# On Ontology-Based Mediation for Cultural Heritage Data

Béatrice Bouchou Markhoff, Cheikh Niang

# On Ontology-Based Mediation for Cultural Heritage Data

Béatrice Bouchou Markhoff[1], Cheikh Niang[1],

[1] Université François Rabelais de Tours, Laboratoire d'Informatique, France
{beatrice.bouchou, cheikh.niang}@univ-tours.fr

**Abstract.** We present our generic semantic mediation system that we have recently begun to apply to several fields of digital humanities. We explain its overall architecture and how it runs, then we report on one application, the Personae project that aims to query multiple prosopographical data sources, before discussing its potential usefulness for other projects in Cultural Heritage.

**Keywords:** Information Integration, Semantic Web

## 1 Introduction

The semantic web is an evolution of the current web towards a large area of resources sharing and exchange. It aims at representing, for machines, the meaning of the mass of information available on the web. The aim is to better automate the use of this information. There are two main notions at the basis of the semantic web, the first one is called linked data [1], which consists of technical principles for building a giant graph of data over the web, and the second one is the notion of ontology. Defined by Thomas Gruber [2] as a «formal, explicit specification of a shared conceptualization", the ontology is the model and the structural framework that allows formalizing and organizing knowledge in the semantic web level.

If ontologies contribute to make the semantic web vision a reality, in return, it is thanks to the semantic web that ontologies gained standard formalisms such as OWL[1], and RDFS[2]. In addition, the semantic web development has promoted the normalization and the more and more accessibility of ontologies that describe precise domains. For instance AGROVOC[3] in the agriculture fields, SNOMED CT[4] in the medical scientific disciplines and CIDOC CRM[5], related to cultural heritage area, are some examples of what we will call *reference ontologies* further on. Knowledge engineering experts, in collaboration with domain experts, develop these ontologies in such a way that they are robust conceptualizations of the knowledge about the given domain, for instance agriculture, medicine, cultural heritage, etc.

[1] http://www.w3.org/TR/owl-ref/
[2] http://www.w3.org/TR/rdf-schema/
[3] http://www.fao.org/agrovoc
[4] http ://www.ihtsdo.org/
[5] http://www.cidoc-crm.org/official_release_cidoc.html

Due to their ability to explicitly represent data semantics, ontologies play also an important role in the data integration process, by capturing the meaning of data drawn from heterogeneous sources [3]. Data integration tackles the problem that involves combining data residing in different sources, and providing users with a unified view of these data sources. This process is important and appears with increasing frequency as the need to share existing data increases, which is, for instance, the case for historical and archaeological domains. Data integration is a broad research topic that has received for many years the attention of researchers in databases and knowledge engineering. In the last decade, with the growth of the semantic web and the standardization of ontologies, data integration has taken a new direction by adopting *ontologies as pivots* for data integration. When the access to heterogeneous data sources is made possible using ontologies, the integration process is called ontology-based data integration [4]. The ontology's formal semantics allows the automation of tasks such as consistency checking, inference, query-answering, etc.

Semantic web and data integration are complementary paradigms when it comes to overcome semantic heterogeneity, in order to share and reuse efficiently data among autonomous interconnected stakeholders. This interoperability is needed in particular for achieving digital humanities purposes, and more especially for the cultural heritage knowledge dissemination. It allows for providing environments and tools for producing, diffusing, sharing and upgrading this knowledge. For instance, here are some projects we are involved, for designing semantic data integration solutions:

- Interconnecting several different historical databases containing descriptions of people and their relationships during the Middle Ages and the Renaissance, in order to perform prosopographical studies (Personae project of CESR[6]).
- Building a dynamic web application for accessing Western-Sahara archives and manuscripts (Biblimos project of the EMAM team of CITERES[7]).
- Building a common search interface for querying multiple databases tracing the Renaissance art in the Val-de-Loire area (Arviva project of CESR).
- Making interoperable several heterogeneous archaeological data sources (ArSol project of the LAT team of CITERES).
- Querying uniformly different data sources, containing both factual descriptions of heritage objects, text documents and references, as well as images, spectral analysis and other instrumental measurements and methods of work related to conservation-restoration careers (the Parcours project of PATRIMA[8]).

In this paper, we present an ontology-based mediation framework, which is a generic data integration solution that may be tailored for each application. Its genericity relies on the use of a reference ontology. To tailor it for digital humanities, we first proposed to build the mediation schema from available general ontologies,

[6] http://umr6576.cesr.univ-tours.fr/
[7] http://citeres.univ-tours.fr/spip.php?article1
[8] http://www.sciences-patrimoine.org/index.php/parcours.html

such as YAGO2[9] or DBpedia. Now, we are studying how to use CIDOC-CRM, which covers more precisely the cultural heritage domain.

The paper is organized as follows: in Section 2 we describe the main principles of our proposal, and its usage scenario. In Section 3 we present one application, the Personae Project. We conclude in Section 4.

## 2 A Semantic Mediation Framework

There are two main classical approaches for data integration, the data warehouse and the mediation [5]. The first one consists in (i) extracting information from the source databases, (ii) transforming it into a common structure and then (iii) loading it into the data warehouse. Its main advantage is that all source data are physically reconciled in a single repository, which allows for efficient query evaluation. Its main drawback is due to data duplication: updates performed on the original data source must be propagated to the warehouse. For instance, the STAR project[10] is one example of such an approach in the domain of semantic web integration for archeological data sources: data from different institutions are extracted, transformed and loaded in a RDF triplestore, as illustrated in the left part of Figure 1.

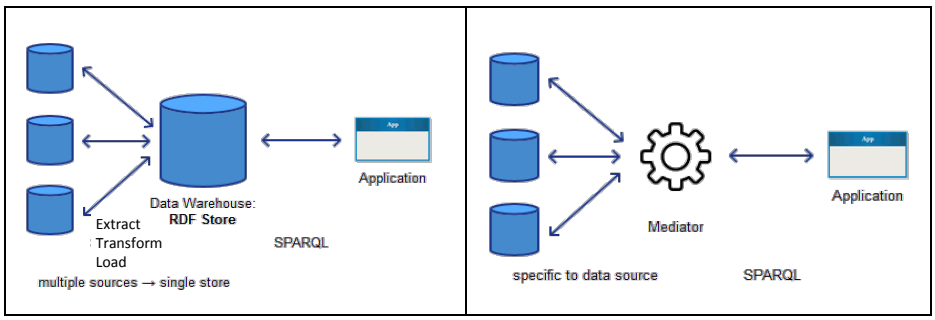


**Fig. 1.** Warehouse versus Mediator.

The second approach, called mediation, allows information to be retrieved dynamically from original databases, at query time. It provides a unified global query-interface and relies on mappings between the global schema and each of the local source schemas. These mappings are used to rewrite the global query into a union of queries that match local schemas. They are directed, either from entities in the global schema to entities in the local sources ("Global As View" (GAV) mappings), or from entities in the local sources to the ones in the global schema ("Local As View" (LAV) mappings). LAV mappings require more sophisticated inferences to resolve a query on the global schema than GAV mappings, but they make it easier to add or retrieve data sources to the mediation framework.

[9] http://www.mpi-inf.mpg.de/yago-naga/yago/
[10] http://hypermedia.research.southwales.ac.uk/kos/star/

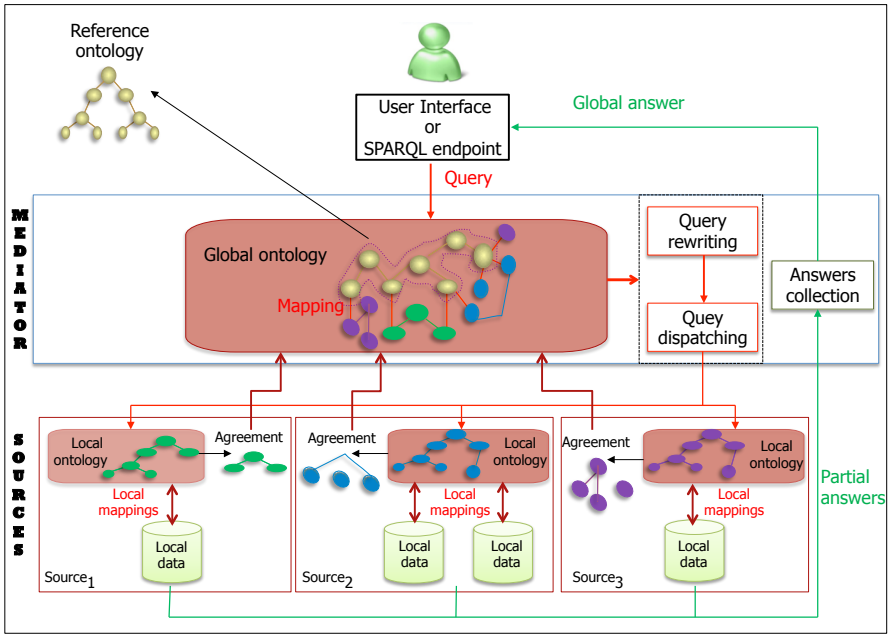


**Fig. 2.** Components of the Semantic Mediation System.

Our proposal, illustrated in the right part of Figure 1, is to automate as far as possible the building of an *ontology-based mediator* in a given application domain. The resulting mediator uses both GAV and LAV mappings, as formally detailed in [6]. Figure 2 gives an overview of our Semantic Mediation system architecture. There are two levels, the mediator in the upper part of the figure, and, in the lower part, the sources that are integrated.

### 2.1 Sources

A source is an institution, a team, etc. who owns structured digitalized data, and who wants to provide a controlled access to (some parts of) this data on the one hand, and, on the other hand, to profit from an access to comparable data owned by other sources. For this second point, the source will simply use the mediator. As a provider of data, the source is represented in Figure 2 by (i) its data and (ii) the access it offers. The access a source provides to the mediation system to which it participates is called "Agreement" in Figure 2.

Starting from its data, which is structured and formatted according to local choices and constraints, the required step for a source is to build a conceptual representation of this data, at least for the part to be accessed via the mediator. This conceptual representation consists of a set of *concepts* (or *entities*, or *classes*) with their *relationships* (also called *properties* in semantic web ontologies). This conceptual

model is expressed as a semantic web lightweight ontology, using OWL2 QL[11]. This language is specially tailored for expressing conceptual models of data, i.e. ontologies with limited semantic constraints but powerful capabilities for acting as *interfaces to access data* stored in non-semantic formats (i.e. not in RDF). It supports the recent notion of Ontology-Based Data Access (OBDA) [7], which is now a very active field of research in database and knowledge engineering domains.

The main principle of OBDA is to allow (i) several relational databases to be integrated, and (ii) their querying to be enriched with the ontological knowledge, while leaving the relational database management system the tasks of efficient storage, maintenance and querying of the data. To the best of our knowledge, important theoretical and practical results have been published essentially for relational database systems, but extensions to other kind of source formats may be considered. A source might also use OBDA principles to build a common interface to *several distinct* databases, when it comes to be possible (cf. $Source_2$ in Figure 2).

A source can either build a complete ontology for representing its whole data, or it can do it only for the part it wants to provide to the mediator. When it has her ontology defined for the whole data, it is frequent that only a part of this data is accurate for the domain of mediation. For instance, in the *Personae* mediator described in Section 3, the database *Bude* is an example of a source that does not provide all its content to the *Personae* mediation process. Nevertheless, other parts of its content may be involved in other mediation processes.

This is why, in Figure 2, we distinguish clearly the local ontology and the agreement. The agreement is the part of the local ontology that provides access to data that is useful for the mediator. For instance, if the mediator is built for enhancing the study of social interactions among people during the Renaissance period, then it might not be useful to integrate all details about the productions of these people (books, letters, music, paintings, etc.), that are described in the different sources. We have presented in [9, 10] algorithms to automate the agreement building process, starting from (i) the local ontology and (ii) a consensual, shared reference ontology, representing the mediation domain. The idea is to align the local ontology with the reference ontology (see [8] for an introduction about ontology alignment). Of course, the process remains human-supervised, to continuously provide to the source the control on what it publishes.

Human supervision is particularly important when using a reference ontology having a wide general scope, such as YAGO2. This one is automatically built from Wikipedia, Wordnet and Geonames, and it proposes an interesting formalization of time and space information [11]. We chose to use it for its multilingual functionality, because our experiments have been done for sources that are all in French. The drawback of such a reference ontology, when used in our system, is to be too general, which increases dramatically the ambiguity due to word polysemy. Nevertheless, when no consensually shared reference ontology exists for the mediation domain, our system is able to use the taxonomy of YAGO2. To this aim, we have developed a module, which is illustrated in Figure 3, to assist the source's owner to choose, between all the existing meanings of the concept label found in YAGO2, the one that best fits to a given local concept.

[11] http://www.w3.org/TR/owl2-profiles/

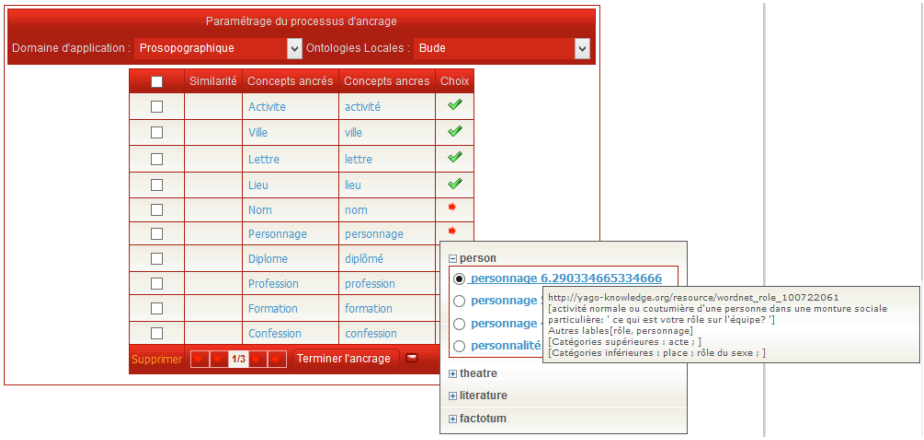


**Fig. 3.** Choice of concepts to keep in the agreement, and choice of the right meaning of the reference concept *personnage*.

### 2.1 Mediator

Once the source has built her agreement for the mediator, all concepts of the agreement are related, directly or indirectly, to a concept of the reference ontology. Then, the agreement is automatically integrated into the global ontology. Our algorithm of integration is detailed in [9, 10], it is the preliminary step for the source be queriable via the mediator. This is illustrated in the general scenario of interaction between a source and the mediator, given in Figure 4. Before interacting with any source, the mediator initializes the global ontology and its repository of source addresses. Then it publishes this repository's address, used by sources to be integrated in the mediation process. Thus, when a source is ready to join the mediation process, it enrolls in the mediator by giving (i) the SPARQL Endpoint of its local OBDA system, in order to be queried from the mediator and, (ii) its agreement and the mappings to the reference ontology issued from the alignment, which are integrated in the global ontology.

The global ontology is a very important part of the mediator. As it is depicted in Figure 2, it can be noticed that it is composed of (i) the agreements from the participating sources and (ii) the reference ontology's concepts that have been associated to the agreement concepts. The key idea of our global-ontology-building algorithm is to borrow from the reference ontology *also the relationships* that exist between these concepts. In this way, once they have been integrated in the global ontology, the local concepts of the agreements acquire new relationships with each other, either because they are associated to the same reference concept or because they are associated to two reference concepts that are related to each other.

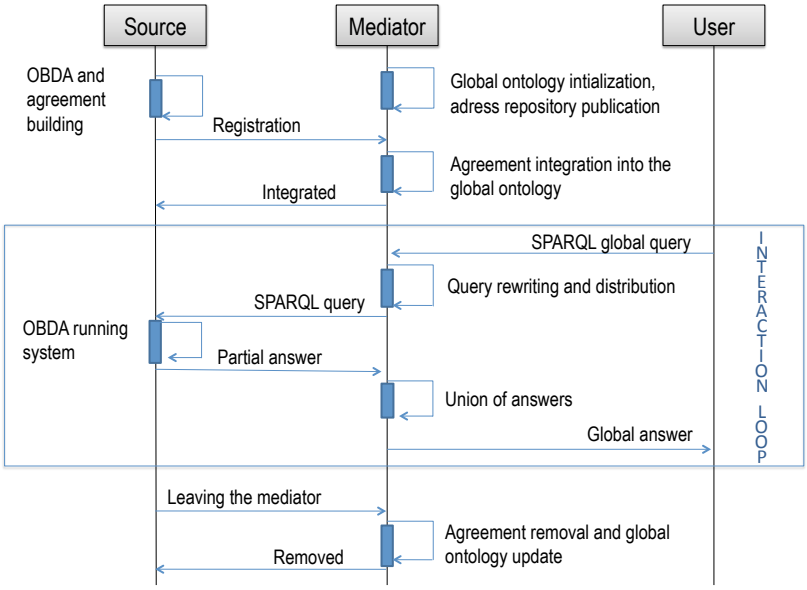


**Fig. 4.** Scenario of interactions.

The other key component of the mediator is the query processing, denoted "interaction loop" in Figure 4. The mediator is queried either by humans through a visual interface or by web services through the SPARQL Endpoint. We call global queries these queries posted on the mediator, which may involve any of the global ontology components (concepts and properties). Our complete algorithm for query processing is described in [6]. As illustrated in Figure 2, it runs in three steps: firstly the global query is rewritten following the principles developed in [7] and the queries involving source concepts are selected. Secondly, the exact queries that must be sent to the sources are built and dispatched to the involved sources. Lastly, all answers from the sources are collected and presented as the global answer.

These generic principles have been implemented for a concrete historical application, the Personae project.

## 3 Application to the Personae Project and beyond

The Personae project, led by the History Institute CESR[12] is an interesting application for our semantic mediation system because the partners want to share their data but each of them has excellent reasons for keeping her own logical data schema. Our proposals are welcomed in this context, because they allow partners to keep their databases independent of the mediator's schema.

[12] Centre d'Etudes Supérieures de la Renaissance http://umr6576.cesr.univ-tours.fr/

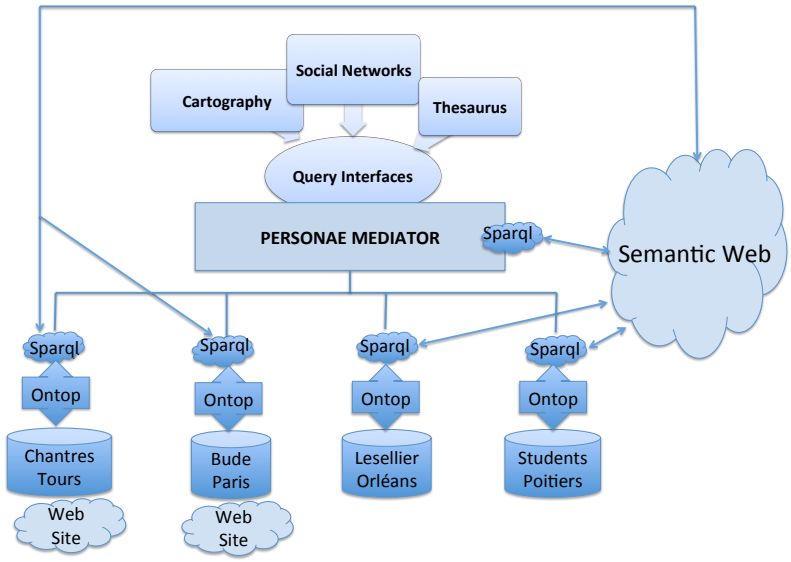


**Fig. 5.** Semantic Mediator for Personae.

### 3.1 Personae

The Personae project associates several groups of Historians who have built databases containing descriptions of people and their relationships during the Middle Ages and the Renaissance, in order to perform prosopographical studies [12]. Starting from natural modes of expression such as local dictionaries, professional dictionaries or thematic dictionaries, the Personae project aims to develop new modes of expression, publication and processing of this kind of data. Personae is focused on the Center West France, which was a high place of knowledge in the late Middle Ages with its three great universities founded in the 14th and 15th centuries (Orleans in 1306, Poitiers in 1431, Bourges in 1464).

Several identified databases form the basis of the project, the ultimate goal being to attract other sources into the planned central portal. One of the existing databases, *Bude*[13], is about manuscripts and printed works from the Middle Ages and the Renaissance, so it covers humanists who wrote, conserved and transmited those texts (about 12000 persons). Another source, the *Chantres* database, comes from the project "Prosopographie des Chantres de la Renaissance"[14], and gathers biographies of professional singers of the 15e and 16e centuries (about 5000 persons). Other sources are involved, for instance the *Lesellier* database, that contains information about French people that were in relation with the papacy, information gathered from registers of letters to the Pope (about 20000 persons). All those sources are structured

[13] http://bude.irht.cnrs.fr/
[14] http://ricercar.cesr.univ-tours.fr/3-programmes/PCR

in relational databases. The integration phase is a cornerstone for the design of new modes of visualization and correlation of prosopographic data.

Figure 5 represents the intended architecture of the Personae semantic mediator. At the top, map-based visualization and navigation tools will use the mediator to exploit the different sources, working with RDF, RDFS and OWL2 QL formats. To this aim, we use some existing useful semantic web tools such as Sesame[15] and Ontop[16], which are also generic enough to provide for the possibility of dealing with other kinds of source format, for instance XML annotated documents. Several dedicated resources are currently built by our historian colleagues, such as a thesaurus of first names (with equivalences in Latin and other European languages of the Renaissance), a dictionary of functions, clerical or military positions, etc. in order to enrich the information retrieval potential offered by the portal. Moreover, it is planned to develop a tool for the declaration and the storage, in RDF, of *coreferences* that historians will find among people described in the different source databases, using the Personae portal.

At the lower level of Figure 5, each local database has an OBDA system installed, i.e., the local administrator has built a conceptual model, expressed as an OWL2 QL ontology and the OBDA mappings between this ontology and the local relational database. In this way, any authorized application can query the source in SPARQL via its ontology. The source can publish this ontology in order to allow any application to access its data, and alternatively, it can give the ontology and grant access only to some identified partners, including the Personae mediator. Irregardless of this access to applications or to the mediator, the source administrator can continue to publish her data as she wants, via a classical web site. It can be noticed that the work sources are required to do on their data (local ontology building and OBDA access) is not harder than OAI PMH[17] requirements, when it opens the source to the semantic web.

### 3.2 Discussion

The prototype developed for the Personae project shows that our ontology-based mediation framework may be used for sharing cultural heritage knowledge. When we began to devise this framework's principles, it was for a domain well described by a consensual reference ontology, AGROVOC, so it was natural to rely on this reference ontology in order to build the source ontological interfaces and the global ontology. When it came to apply these principles to the Personae project, we did not find any reference ontology of prosopographic data. We developed a tool to be able to use a general taxonomy in place of an accurate conceptual description of the domain, but we are convinced that, using a good domain ontology, we could achieve much more interesting agreements, and thus get a global ontology that could be more useful for exploring data sources. For this reason, we are studying the CIDOC-CRM ontology, discovering that the local ontologies, in our framework, are not far from the "surrogate nodes" and the "extracted metadata" of the last architecture introduced in [13]. For a concrete application, we begin to collaborate with Archaeologists, who are using the CIDOC-CRM for publishing their data on the semantic web [14].

[15] http://www.openrdf.org/
[16] http://ontop.inf.unibz.it/
[17] http://www.openarchives.org/pmh/

We are also aware that relational databases are not the only kind of existing data in the cultural heritage field, rather, semi-structured (XML annotated documents) and unstructured (text or images) data are commonly encountered. Another direction of work is to design systems that allow for accessing this kind of data through a lightweight ontology, in the same spirit as OBDA systems for relational databases.

## 4 Conclusion

We presented in this paper the general principles of a generic ontology-based mediation framework that may be usefully applied for Cultural Heritage knowledge sharing and dissemination. We introduced an application for prosopographical studies, and we briefly discussed ways of improving and generalizing our proposal.

References


1. Bizer, C., Heath, T., & Berners-Lee, T.: Linked Data - The Story So Far. In: International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), pp. 1—22 (2009)
2. Gruber T. R.: Toward principles for the design of ontologies used for knowledge sharing. In: Int. J. Hum.-Comput. Stud., 43(5-6), pp. 907--928 (1995)
3. Wache H., Vögele T., Visser U., Stuckenschmidt H., Schuster G., Neumann H., and Hübner S.: Ontology-Based Integration of Information – A Survey of Existing Approaches. In: IJCAI'01 Workshop on Ontologies and Informations Sharing, pp. 108--117 (2001)
4. Calvanese D., Giacomo G., Lembo D., Lenzerini M., Poggi A., Rosati R., and Ruzzi M.: Data Integration through DL-Lite$_A$ Ontologies. In: Schewe K.-D. and Thalheim B., (eds.), Semantics in Data and Knowledge Bases, pp. 26--47. Springer, Heidelberg (2008)
5. Lenzerini M.: Data integration : A theoretical perspective. In PODS, pp. 233--246 (2002)
6. Bouchou B., Niang C.: Semantic Mediator Querying. In: IDEAS'14, Porto, Portugal http://dx.doi.org/10.1145/2628194.2628218 (2014)
7. Poggi A., Lembo D., Calvanese D., De Giacomo G., Lenzerini M., and Rosati R.: Linking Data to Ontologies. In: Journal of Data Semantics, 10, pp. 133--173 (2008)
8. Shvaiko P. and Euzenat J.: Ontology Matching : State of the art and future challenges. In: IEEE Trans. Knowl. Data Eng., 25(1), pp.158--176 (2013)
9. Niang C., Bouchou B., Lo M., and Sam Y.: Automatic Building of an Appropriate Global Ontology. In: ADBIS'11, pp. 429--443 (2011)
10. Niang C., Bouchou B., Lo M., and Sam Y.: A Semi-Automatic approach For Global-Schema Construction in Data Integration Systems. In: IJARS, 4(2), pp. 35—53 (2013)
11. Johannes Hoffart, Fabian Suchanek, Klaus Berberich, and Gerhard Weikum.:YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. In: Artificial Intelligence Journal, Special Issue (2013)
12. L. Stone. Prosopography. In: Daedalus, 100(1), pp. 46—71 (1971)
13. Doerr M., Iorizzo D.: The Dream of a Global Knowledge Network—A New Approach. In: ACM Journal on Computing and Cultural Heritage, 1(1), pp. 5:1--5:23 (2008)
14. Le Goff E., Marlet O., Rodier X., Curet S., Husi P.: Interoperability of the ArSol (Archives du Sol) database based on the CIDOC-CRM ontology. With the collaboration of: Le Bœuf.P. Presented at CAA'14, Paris (2014)