A Cross-Application Reference Model to support Interoperability

Elena Baralis
Politecnico di Torino
Torino, Italy
elena.baralis@polito.it

Tania Cerquitelli
Politecnico di Torino
Torino, Italy
tania.cerquitelli@polito.it

Silvana Raffa
CSI Piemonte
Torino, Italy
Silvana.Raffa@csi.it

Abstract

Data and service requests are currently generated and managed in a distributed fashion. Furthermore, different actors (e.g., service providers, product sellers, governmental organizations) need to exchange data in a wealth of different formats. To allow an effective information exchange, systems need to support interoperability, thus enabling the sharing of information and knowledge. To this end, a general-purpose exchange data model needs to be defined. While many efforts have been devoted to the design of ad-hoc exchange data models, less attention has been devoted to the design of a general reference model. This reference model should define a set of unifying concepts and relationships, which are independent of the specific application domain.

This paper proposes a cross-application reference model (i.e., a meta model) which provides a high level definition of concepts and relationships, in terms of data and metadata (e.g., data provider description, data beneficiary/user identification, data quality dimensions). This reference model provides a general infrastructure for the definition of customized data sharing services in a given application domain. The proposed cross-application reference model has been exploited to build different data models for data exchange in real systems (e.g., the heterogeneous context of the Piedmont registry office of welfare beneficiaries). The performed validation shows the adaptability and effectiveness of the proposed reference model.

INTRODUCTION

Data and service requests are currently generated and managed in a distributed fashion. Furthermore, different actors (e.g., service providers, product sellers, governmental organizations) need to exchange data in a wealth of different formats. To allow an effective information exchange, systems need to support interoperability, i.e., ability of different systems and organizations to work together (i.e., interoperate) for exchanging data and for sharing information and knowledge [8]. Interoperability can be treated at different levels such as technical, semantic/conceptual, and organizational [4]. Achieving technical interoperability means providing efficient solutions to solve technical issues arising from linking different computer systems and services. However, syntactical and structural heterogeneity issues could cause misunderstanding of data. Semantic interoperability techniques can solve these problems by ensuring that the meaning of exchanged information is understandable by any application. Data enriched with semantics enable meaningful processing. Furthermore, by defining business processes and improving the collaboration of administrations wishing to exchange information, organizational interoperability can be achieved. We focus on semantic/conceptual interoperability. To this end, a framework for data sharing and reuse beyond the community boundaries, across enterprises, and among applications is required. In this context a winning feature is the definition of a general exchange data model to support interoperability that can be rapidly customized with reusable domain specific meta-data and ontologies. Consider, for example, different organizations that need to share data and services in a wealth of different formats. Each data provider and each user application exploits its own local data schema. However, when they need to share information and knowledge, a sharing platform is required. The sharing platform can be an integration bus (i.e., infrastructure) in which data providers render their services and data users access useful information and knowledge. The first step in the infrastructure design is the definition of data exchange models. Application domains may require different data exchange models. However, a general reference model could be rapidly customized with domain specific metadata and ontologies.

On March 1, 1977, the Piedmont regional administration, together with the University and the Politecnico di Torino decided to set up the "Consorzio per il Sistema Informativo del Piemonte" (i.e., CSI Piemonte) – a consortium for the regional information system – with the aim to innovate local government through information technology. Over these thirty years, CSI Piemonte has devoted a significant effort to the development of a regional system offering effective services to citizens and businesses to promote the area's development. In this context, the interoperability of regional public authorities becomes a key issue. CSI Piemonte has developed the Piedmont Regional Information System to allow different public bodies to cooperate. In the last ten years, the need to exchange and share information and services among different Public Authorities has been steadily growing. CSI Piemonte is addressing this issue by means of several different projects to allow technical and semantic interoperability. Interoperability may be achieved at growing levels of complexity: (a) interoperability among public subjects of the Piedmont Public Authority (e.g., Piedmont registry office of welfare beneficiaries project), (b) interoperability among Italian regions (e.g, the

ICAR interoperability project), and (c) interoperability between Piedmont and Italian Public Authority (e.g., the SPCoop project). Since all projects are characterized by a common domain model (i.e., interoperability support for public sector information systems), a reference model may allow sharing relevant semantic concepts in the public sector applications and may effectively support CSI developers.

The challenge of this work is to design a reference model to simplify the modeling task and to allow inter-operability among different public systems and public organizations. Thus, this paper presents a cross-application reference model which provides guide lines for the modeling process and model templates in the public sector context. The proposed reference model, also called canonical information model, describes data from the business point of view and introduces a set of metadata (e.g., data provider description, data beneficiary, user identification, data quality dimensions) to enrich information. It is independent of the logical and physical schema exploited by data providers and data consumers. Furthermore, it is easily customizable to support data sharing services in a given application domain. Thus, the model helps domain experts during the design of customized data sharing services. The proposed cross-application reference model has been exploited to build different data models for actual data exchange in real systems (e.g., the heterogeneous context of the Piedmont registry office of welfare beneficiaries, Piedmont Regional Information System). The performed validation shows the adaptability and effectiveness of the proposed reference model.

The paper is organized as follows. Section 2 describes the main features of the cross-application reference model designed to support interoperability. In Section 3 real information systems to validate the proposed approach are presented. Section 4 describes the INTEROP tool and its main features. Section 5 discusses related work, while Section 6 draws conclusions and presents future developments of the proposed approach.

A CROSS-APPLICATION REFERENCE MODEL

The cross application reference model (i.e., canonical information model) identifies a set of concepts useful to model data, services and metadata (e.g., actors and data quality dimensions) to support interoperability. Each concept consists of micro-concepts including different attributes. Due to lack of space, this papers presents the micro-concepts, whereas a complete description of the attributes devised to model concepts is available in [5]. Four concepts have been introduced and detailed in the following. The complete reference model is shown in Figure 1 and Figure 2.

Person, identified by a unique code, can be a natural person or a legal person. In the reference model, this classification is modeled by means of a total and exclusive generalization (see Figure 1). A legal person is any legal entity duly constituted under applicable law, for profit, privately-owned, or governmentally-owned. It also includes any corporation, trust, partnership, joint venture, or association [3]. Different micro-concepts have been devised to characterize a legal person: Tax registry identification, company name, incorporation information, company size, and cutback information. For each micro-concept a set of attributes have been devised (see [5] for further information). To characterize a natural person, different micro-concepts have been identified, such as registry identification (i.e., name, surname, date of birth, city of birth, sex), tax registry identification, citizenship, civil status, passport number, social security number, and local health unit of enrollment. For each natural person, the proposed approach is able to model any residence changes by means of a ternary relationship among location, natural person, and time entities. The identifier of the time entity is the start date (i.e., date in which the residence in a given location started). The many-to-many relationship is also characterized by an attribute (i.e., the end date) which models the date in which the residence of a given person in a given location ends. Other ternary relationships model any domicile changes of natural persons. Furthermore, a legal person can be classified as a public legal person, which pursues public interests, or a private legal person, whose aims are privately-owned. This classification has been represented by means of a total and exclusive generalization (see Figure 1). For each legal person all organizational units are known. This aspect is modeled by means of a one-to-many relationship defined between the legal person and the organizational unit entities. Since an organizational unit can be a (distinct) legal person, a one-to-one relationship (called *can be*) is also defined between organizational unit and legal person entities. For example, consortia are legal persons whose members (also called organizational units) are other legal persons.

A person can play the role of service provider or service consumer. The proposed approach can model the role of the person for different periods of time (i.e., start/end dates). Identifiers of provider entities and consumer entities are automatically assigned by the system (see Figure 2). Since a person carries out different professions, and some of them require to be written in a given register (e.g., medical register, the Rolls), the proposed reference model is able to represent these aspects. Furthermore, the same person can perform different professions at the same time. Thus, the profession entity is characterized by an external identifier (i.e., subject identifier), starting date and profession identification. Since only a subset of professions requires the register enrollment, the participation to the binary relationship *written in*, defined between profession and register entities, is optional. The complete list of skills required to perform any profession is included in the model definition. Finally, the public legal person managing each register is known (see Figure 1).

Goods can be classified as personal properties and real estates by considering the tax office view. This classification is modeled by means of a total and exclusive generalization. The location is known only for real estates. Since a given real estate has only one location and in each location there is only one real estate, a one-to-one binary relationship has been defined between real estates and location entities. The list of owners and possessors of each good in each period of time (i.e., start/end dates) is also modeled. Different goods classification can be modeled by means of the proposed meta-model discusses in Section Upgrade of the cross-reference model.

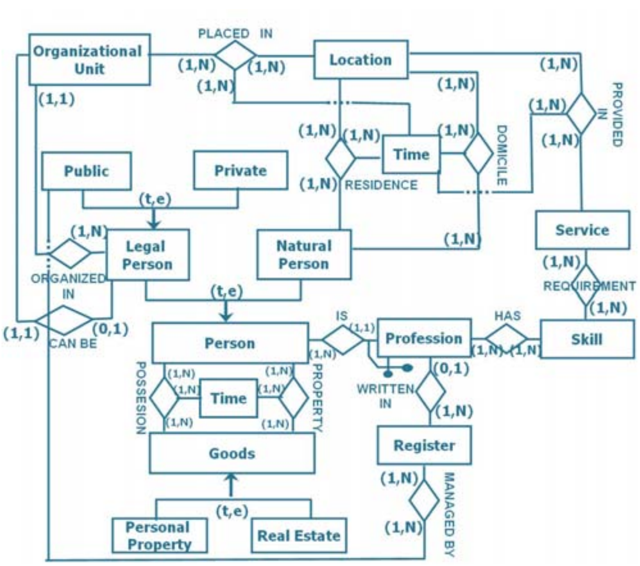


Figure 1. The proposed reference model – Part I

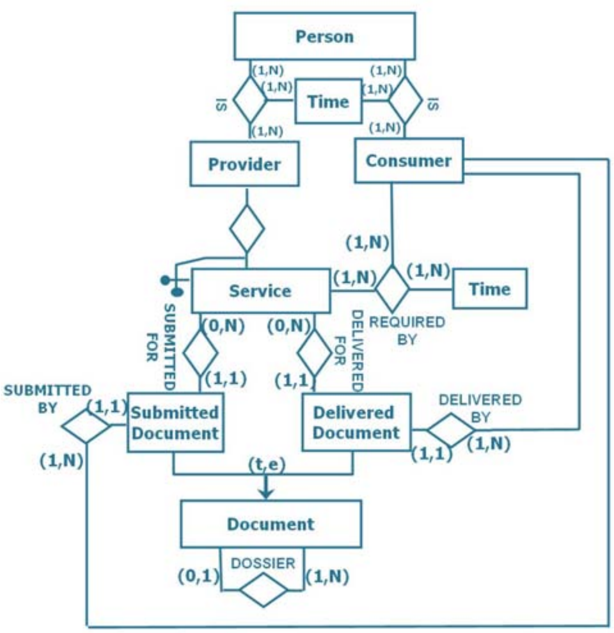


Figure 2. The proposed reference model – Part II

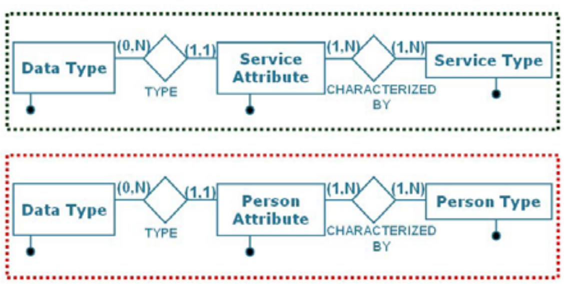
Service is an activity by means of which data are shared or exchanged. However, in this work the service entity models a broader concept. For example different services are modeled by means of the service entity (e.g., services of physicians, services of business consults, services of civil engineers). The same service can be provided by different providers at the same time. To uniquely identify each service, an external identifier (i.e., provider identifier), the starting date and the service identifier are exploited. Furthermore, since the same service can be provided in different locations and in different times, a many-to-many ternary relationship is defined between service, location, and time entities. The binary relationship between services and skill entities models the list of skills required to provide a given service.

Document can contain different information and is classified as delivered or submitted. To model this classification, a total and exclusive generalization is exploited (see Figure 2). For delivered documents, the service for which they are delivered and the user who has delivered them are known. These aspects have been modeled by means of one-to-many relationships among delivered documents and user entities, and delivered documents and service entities. Instead, for submitted documents, the service for which they have been submitted and the user who has submitted them are modeled by means of two one-to-many relationships among the corresponding entities. Finally, for any document the dossier in which it is included is modeled by means of a recursive relationship defined on the document entities.

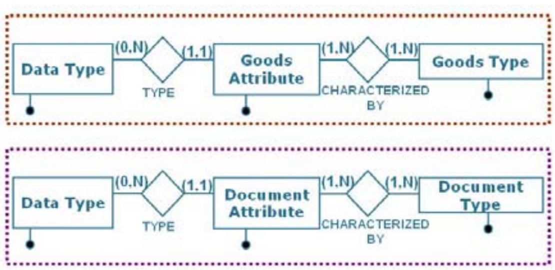
Upgrade of the cross-reference model. When the cross-application reference model is exploited to design a customized data sharing service in a given application domain, only the interesting concepts, micro-concepts and attributes are materialized. According to the application domain in which the reference model is exploited, different aspects of the same concept need to be modeled. Some attributes, some micro-concepts, or some generalizations may be undefined in the proposed reference model. Consider, for example, a local health unit which provides services to welfare beneficiaries. Different person types can interoperate (e.g., physician, welfare beneficiaries, local health unit employees). In this case at least four person types need to be modeled and each one is characterized by a set of different attributes. To address these issues, an *embedded meta-model*, exploited to upgrade our reference model, is proposed. The meta-model, as shown in Figure 3, is represented by means of a ER-schema. Thus, the ER-schema contains metadata extending the original reference model shown in Figure 1 and Figure 2. For each basic concept (i.e., person, service, good, document) the *type* is specified, modeling the interesting classification of the corresponding concept. For example, the local health unit may be interested in the following person types: Physicians, welfare beneficiaries, and local health unit employees. Each person type is characterized by a set of attributes with a many-to-many relationship between person attributes and person type entities, as shown in Figure 3. For each attribute, the corresponding data type exploited to represent it can also be modeled. The proposed upgrade strategy allows extending any entity defined in the original reference model (see Figure 1 and Figure 2).

By means of a set of transformation rules this metadata can be transformed in a new portion of the reference model. For example, records of the person type entity can define either a new generalization of the person entity or a set of new entities, one for each record. Each new entity is characterized by a set of attributes defined by means of the person attribute entity and the many-to-many relationship called *characterized by*. The data type of each attribute is defined by means of the *type* relationship shown in Figure 3. This rule can be exploited for any concept defined in the reference model.

Furthermore, the cross-application reference model can be upgraded by means of a new relationship among concepts. To address this issue, new metadata are included. The metadata are represented by means of a new entity called *Concept* with a recursive many-to-many relationship. The relationship is characterized by an attribute called *cardinality* which defines the relationship cardinality. By means of this template, new relationships among entities can be easily added to the reference model.



(a) Meta-models of the Person and Service concepts



(a) Meta-models of the Document and Goods concepts

Figure 3. Upgrade of the cross-reference model

DATA QUALITY DIMENSIONS

To support interoperability, data need to be characterized by a high quality. Poor data quality causes management issues in terms of both cost and efficiency for data consumers and providers. To efficiently support interoperability we need high quality data, which are intrinsically good, contextually appropriate for the task, clearly represented, and accessible to the data consumers and users [14]. Intrinsic data quality denotes that data have quality in their own right. Some indices have been defined to evaluate such aspect (e.g., accuracy, believability, reputation). Contextual data quality highlights the requirement that data quality must be considered within the context of the task at hand. Useful indices to evaluate contextual data quality are the completeness, the timeless, the value-added, the relevance, and the appropriate amount of data. The importance of the role of systems is emphasized by representational data quality (e.g., consistency in terms of representational consistency, internal consistency and external consistency) and accessibility data quality (e.g., accessibility, access security).

To enrich the reference model shown in Figure 1 and Figure 2, different data quality attributes can be included. Figure 4 shows the ER-schema exploited to model data quality dimensions and its values computed on a single attribute value, a single record of a given table (i.e., entity), or on a given table. In Figure 4, a generic entity called *ENTITY* is reported. This schema can be exploited for any entity defined in the reference model. According to the application domain, a subset of entities are selected to assess data quality indices.

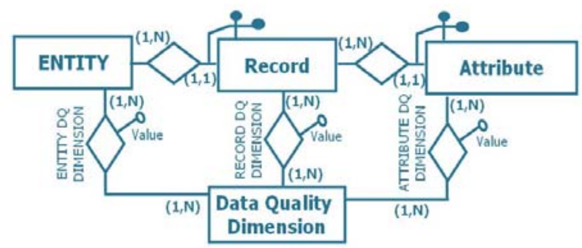


Figure 4. ER schema for data quality dimensions

REAL SYSTEMS

The proposed cross-application reference model has been exploited to build (i) a data model for actual data exchange in real systems, and (ii) a data model for a specific application. Details of the validation are described in the following.

AURA project

The AURA project, which is currently under development in CSI Piemonte, aims at providing a unique and centralized database for the regional registry office of welfare beneficiaries. Personal data and health information (e.g., SSN, payment exemptions, health service physician of choice) need to be shared to provide a unique view of such information. Different actors provide and access data to and from the AURA database. Actors can be classified into legal persons (e.g., local health units, regional health units, Ministry of Finance, etc.) and natural persons (e.g., welfare beneficiaries, physicians, public employees). Since a subset of natural persons (e.g., physicians) carry out professions which require to be publicly registered (e.g., into the medical register), the AURA data model contains the list of registered natural persons. To model such requirements, the *subject* entity with its own generalization, the *profession* entity, the *registry* entity and the corresponding relationships are materialized. To model other natural persons, the upgrade strategy presented above is exploited. A new *generalization on natural person* entity (i.e., welfare beneficiaries, physicians, public employees) is designed. It is characterized by three child entities and each child has a proper list of attributes.

The same subjects can be data/service providers or data/service customers or both at the same time or in different times. Different service data need to be stored (e.g., payment exemptions, health service physician of choice). For each service, a list of documents can be submitted by users (e.g., exam requests) or delivered to users (e.g., diagnoses). Thus, the data model needs to store any type of document for any service along with document provider and consumer. To model these requirement, *provider*, *customer*, *service*, and *document* entities, with the corresponding relationships, are materialized.

To address the data quality issue, a set of data quality dimensions [14], modeled as discussed above have been included in the AURA data model. They guarantee that data are intrinsically good, contextually appropriate for the task, clearly represented and easily accessible to data consumers. The detailed and complete AURA data model schema is reported in [5]. We can observe that the design of the AURA data model has been easily performed by means of the proposed reference model.

Piedmont Regional Information System

The Piedmont Regional Information System, developed by CSI Piemonte, is characterized by a centralized database in which many different types of data are stored. A small portion of this database, called *Test P-DB*, stores a staff registry, a consulting registry, the regional organization structure, and roles of regional workers. To validate the adaptability of the cross-application reference model, the design of the customized *Test P-DB* data model has been simulated. The regional organization structure is characterized by subjects that can be either a natural person or a legal person. A legal person can be either public (e.g., a local authority) or private (e.g., a company). Public legal persons are organized in a set organizational units. An organizational unit is usually located in different branches at different times. All these concepts are presented in the proposed reference model (see Figure 1), hence they are directly materialized. A natural person has a relationship with another subject by means of a relational role, which is characterized by a relationship duration. Each natural person can have more than one role in a lifetime. This requirement is modeled by means of an instance of the *Concept* template which is transformed in a many-to-many relationship between the natural person and the subject entities.

An organization can be higher in hierarchy with respect to another organization unit, it can have a commission in another organization unit, or it can be originated by another organization units, as its evolution. To model these issues, four different instances of the reference model *Concept* template are exploited. These instances are transformed into four recursive relationships on the organization unit entity. Regional workers (who are a subset of natural persons) can be classified into consultants and employees. Position and primary assignment are known for each employee, whereas the role and the contract duration are known for each consultant. By means of the upgrade strategy of the cross-application reference model, a new generalization is defined on the natural person entity. Furthermore, a consultant has a relationship with a private legal person, whereas an employee has a relationship with a public legal person. This issue is modeled by means of two binary (one-to-many) relationships. The first is between consultant and private legal person entities, while the second is between employee and public legal person. The detailed and complete *Test P-DB* data model schema, generated by means of the cross-application reference model, is reported in [5]. This data model presents the same features of the real system. Hence, our proposed approach is easily customized to derive a data model for a specific application.

THE INTEROP TOOL

To efficiently exploit the proposed canonical information model, the INTEROP tool has been designed. Given the XML schema of the cross-application reference model and the specific domain ontology, represented in OWL language [9], the INTEROP tool (see Figure 5) automatically designs the exchange data model for a customized data sharing service. Furthermore, a set of hints (e.g., new domain concepts, new domain relationships) can be provided by domain experts to the INTEROP tool. Since an ontology provides a classification of concepts within a domain and their relationships [6], it is exploited to tailor our reference model to the specific application. The more detailed the ontology is, the more customized the exchange data model will be. Ontology concepts are mapped in the cross-application reference model by means of a set of transformation rules. Transformation rules (i.e., a set of dependencies between application domain concepts and reference model templates) provide a powerful mechanism to deal with automatic design of complex customized data model. Dependencies, instantiated into a set of operative rules, can be classified into two classes: (i) *Ontology versus data model* and (ii) *metadata transformation*. The first set of dependencies define the mapping between concepts and relationships of the domain ontology and the data model templates. The second set of operative rules define how to perform the upgrade strategy of the cross-reference model (e.g., metadata, shown in Figure 3, are transformed in data model templates). Sometimes the specific domain ontology is not available. However, the proposed reference model simplifies the modeling task by providing guide lines for the modeling process and the modeling templates.

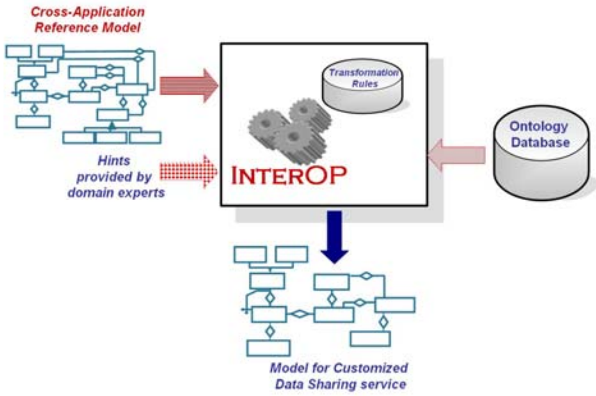


Figure 5. INTEROP architecture

RELATED WORK

Information modeling has proved to be useful in supporting data and process design [7]. By means of a reference model [10], information system research can import basic theoretical knowledge into real applications. A reference model consists of a minimal set of unifying concepts, axioms, and relationship within a particular domain, and it is independent of specific standards, technologies, and implementations [10][11][12]. [2] defines the reference model term as the abstract definition of how to describe and develop a domain of interest by means of: (i) Building blocks (usually abstract concepts) used to build models in the particular domain, (ii) relationships among these building blocks, (iii) a recipe for building specific models.

While many efforts have been devoted to (i) design enterprise specific models, referred to a concrete enterprise context, or (ii) define reference models for the specific domain of interest [11], less attention has been paid to the design of a general reference model independent of the application domain. Two different approaches have been exploited to design general purpose reference models. The first focuses on specifying the process of model development, while the second provides templates for specific models. Since reference models should provide appropriate generalizations of existing domains and aims at delivering blueprints for good system design, they are both descriptive and prescriptive.

[12] proposes a life event reference model at different abstraction levels. It provides templates able to model any life-event model. The reference model can be specialized to design a model of a specific life event in a specific country (e.g. getting married) or tailored to a set of specific user needs (e.g., applying for marriage, obtaining an identity card, obtaining a birth certificate). The approach presented in this paper is similar but more general, since it focuses on the design of a general reference model to support an effective information exchange (i.e., to allow two different organizations to interoperate). Our reference model provides a high-level definition of concepts and relationships, in terms of data and metadata (e.g., data provider description, data beneficiary/user identification, data quality dimensions).

Furthermore, many real systems are designed according to the model driven architecture (called MDA) approach [1]. MDA [1] defines an approach to IT system specification to separate the system functionality (e.g., business aspects) from the technological aspects (e.g., technological details of system functionalities). MDA models can be classified in three vertical layers: (i) Computation-Independent Models (CMI) which provide a domain model. A CIM does not show details of the system structure and uses a vocabulary that is familiar to the practitioners of the domain. (ii) Platform-independent Models (PIM) which do not contain any technology-specific information, and (iii) Platform-Specific Models (PSM) that include information about the specific technologies and platforms exploited in the model development. The last two layers present different levels of abstraction. The first is more general, whereas the second is specific, since it is derived from the PIM using a transformation mechanism. Hence, PSMs possibly contain elements that are specific to the platform. The proposed cross-application reference model can be profitably exploited to build the CMI level of any MDA system. Furthermore, without showing details of the system structure, it defines a set of unifying concepts and relationships which are independent of the specific application domain. The proposed model can be exploited for the definition of customized data sharing services in any given application domain.

CONCLUSIONS AND FUTURE WORK

In this paper a canonical information model has been proposed. It provides model templates and guide lines for the modeling process. A set of general concepts, useful to model data, services, and metadata (i.e., actors, data quality dimensions) to support interoperability, have been identified and characterized. The proposed reference model can be exploited by any domain expert to design data sharing services in public sector domain. To validate the adaptability and effectiveness of the proposed approach, we exploited the reference model to build a data model for actual data exchange in a real system and a data model for a specific application.

We are currently defining a set of operative rules to automatically map domain concepts into the reference model templates. Furthermore, we are currently implementing a Java prototype of the INTEROP tool framework to automatically design and customize data sharing services. Since the IBM Center for Advanced Studies of Rome within the project SPCoop (Semantic Integration of Italian eGovernment Services) [13] are currently developing a database ontology for eGovernment services, we plan to exploit it to extend the validation of the presented reference model.

REFERENCES

[1]. OMG Model Driven Architecture [Online]. Available: http://www.omg.org/mda/.

[2]. SmartGov Project: A Governmental Knowledge-based Platform for Public Sector Online Services [Online]. Available: http://www.smartgov-project.org/.

[3]. Manual on Statistics of International Trade in Services. Eurostat, IMF, OECD, UN, UNCTAD, WTO. Annex II, Glossary, 2002.

[4]. European Interoperability Framework for pan-European eGovernment Services. Luxembourg, European Communities 2004.

[5]. T. Cerquitelli. Technical Report. [Online]. Available: http://tasmania.polito.it/twiki/bin/view/Public/TaniaCeruitelli/ERmodel

[6]. C. Curino, E. Quintarelli, and L. Tanca. Ontology-based information tailoring. In 22nd International Conference on Data Engineering Workshop. IEEE Computer Society, 2006.

[7]. Fowler M. Analysis Patterns: Reusable Object Models. Addison-Wesley, 2001.

[8]. ISO. ISO/IEC 2382-01 Information Technology Vocabulary, Fundamental Terms. 2003.

[9]. D. L. McGuinness and F. V. Harmelen. OWL Web Ontology Language Overview, W3C Recommendation. [Online]. Available: http://www.w3.org/TR/owl-features/.

[10]. O. Thomas. Understanding the term reference model in information systems research: History, literature analysis and explanation. In Book Business Process Management Workshops, volume 3812 of Book Series Lecture Notes in Computer Science, pages 484–496. Springer, 2006.

[11]. O. Thomas and A. Scheer. Tool support for the collaborative design of reference models. a business engineering perspective. System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on, 1, 04-07 Jan. 2006.

[12]. L. Todorovski, M. Kunstelj, and M. Vintar. Reference models for e-services integration based on life-events. In M. Wimmer, H. J. Scholl, and A. Gronlund, editors, EGOV, volume 4656 of Lecture Notes in Computer Science, pages 92–103. Springer, 2007.

[13]. G. Vetere. Semantic Integration of Italian eGovernment Services. [Online]. Available: http://www.dis.uniroma1.it/lenzin/INFINT2007/material/Vetere.pdf.

[14]. R. W. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. In Journal of Management Information Systems, volume 12, pages 5–30. Springer, 1996.

.