Extended Scheme Mediation Integration Model for Information Systems Project Proposal

William Chaves de Souza Carvalho[1], Pedro Frosi Rosa[2], Flávio de Oliveira Silva[3]
[1] Federal University of Uberlandia, william.carvalho@ufu.br
[2] Federal University of Uberlandia, pfrosi@ufu.br
[3] Federal University of Uberlandia, flavio @ufu.br

Abstract. The demand for semantic sharing between information systems is steadily growing in virtually all business domain. In fact, these systems form the basis of business management, lead economic and decision-making strategies, and manage communication with partners. Interoperability has become a priority requirement to and in order to meet the needs of applications have autonomy, scalability, transparency and extensibility characteristics which is crucial for data conflicts not to hamper efforts to integrate and transparently share information from a variety of sources. Research has shown that the evaluation of data semantics is a promising approach to dealing with this problem. This work presents the project of creating a model based on mediation of schemas to facilitate the understanding of shared data and help to identify and adapt data relevant to the user context.

1 Introduction

Information Systems (IS) face two major challenges in their operation. On an hand, there is a need to ensure interoperability between the heterogeneous systems that are developed, in general, independently based on a single domain representation. On the other hand, the operational environment is dynamic which leads to continuous changes in these systems. Thus, it is necessary to optimize the process of reusing the modules between the different systems to accelerate the developments. In analyzing these challenges, we find that the problem of semantic sharing is a major factor that degrades interoperability and prevents reuse between ISs.

The possible way to deal with the problem of semantic heterogeneity is to reduce or even eliminate terminological and conceptual incompatibilities. Therefore, establishing a common understanding (terminological and conceptual), with multiple points of view, can help establish a communication base between users, manage interoperability between systems and improve the engineering process of reuse.

Over the centuries, the constant evolution of communication technologies has made the sharing human impetus increasingly more important. Sharing information is not a new idea and in this paper, we focus on some challenging interoperability issues that emerged with the evolution of communication technologies.

Neither the sharing of information between people nor the exchange of data between information systems are new ideas. Nevertheless, new challenges are created

by the improvement of interconnection technologies of computational agents and improved algorithms for processing and retrieval of large volumes of data. The solutions must meet different requirements and criteria to accomplish the user needs [1]: autonomy of the systems, extensibility of the architecture to control the addition or removal of IS, scalability of the system capacity faced with the increase on the users amount and transparency regarding location and data format. The solutions are based on the notion of interoperability, that is, the implementation of a such collaboration between systems that allows data and service sharing to respond to information requests. Resolving a query in an interoperable environment involves integrating distributed data across multiple heterogeneous systems raises some questions:

1. Which ISs may contain all or part of the desired information?
2. What data are relevant to the expressed needs?
3. How to transform, to adapt and to pair the data in order to obtain a consistent and user-friendly result?

The information integration is preceded by resolution of three types of conflicts:

- Syntactic conflicts result from using different data models from one system to another; beyond what, different concepts are used to structure the same information.
- Schematic conflicts arise from different structures and classifications of information. This is closely related to the choices made during designing the project.
- Semantic conflicts arise from differences in the interpretation of information shared by various application domains: nomenclature conflicts (taxonomic and linguistic problems), value conflicts (unit problems, scales), and cognitive conflicts (meaning problems).

Schematic and semantic conflicts are closely related. The meaning of information must be clearly defined, and semantic conflicts resolved before schematic conflicts are identified and addressed.

Many solutions have been proposed to consider these data conflicts, that guarantee at least the autonomy, the composability and the resolution of syntactic conflicts.

Multi-base approaches rely on the use of a query language and a common representation model [2]. Each SI exports its information as a schema described in the common model (usually an object-oriented model), the multi-base language, extended from SQL (Structure Query Language) or OQL (Ontology Query Language), allows a multi-site query. The multi-base approach is extensible but does not offer localization transparency. The resolution of schematic and semantic conflicts remains entirely on user's responsibility.

Federated approaches rely on integration. Each SI exports a schema in a common model, so the various exported schemas are integrated into federated schemas. A federated schema allows access to shared data in a uniform and global way [3]. Schematic and semantic conflicts are solved by the integration process. Transparency is ensured, just by posing a query on the federated schema. However, extensibility and scalability are criteria that are not adequately met; which indicates that federation is an approach to be considered when integration involves a small number of information systems and when these systems go through few evolutions.

Mediation approaches extend the federation by providing more flexibility. Mediation rely on an essential component, named mediator, which is responsible for meeting the integration needs based on the knowledge made available [4], [5]. The mediator finds the available information and solve the detected schematic and semantic conflicts. A secondary component serves as interface to the ISs and solve syntactical conflicts by providing data to the mediation model.

Two types of mediation can be distinguished according to how data conflicts are resolved:

1. Mediation of schemes that previously builds a database of information about the participating ISs which gives the mediator the means to do his integration work.
2. Context mediation in which the mediator uses semantic information to resolve queries dynamically without prior knowledge of which information systems belong to the context.

Schema mediation is a direct extension of the federated approach [6], [7], [1] with better scalability and often better scalability (object interfaces, rules-based language) [8], [9]. Context mediation seeks to discover data semantically close. This such mediation can find and adapt information to ensure total transparency [10], [11], [12], [13], [21], [22], [27], [28].

This paper presents the main aspects of a under development extended scheme mediation integration model for information systems project proposal. When finalized, the model will be composed by a methodology, a model and an architecture to build a knowledge base to integrate information corresponding to well-defined usage profiles. The mediation model will make it possible to represent the data structure in a common model, adapt and compose the information in order to ensure total transparency. This model will also associate semantics with shared information to compare and identify relevant data. An incremental integration mechanism guided by semantics will allow its adaptation to dynamic and broad environments.

Section 2 provides an overview of solutions based on context mediation and then introduces the key points of for a integration model based on extended scheme mediation. Section 3 describes a typical example of a cooperation environment and explains how our approach can solve problems related to interoperability of IS. Section 4 concludes the article and presents research perspectives.

2 Related Work

Context-mediated solutions are based on models of knowledge representation capable of describing, to a certain extent, the semantics conveyed by information and tools for comparing and unifying the semantics of information independently of the inherent structures. The notion of semantics of an entity can't be represented in an absolute way. An entity must be linked with other entities for a meaning to be associated with it. The semantic description of an entity only makes sense in relation to a specific context. A context can be defined as the finite set of concepts, relationships, constraints, and rules that describe an application domain. Each solution defines contexts on which the semantic aspect of shared information is based. The

information associated with the same context can then be easily classified and compared since they share the same semantics: it is the phase of semantic approximation. This phase is more delicate when it comes to reconcile two information respectively defined on two previously disjointed contexts. It is then a question of reconciling the contexts to estimate if it is possible to reconcile this information and to deduce the way to compose them.

A context can be represented by a conceptual diagram that describes a domain area of application. It is composed of concepts related to each other (inheritance relations, compositions relations, logical relations). A concept is usually defined from the content of an ontology. The ontology [14] is the specification of a conceptualization, the expression in a language of terms and concepts. An ontology can be represented using a semantic network, a terminology graph, a conceptual graph, a set of logical rules, or an object-oriented schema [4]. The content of an ontology remains close to a specific field of application even if some projects are moving towards the definition of generic ontologies [15].

Solutions are categorized into two groups depending on whether the mediation uses a single ontology or inter-ontological relationships. In the first case, single ontology, the contexts of the information are constructed from the same conceptual universe and the contextual reconciliation passes by a direct comparison of the concepts which then have a common basic semantic unit.

The Coin project [12] is based on the description of the semantics of the values of the exchanged data. Each piece of data specifies the semantics of its values by indicating, for example, the type of unit and its meaning. An ontology provides a vocabulary to describe the meaning of these values. Each SI has a local context that describes the semantics of values manipulated locally. A reference context describes the domain of use of the values and the transformation rules making it possible to reconcile the semantics of values. These rules use the frame logic to derive a translation path from a source context to a target context and adapt the representation of the information [6]. The user can query the system without worrying about the format of the data presented to him in his local format. The identification of semantic conflicts is restricted to differences in the interpretation of values. Localization and composability remain the responsibility of the user.

On2broker [13] is particularly interested in extracting data from the content of web pages. Interoperating here consists in extracting the relevant data from the content of the web pages, discovering the links with information located on other pages, combining this data. A semantic model adds information to the content of web pages in the form of proprietary HTML tags in the manner of XML. Ontology is defined as object classes and deduction rules that extend the F-logic language [16]. These are references to these classes that are linked to the content of web pages. A research and analysis process continually scam participating web sites to build a global fact base that will be used to infer ontology rules about the facts. Ontology takes the appearance of a global schema to access data. The construction of a fact base compromises the scalability of the system.

The Observer project [10], [17] is based on a hierarchy of ontology servers described in terminology logic to define utilization domains. Each SI must define its own data in relation to the concepts of one or more ontologies. Rewriting rules and transformation functions define the local interpretation of information. A user asks a

terminological query on the concepts of an ontology. This request can be sent over the entire domain and only servers whose context agrees with that of the request provide answers. Semantic reconciliation is based on predefined 2-2 inter-ontology relations and synonymic relations.

InfoSleuth [18] is based on an agent architecture. Resource agents store the information provided by the data sources using the vocabulary of ontological agents. Each resource must be registered to a broker agent to ensure the location of the information. A user agent provides a query interface to users and manages a query using broker functionality. Information discovery is supported by specialized agents that extend the architecture to handle multiple ontologies simultaneously through predefined inter-ontology relationships.

The use of ontologies to describe a universe of speech is the common point of many solutions. Defining an ontology server as a semantic reference is nevertheless not without problems:

1. Identifying the concepts of an ontology requires a consensus on the part of the interveners.
2. Modeling of the ontology is a delicate step.
3. Semantic universe is limited by the content of the ontology.

Taking into account several ontology servers brings out the problem of matching between heterogeneous representations [19].

3 Semantic Approach

Our proposal will rely on incremental integration of relevant information dynamically discovered on information systems. It will take advantage of the robustness of the schema mediation approach and will combines it with the semantic matching techniques of context mediation. The integration work is simplified once the information has a contextual description that guides their integration it:

1. Allows the understanding of shared data.
2. Supports contextual reconciliation for comparing contexts and identifying information in semantic relation.
3. Ensures the location of the information.
4. Facilitates the generation of the integration information and adapts the data to a reference context.

Our model will adopt a technique of contextual reconciliation independent of an ontological language, based on the specification of local contexts. That way, information identified on heterogeneous but overlapping contexts can be compared without the definition of inter-ontology relationships. Contextual reconciliation will be based on a technique of comparing the concepts that make up two contexts. So, the similarity of concepts can be evaluated according to their taxonomic similarity, their intrinsic properties, and their semantic neighborhood (that is, relationships between concepts). We define a concept as an elementary semantic unit that could be represented using the triangle known for applications of artificial cognition [20].

A concept is a link between three notions: the referent, the signifier and the signified. The referent is what a concept represents, it is here an informative object

which constitutes the class object of a shared information. The informational object is the basic element for exchanging data through our mediation model. The signifier gathers all the knowledge allowing to give a meaning to the concept, the taxonomic information (synonyms, antonyms), the properties (semantics of values, constraints) and the relations with the neighborhood. The signified is the conceptual class that represents the abstraction of an entity.

As SIs can exports mediation model object classes, which provide a local context, these classes could be defined in relation to their local context, so the mediation model will describe both the structure and the semantics of the data. In other hand, semantic reconciliation between the contexts of two SIs allows to import the relevant classes and integrate them.

3.1 A Cooperation Scenario

This subsection describes an example of interoperability between heterogeneous systems. An association that wants to provide an information service on the concerts of a variety artists across Brazil.

S1(nbC, artistN, dateC, freeP, soldP, priceP)
nbC (integer): Identifying Number concert id (local to system)
artistN (string): Artist name
dateC (date): Date of the concert
freeP (integer): Number of free places
soldP (integer): Number of seats sold
priceP (float): Price of a seat (in Brazilian Reals)
S2(id, name, session 1..n, nbP, totP, ticket)
id (integer): Number identifying a concert (local to the system)
name (string): Name of the artist
session 1..n (date): Date of the session
nfP (integer): Number of free places
totP (integer): Total number of places
ticket (float): Price of a seat (in dolars)
S2 has a function of transforming the Brazilian Reals into dolars in the form of a macro named rtod.

Fig. 1. Schemes of s1 and s2 Information Systems

The information offered by this association includes for each concert the name of the artist, the date of the concert, the number of places still free, the price of places and the room where it takes place.

A user can query this system to obtain an integrated view of the information distributed on various information sources. To simplify the example, we consider the existence of two systems *s1* and *s2* relating to two concert halls and likely to give information on concerts.

The system *s1* is administered by a relational DBMS (a table S1) while *s2* uses a spreadsheet to store the data (a worksheet S2). The organization of the information is described in Fig. 1.

Sources *s1* and *s2* pose various problems of heterogeneity. The data models are different, name conflicts appear between similar entities, value conflicts appear in the price of places, cognitive conflicts appear in the way of considering concert dates. Each system begins by exporting the schema of its information in a format specific to our model.

The local representation of the data is translated into the mediation model in the form of object classes. A domain specialist constructs the context of the information source and links classes to that context to define the semantics of each class. The context can then be recorded at the level of a directory or register.

The association must discover potentially interesting sites and extract the information in line with its needs. A domain expert is responsible for defining the context of the application that becomes the reference context when a system plays the role of consumer information.

The context specification is compared with all or a subset of the contexts of the sites registered in the directory. The source contexts are reconciled with the reference context, the relevant object classes are imported and adapted to the semantics of the reference context. Imported classes can then be interpreted in the application domain.

Virtual classes provide full access transparency, providing a seamless and integrated interface for shared information. The construction of virtual object classes consists in identifying the classes of similar or related objects by using the contents of the reference context and generating matching rules between these classes.

Object classes and integration rules form a static knowledge base for resolving queries. The dynamic aspect is based on the possibility of incrementally increasing the content of this knowledge by using contextual reconciliation to import and integrate new object classes.

3.2 Language and class representation

The description of the shared information will use a mediation language that both describes object classes in an object-oriented model [16], [26] and the semantics of these classes through a context derived from terminological logic [27].

An object class represents a set of objects that respond to a common structure. A class is defined by its name and by the list of its methods that characterize the properties (attributes and operations) of the objects of the class.

The object is represented both by its state (the properties values) and a unique internal identifier.

Information sources *s1* and *s2* provide classes `cs1` and `cs2` for cooperation. The purely local information such as the identifiers disappear, the attributes of the site relation *s1* and the column headings of the site spreadsheet *s2* become the methods of the classes `cs1` and `cs2`.

The formats of numbers and dates are standardized in the model primitive types. The representation of the number of numbered sessions of *s2* uses the concepts of the model using a session method set with the session number.

The Brazilian Reals to American dolars conversion function of *s2* also becomes a parameterized method as `rtod`.

```
class cs1[repository _'s2.uberlandia.br';
       description _'Uberlandia Municipal Theater Events';
       artistN := string;
       dateC := date;
       freeP := integer;
       priceP := float].

class cs2[repository _'s1.pittsburgh.us';
       description _'Pittsburgh Heinz Hall Information System';
       name := string;
       session(integer):= date;
       nfP := integer;
       totP := float;
       ticket := float;
       rtod(float) := float;
```

3.3 Context Representation

A context in our model will be a semantic network, that is to say a set of terminological concepts [23], [24], [25] related to each other. To understand a concept is to allow the interpretation of the classes of objects associated with it. A context is composed of several semantic elements: the concept that forms the basic semantic unit, the role that models a binary relation between concepts, specifies relations of generalization / specialization between the concepts on the one hand and the roles on the other hand, the taxonomy which defines relations of synonymy and antinomy between the terms of the terminology of the context, the interpretation which allows an object class to instantiate a concept.

4 Conclusion and further work

In this paper, we have presented an extended scheme mediation integration model for information systems project proposal. Our approach uses a model that describes the shared information by taking their semantics into account in the context of usage contexts. The definition of local contexts location and use of shared and heterogeneous information turn our work in the direction of further formalizing the notion of semantic distance and moving towards the implementation of architecture. The next steps of creating our model will evolve tailoring a methodology based on contextual matching and flexible architectures which will allow incremental integration of relevant information in the context of information systems. Further steps will also be responsible to formalize the definition of local contexts, contextual matching and semantic distance to realize the interpretation, finding and use of shared and heterogeneous information. Our current work goes in the direction of further formalizing the notion of semantic distance and moving towards the implementation of an architecture.

References

1. Liu, L., Pu, C.: Approach to Large-scale Interoperable Database Systems. Technical report TR95-16, University of Alberta (1995).
2. Floresco, D., Rachid, L., Valdurez, P.: Using Heterogeneous Equivalences for Query Rewriting in Multidatabase Systems. Proceedings of the Third International Conference on Cooperative Information Systems, (1995).
3. Shet, Amit P., Larson, James A.: Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Surveys, Vol. 22, n3, September (1990).
4. Wiedershold, G.: Mediators in the Architecture of Future Information Systems. IEEE Computer Magazine, Vol. 25, n3, March (1992).
5. Wiedershold, G.: The Basis for Mediation. Proceedings of the Third International Conference on Cooperative Information Systems, (1995).
6. Cody, William F., Has, Laura M., Nblack, W., Arya, M., Carey, Michel J., Fagim, R., Flickner, M., Lee, D., Petkovick, D., Schwarz, Peter M., Thomas, J., Roth, Mary T., William, John H., Wimers, Edward L.: Querying Multimedia Data from Multiple Repositories by Content: the Garlic Project. Proceedings of the third IFIP, (1995).
7. Tonasic, A., Rasschid, L., Valdurez, P.: Scaling Heterogeneous Databases and the Design of Disco. Proceedings of the 16th International Conference on Distributed IEEE Computing Systems, IEEE Computer Society Press, May (1996).
8. Garcia Molina, H., Quas, D., Papakostantinou, Y., Rajarman, A., Sagiv, Y., Ullman, Jeffrey D., Widon, J.: The TSIMMIS Approach to Mediation: Data Models and Languages. Proceedings of the Second International Workshop on Next Generation Information Technologies and Systems, (1995).
9. Papakonstatinou, Y., Abitebul, S., Garcia Molina, H.: Object Fusion in Mediator Systems. Proceedings of 24rd International Conference on Very Large Data Bases, (1996).
10. Kashyap, V., Sheth, A. Semantics-Based Information Brokering. Proceedings of the 3rd International Conference on Information and Knowledge Management, Gaithersburg, ACM Press, (1994).
11. Bishr, Y.: Semantic Aspects Interoperable GIS. Phd Thesis, Enschede Netherlands, (1997).
12. Bressam, S., Goh, C. Hian, Fiynn, K., Jakobisiack, M., Husein K., Kon, Henry B., Lee, T., Maddnick, Stuart E., Pena, T., Qu, J., Shum, Annie W., Siegel, M.: Context Interchange Mediator Prototype. Proceedings of ACM SIGMOD International Conference on Management of Data, (1997).
13. Deker, S., Erdmamn, M., Fensel, D., STUDER, R.: Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. Eighth Working Conference on Database Semantics, DS-8, Kluwer, (1999).
14. Grubber, T.: A Translation Approach to Portable Ontology Specifications. International Journal of Knowledge Acquisition for Knowledge-based Systems, Vol.5, n2, June (1993).
15. Guarino, N.: Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration, Summer School on Information Extraction, Frascati, Italy, July 1997 (1997).
16. Kiffer, M., Lausem, G., Wu, J.: Logical Foundations of Object-Oriented and Frame-Based Languages. Journal of ACM, Vol.42, n4, (1995).
17. Mena, E.: OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. Phd Thesis, University of Zaragoza, (1999).
18. Fowler, J., Perry, B., Bargmeyer, B.: Agent-Based Semantic Interoperability in InfoSleuth. ACM SIGMOD Record, Vol.28, n1, (1999).
19. Nicolle, C., Cullot, N., Yetongnon, K.: A Translation Process Between Cooperating Heterogeneous Database Systems. Proceedings of PDCS97, New Orleans, (1997).

20. Grumbach, A.: Cognition artificielle. Du réflexe à la réflexion. Ed. Addison-Wesley France, (1994).
21. Vitvar, T., Kopecky, J., Viskova, J., Fensel, D.: WSMO-Lite Annotations for Web Services. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 674–689. Springer, Heidelberg (2008)
22. Farrell, J., Lausen, H. (eds.): Semantic Annotations for WSDL and XML Schema. W3C Recommendation (2007).
23. McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language Overview. W3C Recommendation (2004).
24. Hibner, A., Zielinski, K.: Semantic-based Dynamic Service Composition and Adaptation. In: IEEE SCW 2007, pp. 213–220 (2007)
25. Gomadam, K., Verma, K., Brewer, D., Sheth, A.P., Miller, J.A.: A tool for semantic annotation of Web Services. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005, vol. 3729, Springer, Heidelberg (2005)
26. Gmez-Prez, J. M., Erdmann, M., Greaves, M., Corcho, O., & Benjamins, R. A framework and computer system for knowledge-level acquisition, representation, and reasoning with process knowledge. International Journal of Human Computer Studies. (2010)
27. Krotzsch, M., Simancik, F., & Horrocks, I. Description logics. IEEE Intelligent Systems. (2014)
28 Kurniawan, K., & Ashari, A. Service orchestration using enterprise service bus for real-Time government executive dashboard system. In Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015. (2016)
29. Parlanti, D., Pettenati, M. C., Bussotti, P., & Giuli, D. Improving information systems interoperability and flexibility using a semantic integration approach. In Proceedings - 4th International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution, Axmedis 2008. (2008)