Semantic Data Integration in a Newspaper Content Management System

A. Abelló[1], R. García[2], R. Gil[2], M. Oliva[2], F. Perdrix[2,3]

[1] Univ. Politècnica Catalunya
aabello@lsi.upc.edu

[2] Universitat de Lleida
{oliva,rgil}@diei.udl.es
roberto@griho.net

[3] Diari Segre S.L.U.
perdrix@diarisegre.com

Abstract. A newspaper content management system has to deal with a very heterogeneous information space as the experience in the Diari Segre newspaper has shown us. The greatest problem is to harmonise the different ways the involved users (journalist, archivists…) structure the newspaper information space, i.e. news, topics, headlines, etc. Our approach is based on ontology and differentiated universes of discourse (UoD). Users interact with the system and, from this interaction, integration rules are derived. These rules are based on Description Logic ontological relations for subsumption and equivalence. They relate the different UoD and produce a shared conceptualisation of the newspaper information domain.

1. Introduction

From our experience in the newspaper content management systems domain[1], it has been possible to develop an experience of Semantic Web technologies in a real setting in the Diari Segre [1]. The main contributions are:

- An ontological framework for the newspaper domain [2].
- A semantic search and exploration user interface [3].

However, despite these achievements and as result of the experience acquired thereof, some aspects of this practical approach to a semantic newspapers have to be improved. Fundamentally, the main issue is the gaps among the different users' conceptual models and the ontologies that try to formalise a shared conceptual model.

Users of a newspaper system need to collect, to organise and to share lots of information about news. It is very difficult that different users classify a piece of news with the same topic, subject or keywords. This situation leads the users of the newspaper system to easily miss the needed information. The main problem with the Diari Segre newspapers content management system is the gap between journalists' keywords and the topics used by archivist for classification. This gap prevents journalist from finding the content they need during their daily work. Thus, it makes them asking archivists to locate content for them, which is an overhead for them.

[1] This work is funded by the Spanish Ministry of Science and Technology, grants FIT-150500-2003-511, TIC2002-1948, and TIN2005-05406

Altogether, this is a data integration problem caused by the interaction of different conceptual models. In order to overcome it, our approach is to formalise these conceptual models using ontologies. Once formalised, it is possible to employ computerised methods based on Description Logics in order to build up an integration service based on users' interaction.

2. Semantic Methodology for Data Integration

[4] argues for the importance of interactive and iterative integration. They present a tool that helps such process by looking at the instances to guide the conflict resolution at the schema level. In our case, we do not aim at solving any conflict, by assuming that they exist due to the different points of view of users. Thus, this will be handled by making their different points of view explicit, i.e. those known instances, also known as Universe of Discourse (UoD).

Regarding the criteria proposed in [5] to classify semantic integration approaches, ours would take the following values:

- Who generates the mappings: Agents themselves
- When define Agent-to-Agent mapping: Auto-generated at agent interaction time.
- Topology: Mediated.
- Degree of Agreement: Agree on subsumption/overlapping of Universes of Discourse (UoD).

Therefore, we have introduced a new value for the degree of agreement, and our proposal does not fit in any of the five architectures proposed there. Those values given to this attribute in [5] are all based on the ontologies, while we do not assume any agreement on the ontologies being used, but on the instances known by each user.

References

1. Diari Segre media group, http://www.diarisegre.com
2. García, R.; Perdrix, F. and Gil, R.: " Ontological Infrastructure for a Semantic Newspaper". In "Semantic Web Annotations for Multimedia Workshop, SWAMM 2006". 15th World Wide Web Conference, Edinburgh, UK, 2006
3. Castells, P.; Perdrix, F.; Pulido, E.; Rico, M.; Benjamins, R.; Contreras, J. and Lorés, J.: "Neptuno: Semantic Web Technologies for a Digital Newspaper Archive". Springer, LNCS Vol. 3053, pp. 445-458, 2004
4. Sattler, K.-U.; Conrad, S. and Saake, G.: "Interactive example-driven integration and reconciliation for accessing database federations". Information Systems 28(5), pp. 394-414, 2003
5. Uschold, M. and Grüninger, M.: "Architectures for Semantic Integration". In Kalfoglou, Y. et al (eds.): "Semantic Interoperability and Integration", Dagstuhl Seminar Proceedings, Num. 04391, Schloss Dagstuhl, Germany, 2005