Modelling the Public Sector Information through CIDOC Conceptual Reference Model

Lina Bountouri[1], Christos Papatheodorou[1,2], and Manolis Gergatsoulis[1]

[1] Database & Information Systems Group (DBIS),
Laboratory on Digital Libraries and Electronic Publishing,
Department of Archive and Library Sciences, Ionian University, Corfu, Greece
[2] Digital Curation Unit,
Institute for the Management of Information Systems (IMIS),
Athena R.C., Athens, Greece
{boudouri,papatheodor,manolis}@ionio.gr

Abstract. Nowadays, due to the growing development of eGovernment information systems, there is an increasing need to handle Public Sector Information (PSI) in a homogeneous way. Ontologies are currently a powerful tool to act as semantic reference models for the development of information systems and as semantic mediators for achieving interoperability. In this paper, we analyze the procedures that lead to the PSI's production and management and we present all the concepts and agents that relate to it. Based on this analysis and given that CIDOC CRM ontology is able to define the rich semantics of the historical records' production and management, we propose the CIDOC CRM to represent the public records' conceptualization and to act as a reference model for PSI.

Keywords: Public Sector Information, eGovernment, Ontologies.

1 Introduction

The *Public Sector Information* (*PSI*) or *Government Information* is the information created, collected and freely disseminated in the form of public records by the *Public Administration* (*PA*). Indicative examples of PSI are financial and business information, legal and administrative information, scientific and cultural information etc. The management of PSI deals mainly with the facilitation of the transactions with PA, the access to PSI and its use and reuse so as to act as the basis for the provision of added value services within PA and/or to external users (citizens and business).

Public sector is considered as the largest information provider in almost every country and various PSI systems have been developed aiming to satisfy the augmented needs. An important part of such systems is the adoption of standards for the documentation, organization and dissemination of information, for the administrative terminology as well as standards/protocols for their communication and interoperability.

The public records carry significant information and have their own characteristics. They are documented through standards, usually depending on the

R. Meersman et al. (Eds.): OTM 2010 Workshops, LNCS 6428, pp. 404–413, 2010.

country it produces them, such as eGMS in the UK [13] and AGLS in Australia [12]. Furthermore, they may be penetrated by different documentation logics and records management policies, for example differences have been observed in their philosophy, since some of them are more oriented to the management of PSI, while others are oriented to facilitate the citizens to access PSI.

As a consequence, interoperability issues related to the Government Information are not only technological, but they also cover a wide range of aspects, such as the adopted policies, the lack of agreement on common standards and vocabularies etc [10]. The dimensions of interoperability needed to be addressed between the PSI information systems are the organizational view (i.e. by making the services widely available), the technical view (i.e. by achieving data integration) and the semantic view (i.e. by achieving the PSI's semantic conformance). Besides, even if various conceptual models have been proposed to semantically define the wider PSI domain or parts of it, most of them are oriented: a) to define concepts needed specifically for the provision of eGovernment services [7], or b) to represent the Public Administration views [2], or c) strictly to deal with records management and not related to the basic notions of PSI [8]. It is important to notice that even if these efforts serve their goals in a specific context, they cannot deal in parallel with the important building blocks of the PSI's production and management: a) the events and functions in the citizen's, business' and government's life on which the production of PSI is based, b) the archives and records management policies, and c) the need to provide added value services to the internal and external users.

In our research, we deal with the semantic interoperability issues, which could also promote the PSI's semantic integration. Our purpose is not to represent all the existing concepts related to PSI systems, but to a) bring out the main PSI's semantics and b) propose the use of a tool that can act as the basis for the semantic alignment of the PSI metadata and the provision of interoperable services, always based on the fundamental archives and records management practices and, at the same time, without leaving unexplored the role of the eGovernment services. In order to achieve these goals, we propose the use of the ontology CIDOC Conceptual Reference Model [4] as a reference model of the PSI domain, since *ontologies* provide rich constructs to express the meaning of data, promote reasoning and are widely used as mediators between heterogeneous sources. CIDOC CRM, specifically, is an event-based ontology and this characteristic, as it will be analyzed in the following sections, enables the representation of functions that generate the PSI. Hence, the added value feature of this work is that it associates the public records not only with their producers, but also with the PA's functions which generate, modify and use the records.

Notice that we already have explored the role of the CIDOC CRM for the archival information in [15] and since the *public archives* are *records* of continuing value selected for permanent preservation, while the *public records* refer to the documents that are still in current use, their strong relationship is obvious. Given that both the public archives and records provide evidence for the daily functions of the PA, the objective of this work is to integrate the concepts penetrating the management of the two types of documents and to promote interoperability between information and service providers. Reusing CIDOC CRM for PSI's conceptualization and alignment promotes the PSI's incorporation to the wider archival semantic environment.

In this paper, we firstly derive the domain specific requirements for the PSI, analyze its conceptualization, based on the study of the international standards and practices, and present all the involved agents and ideas related to its management. Secondly, we introduce the CIDOC CRM to represent the semantic structure of the involved knowledge, by demonstrating how its conceptualizations can be used to document PSI, by assigning the main PSI's notions to specific CIDOC CRM paths. Finally, related efforts and conclusion are presented.

2 The Production and Management of the PSI

The delivery of PSI's services typically involves the interaction between the citizens, the business and the PA in a complex scenario, not only in terms of technology, but also of how the relationships and the processes are organized and how the necessary data are structured and handled. In this section we analyze these PSI's concepts and relations. A more extended analysis can be found in [3].

PSI refers to the operation of the *Public Administration*, having the form of the public *records*. In particular, the PSI is produced either during the *PA*'s internal procedures or during its communication to external users, such as *citizens* and *business*. In both cases, specific *functions* are executed and *records* are produced to accomplish the tasks to be completed. The *PA* consists of government's agencies that control and supervise public programmes and have executive, legislative, or judicial authority over other institutions within a specified area. These agencies also set the strategies, recommend the creation of *laws* and generate the *mandates*. The various agencies in the public sector are typically engaged in the organization and the production of *services* [17].

A *function* is any high-level purpose, responsibility or task, assigned to the accountability agenda of a corporate body by *legislation*, *policy* or *mandate* [14]. *Functions* are decomposed into a related set of *activities*, which are the tasks performed by a corporate body to accomplish each of its *functions*. *Activities* encompass *transactions*, which in turn produce *records*. The importance of the notion of *function* in PSI is proved through the archives' and records' management practice, which strongly connects the creation of *records* to *functions*. *Functions* are considered as a more stable point of reference than administrative structures, because administrative structures are often merged or devolved when restructuring takes place [5]. Moreover, they are strongly related with the *citizens* and *business* through specific *business situations* and *life-events*. Due to that fact, records management and business classification schemes document or are based on *functions*, such as eGMS [13], AGIFT [11] and ISDF [14].

A *record* is in line with [9], information in any form or medium, created or received and maintained by an organization or person in the transaction of business or the conduct of affairs. *Records* are essential parts for the accountability of the governments to maintain the democracy and to provide the access to the public. The public *records* provide authentic and reliable evidence of the past and current *functions* and of the *transactions* that result from them. Public *records* have their own characteristics that enable their identification and management. According to [13,11,19], some if these are: (a) the *language* of the *record*, which is of crucial importance, especially in multi-lingual environments (i.e. the E.U.), (b) the unambiguous *identifier* of the *record* within a specific context, (c) the *place* where the *record* can be reached or found, and (d) the *title* given to it.

What is more, *citizens* and *business* are an important part of the PSI's documentation, since they have daily communication with the *PA* and the right to require *services*. The *citizens* participate in the *life-events*, which are everyday life situations in which a citizen uses *PA services* to confront them. Some of the most common *life-events* are: moving home, bereavement etc. *Business* participate in *business situations*, which are situations where they trigger the public services or interactions with the public authorities, such as founding a company, (re-)constructing factory premises etc. *Life-events* and *business situations* are highly important eGovernment concepts participating in the PSI's management and this is proved by the fact that specific ontologies have been created for these concepts [7].

Furthermore, the actions taken by the *PA* are based on *laws* and *mandates* recommended by the *PA*. In the PSI's context *laws* and *mandates* are specific warrants that require the resource to be created or provided [12] and they introduce and clarify the *function/activity* to produce the public *records* [14]. Another important notion for the PSI's creation and management is the notion of *policy*, which is a plan of action adopted by the *PA* aiming to achieve particular targets. In the PSI context, *policy* comprises, among others, activities related to *records*'s management and dissemination, such as accessibility and reproduction.

A concept of crucial importance in the PSI's production and management is the *time*, which is usually formed as a date associated with a specific event in the life cycle of the PSI (i.e. production, copyright, modification) [13]. The *Public Administration* is also related to other concepts, such as *place*, defining for example where the *PA* units are located or where the *records* can be found. Various *names* are used to identify individuals of the *PA*, such as the *businesses*, the *citizens*, the *places* etc.

A PSI reference model, in order to act as the basis for the creation of metadata, to improve the communication between interest parties and to be a mediator for the alignment of PSI sources, must be able to describe all the mentioned conceptualizations, their complexity and interrelations.

3 Using the CIDOC CRM for the PSI's Modelling

3.1 The Followed Methodology

The *CIDOC Conceptual Reference Model* (*CIDOC CRM, ISO 21127:2006*) is a core ontology, which consists of a hierarchy of 86 *entities* (or *classes*) and

137 *properties*. Its main target is to promote a shared understanding by providing a common and extensible semantic framework to which information can be mapped, to facilitate the information integration for cultural heritage information sources and to help the implementers to formulate the requirements for information systems, serving as a guide for conceptual modelling [4]. CIDOC CRM expresses semantics as a sequence of path(s) of the form entity-property-entity. It is an event-based model and its main notions are the temporal entities. As a consequence, the presence of CIDOC CRM entities, such as actors, dates, places and objects, implies their participation to an event or an activity.

The methodology followed was based on the identification of specific CIDOC CRM classes and properties and their between paths to represent the main PSI's concepts and relations presented in Section 2. More tools, such as standards [12] and ontologies [16] can be used to further specialize the CIDOC CRM's semantics for PSI, however, it is out of the scope of this paper to deal with more custom PSI's aspects. Given that CIDOC CRM is a semantically rich model, further classes and properties can be used to represent additional, but of secondary importance, PSI's notions not presented in this paper due to space reasons.

As part of our methodology, we studied in depth the Government Information metadata to explore their "hidden" semantics and their interrelationships. We did not restrict our study to Government Information, but we also studied the records management policies (i.e.[6]) and the archival description standards (i.e. [14,9]). This action was taken given that the PSI is encapsulated in the public records and archives; hence, ensuring the effective documentation and management of the records and archives by applying the relevant standards and policies results over time in the efficient management and exploitation of the PSI. Duranti [18] emphasizes the importance of records for the PSI, stating that they play a crucial role in most human activities and they are essential to all business and social exchanges, being the basis of the legal system.

The PSI model is partially presented in Figure 1 and analyzed in the two following sections. In Figure 1, the entities of CIDOC CRM are presented in the upper part of the circles and in the lower part their corresponding PSI semantics are presented. The relations between the entities are indicated through arrows.

3.2 The PSI Model: The Entities

The CIDOC CRM classes that represent the three main concepts of PSI (*PA*, *function* and *record*) are respectively instances of the classes E40 Legal Body, E7 Activity and E22 Man-Made Object. E40 Legal Body includes instances that represent the institutions or groups of people that have obtained a legal form as a group, can act collectively and can be held collectively responsible for their actions. The instances of E7 Activity are actions intentionally carried out by instances of E39 Actor and its subclasses (such as E40 Legal Body) that result in changes of state in the cultural, social, or physical systems. Hence, its instances can be used to represent the *functions* carried out by the *PA* for the creation and modification of PSI. Notice that the *functions* include *activities* and *activities* include *transactions*, which are also represented as instances of E7 Activity.

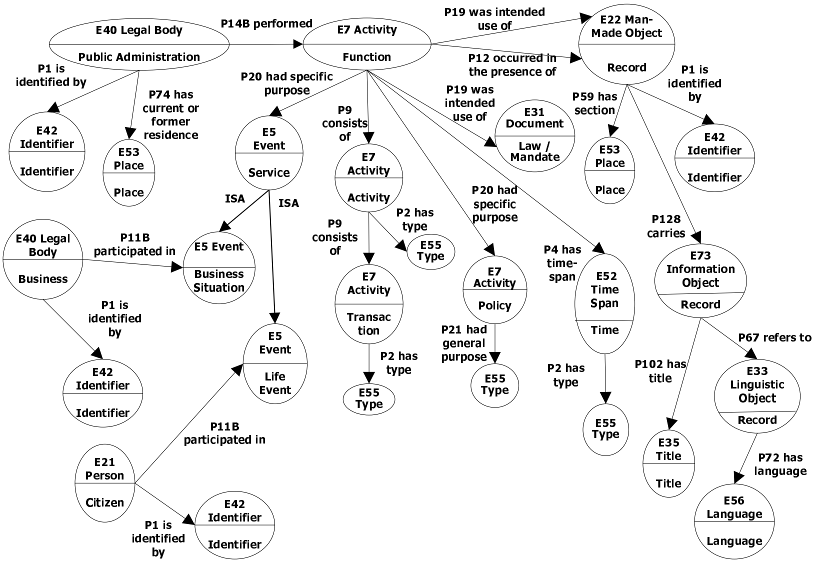

Fig. 1. The PSI model based on CIDOC CRM

The *records* are represented through E22 Man-Made Object that defines the physical objects purposely created by human activity. A *record*, apart from being an object created on purpose, it is also an information carrier. The informational view of the *record* can be represented as an instance of E73 Information Object. In case the *record* includes text, it is also an instance of E33 Linguistic Object.

Citizens and *business* are denoted as instances of E21 Person, which includes the instances of real persons who live or are assumed to have lived, and E40 Legal Body. In addition, the *services* provided by the *PA* to the society are modelled as instances of E5 Event, which comprises the changes of states in cultural, social or physical systems. *Life-events* and *business situations* are depended on the *services*, hence can be also modelled as instances of E5 Event, being in an ISA relationship with the class E5 Event representing the *PA*'s *services.*

Identifiers are represented as instances of E42 Identifier that includes the strings or codes assigned to instances of almost every CIDOC CRM class, in order to identify them uniquely and permanently within the context of one or more organizations. For example, a *citizen* may have a unique identifier as an employee in the public or private sector, as a member of an insurance/health system or as a tax payer. Moreover, the *titles* of the public *records* are instances of E35 Title because this class includes the names assigned to works.

As mentioned in Section 2, *laws* and *mandates* recommend and control the creation of public *records*. To represent this notion, E31 Document is used, since its instances are information objects that comprise the immaterial items making propositions about reality. The *policy* adopted by the *PA* can be modelled as instances of E7 Activity, given that it is a set of actions intentionally carried out by instances of E39 Actor, which result in changes. To further specialize the type of the *policy*, i.e. to define that a *policy* is targeted to the PSI's preservation, instances of E55 Type can be associated to instances of E7 Activity.

E53 Place covers the *places* where *PA*'s bodies are located and where the public *records* can be found. These places could be organized in a taxonomic hierarchy indicating the geographic divisions. Another important PSI's notion is the dates that surround the various activities (such as creation dates, deletion dates etc). *Time* is represented in CIDOC CRM via the E52 Time-Span.

It is worthy of note that the instances of E55 Type can be associated to almost every CIDOC CRM class used in this model, to further specify its meaning. For example, it can be used to specialize the type of the *time*, by associating instances of E52 Time-Span to instances of E55 Type. An additional example is its use to further specify the type of *functions*, *activities* and *transactions* (instances of the class E7 Activity). The same applies to instances of E41 Appellation, which are used in order to provide names to instances of a large number of CIDOC CRM classes, such as E21 Person, E40 Legal Body etc.

3.3 The PSI Model: The Semantic Relations

The classes described in Section 3.2, form a rich semantic network through the use of the CIDOC CRM properties that relate their instances. In order to express the main concepts of the PSI's generation and management (*PA*, *function* and *record*) and their interrelations, instances of E40 Legal Body are related to instances of E7 Activity via the property P14B performed and instances of the E7 Activity are related to instances of E22 Man-Made Object via the property P12 occurred in the presence of, with the purpose of representing that the *PA* (E40 Legal Body) performs *functions* (E7 Activity) in an environment which produces/manages *records* (E22 Man-Made Object). To define that the public *records*, created by the *PA*'s *functions*, are evidence of these *functions*, the property P19 was intended use of is introduced between instances of the class E7 Activity (representing the *functions*) and instances of the class E22 Man-Made Object (representing the *records*) to express that specific objects are created for use in the *functions*. Besides, instances of E22 Man-Made Object are related to instances of E73 Information Object via the property P128 carries to express that the produced *record* carries information. Additionally, instances of E73 Information Object are related via the property P67 refers to with instances of E33 Linguistic Object to denote that sometimes the produced *record* can be expressed in natural language(s), independently of the medium that carries it.

The class E40 Legal Body representing the *PA* is also related to instances of the classes E42 Identifier and E53 Place through the properties P1 is identified by and P74 has current or former residence respectively, to define specific *identifiers* that the *PA* may have, i.e. identifiers for the ministries, their departments etc, and specific *places* where the offices, departments etc of the *PA* are located.

Furthermore, the decomposition of the *functions* to the *activities* and then to the *transactions* is declared through the property P9 consists of that relates instances of E7 Activity, representing *functions*, to instances of E7 Activity, representing the *activities*, to instances of E7 Activity, representing the *transactions*.

At this point, notice that instances of E7 Activity, representing the *functions*, are also related to instances of E5 Event via P20 had specific purpose to define that some of the *functions* are intended to serve *life-events* and *business situations*. E40 Legal Body (representing the *business*) and E21 Person (representing the *citizens*) are both related to instances of E5 Event representing the *life-event* and *business situation* respectively. Both are related to these instances by the use of the relationship P11B participated in. Besides, instances of E40 Legal Body and E21 Person are related to instances of E42 Identifier, through P1 is identified by, to denote the relationship between them and the identifiers assigned to them.

Aiming to express that the *functions* are based on *laws* and *mandates*, the property P19 was intended use of is used between instances of the class E31 Document (representing the *functions*) and instances of the class E7 Activity (representing the *laws* and the *mandates*). E7 Activity (representing the *PA*'s *policy*) is related to E7 Activity (representing the *PA*'s *functions*) through the relationship P20 had specific purpose, stating the relation between the *policies* and the every day operation of the *PA*. In particular, through this CIDOC CRM path it is expressed that some of the *PA*'s *functions* usually affect the generation of *policies*. With the purpose of specifying the type of the *policy* followed (e.g. in the PSI context, preservation, appraisal, disposal, rights, accessibility and reproduction policies could be applied) the instances of the class E7 Activity (representing the *policy*) are related to instances of the class E55 Type through the property P21 had general purpose, since this property involves activities intended as preparation for some type of event.

For the expression of the *time* of the *functions*, the P4 has time-span is used to relate instances of E7 Activity to instances of E52 Time-Span. Notice also that the instances of E41 Appellation are reached through almost every class via P1 is identified by and can express names which refer to and identify specific instances [4], i.e. the names of the *citizens*. E55 Type can also be reached through almost every CIDOC CRM class through P2 has type. In the PSI context, this class can be used to express sophisticated semantic needs.

The CIDOC CRM classes and properties used to model PSI can be further analyzed to a taxonomy of classes and subclasses. For instance, the *PA* can be analyzed to Central and Regional Administration and then the Central Administration to be analyzed to the classes Ministries and Supervised Public Organizations, and the Regional Administration to the hierarchy Regions, Prefectures, Municipalities. Nevertheless, this analysis is not of interest for this paper, which focuses on the representation of the main concepts of the PSI' s production and management processes and not on the provision of analytic conceptualizations, which usually form part of thesauri, vocabularies etc.

4 Related Work and Discussion

Lately, the PA is facing many challenges like improving its services and reducing costs. Due to that fact, many related efforts are currently running to facilitate the PA's tasks in a national and international level.

In [1] a PSI framework for data sharing and reuse to support interoperability is proposed. This framework includes a cross-application reference model, which provides instructions for modelling the processes in the PA's context and can be customized with domain specific metadata and ontologies. In [8] a conceptual metadata schema model for records management is defined, based on the ISO 15489 and ISO 23081 guidelines. This schema's target is to maintain the international compatibility and standard management procedures. It is a record-centered model consisting of three basic elements: "Records", "Business" and "Mandate". The first model proposed is oriented to cover mostly eGovernment needs, without taking into consideration records management policies. The second model is orientated to records management and due to that fact, even if it can deal with parts of PSI, it is not related to the basic notions of eGovernment, such as services offered by the PA (life-events and business situations).

It is important to notice that the existing conceptual models for PSI do not explicitly associate the notion of the public records' production with the PA's functions as well as with the citizens or business operations, but they consider independently either the PA's characteristics and their relation with the business situations and life-events, or the public records with their producers. In [2] a PA ontology is proposed representing some of the PA's views (legal, organizational, business, IT, end-user). Nevertheless, this ontology is adapted as a part of a mechanism inside a life-event portal, it includes very broad concepts and, since it is not based on records and archival policies, it is inadequate for use as an interoperability tool between the PSI's documentation.

In our research, CIDOC CRM does not intend to replace metadata schemas that describe the PSI, but to define a conceptual view of it, which could complement the wider aspect of PSI's management, describing the most important concepts and their relations. Our main target is to bring out the main semantics related to the PSI's production and management.

To conclude, this emergence is promoting the use of CIDOC CRM as a semantic reference point for the PSI metadata in various applications, such as government portals, reasoning systems and integration architectures. In regard to the integration, the CIDOC CRM could be adopted act as the mediator between diverse PSI sources in a metadata interoperability scenario. An indicative example is its use in governmental portals which offer one-stop services to citizens and for this purpose integrate information form different PSI sources (municipalities, ministries, etc.) having possibly different metadata schemas, all of them mapped to the ontology. The satisfaction of a citizen's application might require the integration of information form several distributed sources. In a such case particular queries translated to suitable forms for each source should be promoted by the mediator, using the appropriate mappings, and then the answers should be integrated before returned to the user.

References

1. Baralis, E., Cerquitelli, T., Raffa, S.: A Cross-Application Reference Model to support Interoperability. In: 2nd Eur. Summit on Interoperability in the iGovernment, Rome (2008)

2. Bercic, B., Vintar, M.: Ontologies, Web Services, and Intelligent Agents: Ideas for Further Development of Life-Event Portals. In: Traunmüller, R. (ed.) EGOV 2003. LNCS, vol. 2739, pp. 329–334. Springer, Heidelberg (2003)
3. Bountouri, L., Papatheodorou, C., Soulikias, V., Stratis, M.: Metadata Interoperability in Public Sector Information. Journal of Information Science 35(2), 204–231 (2009)
4. CIDOC CRM SIG. Definition of the CIDOC CRM. Technical report (January 2010)
5. Shepherd, E., Yeo, G.: Managing records: a handbook of principles and practice, ch. 3, p. 74. Facet Publishing (2003)
6. International Organization for Standardization. ISO 15489 1:2001, Information and Documentation: Records Management Part 1: General. ISO, Geneva (2001)
7. Kavadias, G., Tambouris, E.: GovML: A Markup Language for Describing Public Services and Life Events. In: Wimmer, M.A. (ed.) KMGov 2003. LNCS (LNAI), vol. 2645, pp. 106–115. Springer, Heidelberg (2003)
8. Han, S.-K., Lee, H.-S., Jeong, Y.-S.: Conceptual model of metadata schema for records management. In: Proc. of 2nd Int. Symp. on Know. Processing and Service for China, Japan and Korea, Metadata and Ontology, Beijing, pp. 21–31 (2006)
9. International Council on Archives. Committee on Descriptive Standards. ISAD(G): General International Standard Archival Description. ICA, 2nd edn. (2000)
10. Interoperability Solutions for European Public Administrations. European Interoperability Strategy (EIS): Document for public consultation (February 2010), http://ec.europa.eu/idabc/servlets/Doc?id=32595
11. National Archives of Australia. AFIGT (2005), http://www.naa.gov.au/records-management/create-capture-describe/describe/AGIFT/index.aspx
12. National Archives of Australia. AGLS Metadata Standard Reference (2008), http://www.agls.gov.au/documents/aglsterms/
13. Cabinet Office. e-GMS v.3.1 (August 2006), http://www.cabinetoffice.gov.uk/media/273711/egmsv3-1.pdf
14. International Council on Archives. Committee on Best Practices and Standards. ISDF: International Standard for Describing Functions (2007), http://www.ica.org/sites/default/files/ISDF ENG.pdf
15. Stasinopoulou, T., Bountouri, L., Kakali, C., Lourdi, I., Papatheodorou, C., Doerr, M., Gergatsoulis, M.: Ontology-Based Metadata Integration in the Cultural Heritage Domain. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 165–175. Springer, Heidelberg (2007)
16. Stojanovic, L., Kavadias, G.,Apostolou, D., Probst, F., Hinkelmann, K.: Ontology-enabled e-Gov Service Configuration (June 2004), http://ec.europa.eu/information_society/activities/egovernment/docs/pdf/ont ogov.pdf
17. US Census Bureau. North American Industry Classification System (2002), http://www.census.gov/epcd/naics02/
18. Wilson, G.: Keeping the Records Straight: Pr. L. Duranti put 17th-century monks to work for the Pentagon (1998), http://www.publicaffairs.ubc.ca/ubcreports/1998/98mar05/98mar5pro.html
19. Archives New Zealand. NZGLS Metadata Element Set Version 2.1. (2004), http://www.e.govt.nz/standards/nzgls/standard