a big model in ER style and highlights the common entities between information systems.

Vocabularies and Ontologies of the Electronic Government (e-Vog) intend to be a set of standards, tools and methodologies to allow: information exchange under semantic agreement, so as to favor data matching from various sources; elicitation of the tacit knowledge from government business areas by using ontology as a tool for conceptual modeling [17]. The e-Vog is a initiative under construction and has not yet made its products available, except the Repository of Vocabularies and Ontologies of Electronic Government, a site to access all ontology references of the Electronic Federal Government [17]. As stated before, FACIN indicates GDM, an ER model, to be adopted in the implementation of the data vision in the federal agencies, seeking to make feasible an integrated metadata platform. Besides that, FACIN points to take care of the de definition of master data. On the other hand, the e-Vog initiative indicates the use of ontologies for conceptual data modeling as being in this aspect more aligned to the proposal of this work.

These FPA initiatives are still evolving, but currently they do not provide ready and appropriate models for CGU needs. Either way, they do offer some directions that can be observed for the development of CGU information architecture and DD solution.

3 Methodology

One of the first activities of this study was a systematic literature review about studies that dealt with ontology for data management and, after that, the analysis of some that were selected. The systematic review process was based on the Kitchenham e Char-

ters' guide [12]. The main results of the systematic literature review allow us to conclude that:

- the OBDM has received increasing attention since 2006, with numerous ontological data model proposals with an adjacent technical architecture applied to real problems;
- the reporting of implementation experiences corroborates the proposed models applicability;
- the majority of the studies with a model proposal deal with a specific domain;
- most of the authors indicate the need for expansion of use and evolutions in the presented models.

Considering that the ICT field is somewhat lacking in fundamental theory, a mixed research approach is recommended, with a theoretical base and practical verifications, especially in regards to data integration [7]. Knowledge can be developed (drawn) from academic research and (also) practice [7]. Based on that, other sources for the study include guides for best Information and Communications Technology (ICT) governance practices as well as some related Federal Government initiatives. The ICT guides considered are: Control Objectives for Information and Related Technologies - COBIT 4 [10] and the Togaf 9 [20] [9].

As regards the development of the information architecture, the ontology based conceptual data model, it followed the informal phase of Enterprise Approach of ontology construction [21]. For this, it was decided to use the questions and problems mentioned in the introduction 1 of this work as competency questions to identify the ontology based model scope, a list of potentially relevant concepts [11]. One of the ways to determine the scope of the ontology is to sketch a list of questions that a knowledge base based on the ontology should be able to answer, competency questions [8].

As regards the development of the DD solution architecture, it followed customized steps of Togaf Architecture Development Method - ADM - phase "A. Architecture Vision" [9]. The TOGAF ADM is a generic method that describes a method for developing an enterprise architecture. ADM is phases cycle composed, ranging from the phase "A", of the initial architecture view, to the phase "H" which identifies further necessary changes. The phase "A" objective is to create a vision of the architecture proposal (version 0.1). It intends to help the approval decision and, also, to help the understanding of its impacts.

4 Case Study and Preliminary Results

The systems and activities that assist CGU functions deal with varied and complex data. The proper use of CGU data sources requires effective management by a common element for the documentation of all the data in the organization: the ontology-based data model.

4.1 Ontology-based Data Model

The current stage of this research did not allow the development of a proper CGU ontology, in which the concepts of its elements are shared by the users of the proposed system. Therefore, we have currently opted for the proposal of an ontology based conceptual data model, which can evolve to become an ontology on a future stage. The development of the conceptual data model and its prototype for this research was done through a few options: i. the adoption of OBDM approach of integrating information; ii. the customization of the multi-domain model [3], carefully observing the FACIN recommendation of taking care of the sovereignty of information, master data [5]; iii. the use of CGU domain based concepts, CVEG based concepts, in compliance with the MR-FACIN recommendation [5], and other available ontologies; iv. the use of the Protégé [16] tool, web version, to draft the ontology based data model prototype.

The basic structure of CGU proposed model comprises a multi-domain data integration layer - MD, a business domain layer - BD, a generic task layer - TA and an application layer - AP. In the proposed model, business data domains are subsumed in respect with MD layer data domains. The application data domains subsume BD layer data domains and TA layer data domains. The MD layer data domains were selected from some enterprise ontologies used as references for this work, even they not belong to governmental area. This is possible due to the multi-domain concepts generality.

Figures 2 to 4 were taken from the prototype structure of the ontology-based conceptual model built on the WebProtégé tool. Figure 2 identifies data domains that will compose the MD layer. They address the fundamental concepts, some already known in data modeling, that allow the systems interoperability. It means that this layer can help answer questions about master data like Q1 and Q2 in the Section 1.

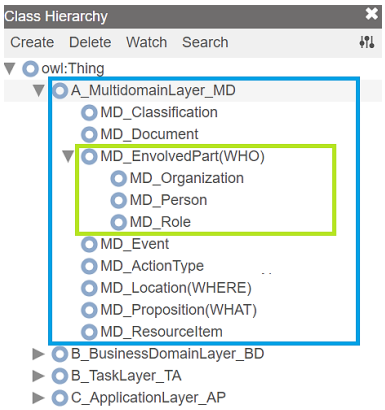

Fig. 2. Data Domains for the Multi-Domain Data Integration Layer

Figure 3 left side identifies some concepts that make up BD layer, they represent functions performed by the agency and its specializations, which have been derived

from the information in [14]. This layer is consistent with the GDM view as it addresses only the scope of the business domain. Figure 3 right side shows how the data domains are related to describe the CGU functions in a general way: Subject (involved parties) + action (type of action from the agency) + object (resource item or other) and complement (locale and others). All the elements which are necessary to the description in this general way must be found in the MD layer or in the in the BD Layer. This layer can help answer questions like Q3 in the Section 1.

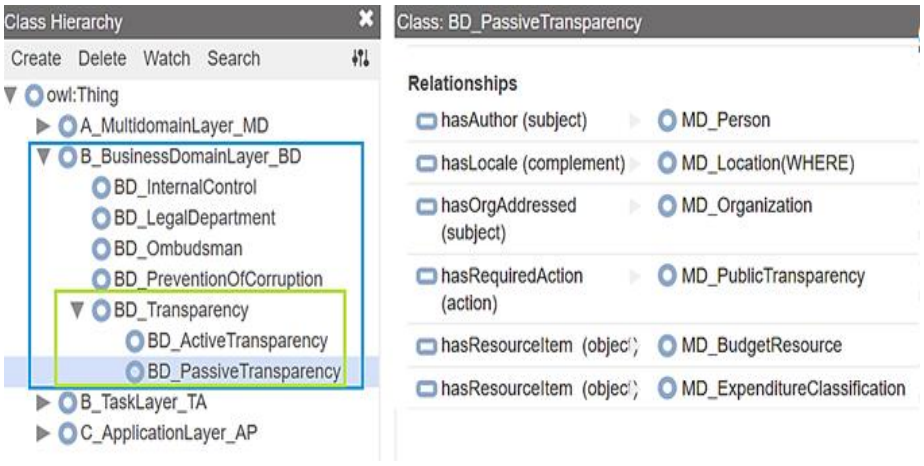

Fig. 3. Business Domain Layer and Concepts Relationships

Figure 4 left side identifies some elements that make up the TA layer. The layer is organized by subjects related to tasks that are independent from the business domain, that could become reusable corporate services. The listed subjects are simply illustrative, able to evolve as applications are cataloged. The TA data domains may subsume from MD layer data domains, for example "Document", or may be a proper subject like "File". This layer can support the reuse of general services and data sharing, for instance, user's access authorizations task in "Security" subject.

Finally, the AP layer will have two divisions: BusinessApplication and SupportServiceApplication. The CGU applications that most meet the business functions will be under the first division and the ones destined to offer support services will be under the last one. Figure 4 right side presents the two main divisions of the AP layer and some examples of existing applications in the Agency. The data domains of each application are subsumed in respect with BD domains or TA domains.

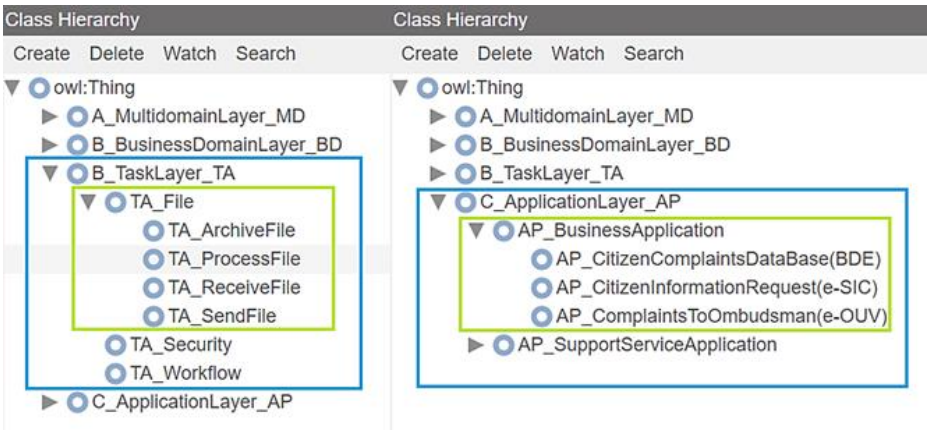

Fig. 4. Task Layer and Application Layer

The proposed conceptual model, top-down, is made of layers. Its is different from GDM type models that are created by ER models composition of applications that interact between themselves, with a bottom-up approach. Nevertheless, each integration point identified in the GDM type models must find a corresponding concept or relationship in the proposed model MD layer. Besides that, each data source of the CGU applications, when submitted to reverse engineering, should find a corresponding concept or relationship in any layer of the proposed model, as a way to verify its completeness.

Over the rest of the research project, the prototype will be evolved and data sources from two CGU applications will be mapped to the proposed model concepts. The directions presented in [13] regarding semantic database will be used to map the applications database tables to the proposed ontology based model. Next, the prototype will be explored by CGU experts (like database administrators, application developers, business intelligence analysts and business specialists) to answer a qualitative survey in order to validate the proposed model.

4.2 Architecture for Data Dictionary Solution

The architecture considered suitable for CGU needs is presented in the Figure 5 and its main elements are described in the table 1. The proposal is oriented by OBDM's base idea [2] of separation between ontology and data source, with their respective mappings. This version 0.1 architecture vision, that covers business, application, data and technology domains, was draft in an Archimate tool. Archimate is an open enterprise architecture modeling language to support the description, analysis and visualization of architecture within and across business domains [15]. In future experiment stages in this project, there might be a simplification in some components, such as the removal of the extraction and automatic load of WebProtégé [22] ontology elements to the ontologies repository.

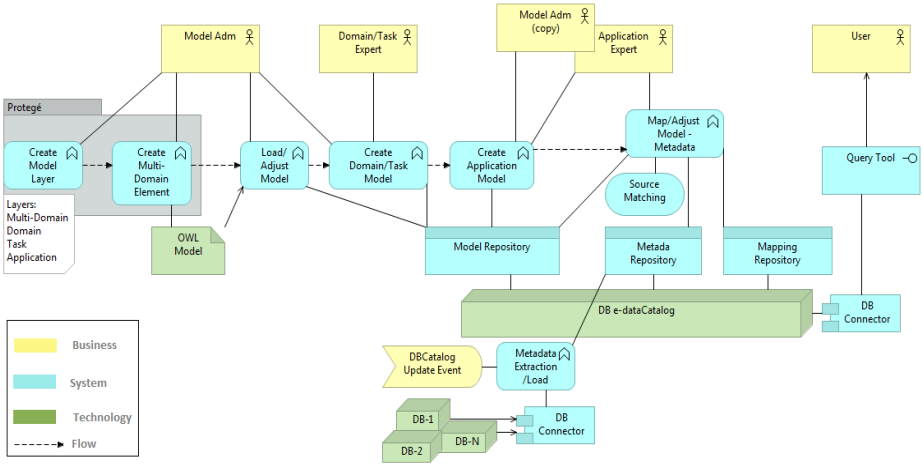

Fig. 5. Proposed Architecture for DD Solution

Table 1. Architecture Components

Element	Description
Model Adm	Actor responsible for maintaining the model basic structure and assists other managers so as to keep it coherent.
Domain/Task Expert	Specialist actors that describes a domain or task/service.
User	Actor who uses the DD information.
Create Model Layer	Model structure creation function.
Create MD Element	Master data domains creation function.
Load/Adjust Model	Model data extraction and relational repository loading functions. Adjustments are allowed to the Model's Administrator.
Create Domain/Task Model	BD and TA data domains creation function. It allows concepts and relationships creation.
Create Application Model	AP data domains creation function. It allows concepts and relationships creation.
Metadata Extraction /load	Function that automatically extracts technical metadata from structured data sources to load on the metadata repository.
Map/Adjust Model - Metadata	Function that allows the mapping between structured data sources (tables) and the model concepts. It can uses table matching solutions, through structure or content analysis.

5 Conclusion

The proliferation, sometimes uncontrolled, of electronic data in the government area, as well as in other areas of society, is a fact. As control mechanisms and the transparency of government actions evolve, becoming more complex, so does the need of a robust management capacity of this fundamental asset. The recent tendency of data management with ontology support provides a promising path to solve problems in data understanding, localization, reuse and integration, improving the interoperability between systems and the quality of the information created.

This study considers the academic advances and industry initiatives in order to elaborate a proposal for the implementation of the management of multi-domain data with an ontological approach for the CGU, using quantitative and qualitative research. We hope that the proposal can help the CGU PO2 process – Define the Information Architecture - effective implementation, as well as aligning the CGU initiative to the ones existing in the FPA.

The prototype of the conceptual model and the proposed architecture should be a little more detailed and validated with its application in a reduced scope of CGU applications, to be defined at a later date. Besides, there is the concern of doing a more minimalistic detailing with the maintenance of a central structure that can be adapted to other FPA agencies.

References

1. Daraio, C., Lenzerini, M., Leporelli, C., Moed, H. F., Naggar, P., Bonaccorsi, A., & Bartolucci, A.: Sapientia: the ontology of multi-dimensional research assessment. In Issi (2015).
2. Daraio, C., Lenzerini, M., Leporelli, C., Moed, H. F., Naggar, P., Bonaccorsi, A., & Bartolucci, A.: Data integration for research and innovation policy: an ontology-based data management approach. Scientometrics, 106 (2), pp. 857-871 (2016).
3. Fitzpatrick, D., Coallier, F., & Ratté, S.: A holistic approach for the architecture and design of an ontology-based data integration capability in product master data management. In: *IFIP International Conference on Product Lifecycle Management*, pp. 559-568. Springer, Berlin, Heidelberg (2012).
4. Governança Corporativa, G. T.: FACIN - Framework de arquitetura corporativa para interoperabilidade no apoio à governança - modelo de conteúdo v1.5, retrieved 2018/06/01, from http://www.participa.br/ articles/public/0056/4482/ModelodeConteudov1.5.pdf (2017).
5. Governança Corporativa, G. T.: FACIN - Framework de arquitetura corporativa para interoperabilidade no apoio à governança - modelo de referência v3.1, retrieved 2018/06/01, from http://www.participa.br/articles/public/0056/4483/Modelo_de_Referencia_v3.1.pdf (2017).
6. Governança Corporativa, G. T.: FACIN - Framework de arquitetura corporativa para interoperabilidade no apoio à governança - visão executiva v2.1, retrieved 2018/06/01, from http://www.participa.br/ articles/public/0056/4481/FACINVisaoExecutivav2.1.pdf (2017).

7. Gregor, S.: Building theory in the sciences of the artificial. In Proceedings of the 4th international conference on design science research in information systems and technology p. 4. ACM (2009).
8. Grüninger, M., & Fox, M. S.: Methodology for the design and evaluation of ontologies. Citeseer (1995).
9. Haren, V.: Togaf version 9.1. Van Haren Publishing (2011).
10. ISACA, C.: Cobit R 4.1 process assessment model. ISACA IL, USA (2011).
11. Jones, D., Bench-Capon, T., & Visser, P.: Methodologies for ontology development (1998).
12. Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., & Linkman, S.: Systematic literature reviews in software engineering - a tertiary study. Information and Software Technology, 52 (8), pp. 792-805 (2010).
13. Köhler, J., Philippi, S., & Lange, M.: Semeda: ontology based semantic integration of biological databases. Bioinformatics, 19 (18), pp. 2420-2427 (2003).
14. Ministério da Transparência e Controladoria-Geral da União – competências page, http://www.cgu.gov.br/sobre/institucional/competencias-e-organograma, last accessed 2018/06/01.
15. Open Group Archimate Overview Homepage, http://www.opengroup.org/subjectareas/enterprise/archimate-overview, last accessed 2018/06/01.
16. Protégé Homepage, https://protege.stanford.edu/, last accessed 2018/05/20.
17. Repositório de Vocabulários e Ontologias de Governo Eletrônico Homepage, http:// vocab.e.gov.br/, last accessed 2018/06/01.
18. SERPRO: Guia metodológico para integração de dados e processos modelo global de dados, retrieved 2018/05/20, from http://modeloglobaldados.serpro.gov.br/modelo-global-de-dados/guia-metodologico (2017).
19. Smith, B.: Ontology (science) (2008).
20. Togaf Version 9.2, available in http://pubs.opengroup.org/architecture/togaf9-doc/arch/index.html, last accessed 2018/06/01.
21. Uschold, M., & King, M.: Towards a methodology for building ontologies (1995).
22. WebProtégé Tool page, https://webprotege.stanford.edu/, last accessed 2018/05/20.