Towards Semantic Integration of Heterogeneous Data Based on the Ontologies Modeling

Cheikh Ould El Mabrouk(✉) and Karim Konaté(✉)

Department of Mathematics and Computing, University Cheikh Anta Diop, Dakar, Senegal
cheikhmab@yahoo.fr, kkonate911@yahoo.fr

Abstract. The integration of the heterogeneous data is a major problem encountered today by the users of the Web. A typical integration scenario is that two heterogeneous systems A and B are built for different business purposes for different users at different times by different software developers using different information models. The two systems often have heterogeneous semantics, data structures and business rules are different.

It involves in particular the differences between systems infrastructures, the conceptual schematizations of the data and its meanings. Indeed, the ontology specifies its systems of knowledge representation. It allows the modeling of knowledge in an explicit and formal way by concepts and relations between these concepts. The semantic integration comes after syntactic integration and the mechanisms of translation connection.

In this paper, we proceeded to a semantic integration of the heterogeneous data based on the management of the heterogeneousness and the semantic ontology of the knowledge.

Keywords: Integration · Ontologies · Semantic Web · Heterogeneousness · Distribution

1 Introduction

The integration process of heterogeneous information sources in the ontologies context of the semantic Web rely on several approaches. For that purpose, the last years have seen the several researches works realization in the Web databases domains and the ontologies concerning the integration design realization of different databases in open environments such as Semantic Web.

This new Web generation is an important evolution compared to the other previous generations will be based on ontologies and semantic knowledge management. Furthermore, the ontologies are used in several areas of information processing. In addition to, the semantic Web infrastructure has to allow their integration giving the impression to the user that it uses a homogeneous system. However, the typologies information distribution systems are generally based on software and sources of heterogeneous data on several computers. Often, within a company, there are several heterogeneous systems that are not designed to integrate data or applications. Then, the development and

É. Renault et al. (Eds.): MSPN 2019, LNCS 11557, pp. 188–200, 2019.
https://doi.org/10.1007/978-3-030-22885-9_17

deployment tools evolve over time. This raises integration problems, because infrastructures information systems do not easily exchange computerized data. In particular, the semantic integration problems become then ontological integration problems [1, 2].

A machine and human readable language is therefore preferred to model both the semantics of the systems and the heterogeneity between them. Data Semantics models can be expressed in various forms ranging from schemas to system documentation. Some well-known information modeling languages are evaluated below for the criteria of machine and human readability. Gathered data often comes from heterogeneous (different) sources; therefore, integration activities are needed and very important. In a business context, integration activities are commonly referred to as Enterprise Integration. This means the ability to integrate information and functionalities of different Information Technology systems.

In this paper, we divide our work as follows: the first part is this introduction. The second part constituted by the problem of data integration. The third part makes a state of the art on the heterogeneousness of the information system. The fourth part formulated the set of heterogeneousness management approaches. The fifth part of this article focuses mainly on a conception of the ontology models and the data integration. Finally, the sixth part summarizes all of our works and gives some perspectives for their continuation.

2 Problem of Data Integration

The systems of the information (SI) consist of Hardware and Software such as: operating systems, communication protocols, local area network, network links, DATABASE MANAGEMENT SYSTEM (DBMS), programs and software packages. Data integration addresses problems related to the provision of interoperability to information systems by the resolution of heterogeneity between systems on the data level. The heterogeneousness of the information system is an inevitable problem in the fact that the data and the applications of SI can be developed and deployed in independent ways and according to approaches and different methodologies of design and realization [3].

Data integration is an area of research that addresses a pervasive challenge faced in applications that need to query across multiple autonomous and heterogeneous data sources.

In this context, the systems of integration have to allow the user to access, via a single access interface, data stored in several sources of data. These sources have been independently designed by different designers. It entails the heterogeneousness of data, that is to say, that the data relating to the same meaning are differently represented on different information systems [4]. This heterogeneousness systems from different choices which are made to represent facts of the real world in a scheme of design and development.

3 State of the Art on the Heterogeneousness of the Information System

Heterogeneous (means from Greek "other" or heteros and geneous or "nature") is the characteristic to contain dissimilar constituents. A common use of this word in terms of information technology is to describe a product as a measure to contain or to be a part of a «heterogeneous hardware/ software», made up of various manufacturers products that can (interoperate) [5].

In this context, the heterogeneousness of data concerns at the same time the physical system, the syntactic and semantic structure. This informational heterogeneousness results from the fact that the sources of data may have different structures and/or different formats to represent their data. And the sources of data are independently designed, by different designers, having different application objectives. Everyone can, thus, have a different point of view on the same concept and the object [2]. As a result, it exists several types of heterogeneousness are due to the technical differences of management of SI such as the differences between the physical hardware, the software systems. The authors' work [3, 6, 7, 27, 28] can distinguish mainly three types of heterogeneousness as below:

- Technical heterogeneousness: refers to the difference between physical hardware, network infrastructure, cables, operating systems and application platforms. It refers to the resolution of structural heterogeneity; for instance, the heterogeneity of data models, query and data access languages, protocols, and hardware platforms.
- Syntactic heterogeneousness: corresponds to the different presentations in the data formats and interfaces of the applications. It refers to the resolution of semantic mismatch between schemata. A mismatch of concepts appearing in such schemata may be due to a number of reasons. For instance, different schemas may represent the same information in different ways. The major issues that make integrating data difficult include. The similar semantics of data representation might be quite different in each data source. Moreover, they may contain conflicting data. In addition, heterogeneity may also occur at lower levels, including access methods, underlying operating systems, etc.
- Semantic heterogeneousness: corresponds to the differences related to the interpretation or the explanation in the sense associated with the data and functions of an application. Data sources are independent elements that are not designed for a data integration system. They cannot be forced to act in certain ways. As a natural consequence of this, they can also change their data or functionality unannounced.

4 Approaches to Heterogeneousness Management

The web services development today has allowed the putting on line of an imaginable number of heterogeneous and distributed information (data, files, video, images, sound …). Each type of information offers autonomous access interfaces of other types and often heterogeneous between different web technologies (HTML, PHP, JAVA, Web Services ….), and by the communication infrastructure of the information system (Systems, Networks, DBMS (DATABASE MANAGEMENT SYSTEM))

heterogeneous [8]. Therefore, the integration and the exchange of the heterogeneous data allow having a logical integration at the level of the access to the data.

In this case, the logical integration of the data takes place at the level access to the data. At the global level, the applications have a uniform view of the data physically distributed, through a representation of the data. At the local level, information systems keep their autonomy of the data representation (identification, type, length …), and to allow access the data via other applications [2, 9]. These approaches aim to solve the above-mentioned heterogeneous problems and to move to the limits semantic integration ontologies.

4.1 Management of the Technical Heterogeneousness

The management of the technical heterogeneousness allows opening a dynamic management of the tasks and the human-machine interface. The evolution of the material and the IT software have led to new needs in terms of adapted networks and highly distributed architecture (offering to all the possibility of enriching the company's information system) [10].

In this context, software allows in this case to launch several tasks invited on host machines and be placed next to several completely isolated operating systems. Then, facilitate communications between the protocols of operating systems and networks.

For that purpose, all the potential machines are provided with linking network cards which are in fact processors producing the option to exchange information with the outside in order to establish easily effective interactions between the systems. As a result, several technical solutions are used for the operation of the heterogeneous hardware and software. The HAL stands for "Hardware Abstraction Layer", or hardware (or software) abstraction layer, whose function is to isolate the specificities of the hardware through a number of hardware-specific functions: [11, 12].

(1) - Partitioning, isolation of the physical and/ or software resources.
(2) - The ability to manipulate remote machines by transcribing data, pausing, stopping and starting, remote programming capabilities…
(3) - Possibility of "logical" networking of remote machines but also interfacing them with physical networks.
(4) - The high availability for backup security: outsourcing and third-party centralization can be applied to any SI.

4.2 Management of the Syntactic Heterogeneousness

The syntactic heterogeneousness is when data of interest is available in various formats, different representation schema or a query translation [13]. Thus, the interaction between several information systems requires efficient management of the exchange of computerized data between the heterogeneous applications of these latters.

The difficulty mainly arises from the data representations incompatibility. As for example, «id-student of alphabetical type and length 10A» and «code-student of digital type and length 6 N» are two different representations from the same meaning identification of a student.

The authors' work [6] proposed mechanisms of translation between the diagrams of representations of data (designation, type, length…).

To resolve this problem of syntactic heterogeneity, translation mechanisms between data representations must be implemented. We have proceeded to prove by recurrence the two approaches presented by [6] (in position by type of distribution architectures point-to-point, EAI (Enterprise Application Integration) and ESB (Enterprise Service Bus)).

We distinguish two approaches according to the architecture of distribution (Fig. 1):

A) - **A point-to-point architecture for data translation approach**

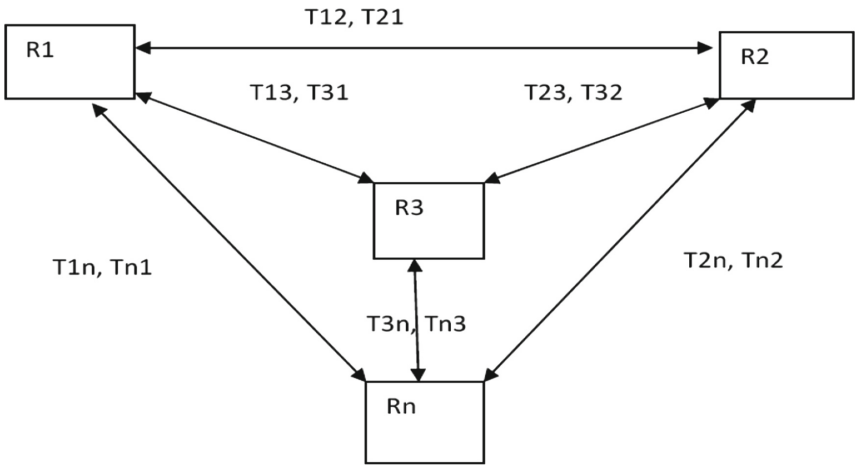

Fig. 1. Data translation by point-to-point architecture

In this approach to data representation of point-to-point architecture, the translations are directly established between two representations. To link two representations (R_i, R_j), two translations T_{ij} (from representation ***i*** to representation ***j***) and T_{ji} (from representation ***j*** to representation ***i***) carry out the syntactic correspondences.

As a result, each representation (R_i) there are translations as follows:

$$T_{ij},\ j\ = 1, 2, .., \ i-1,\ i+1, .., \ n$$

$$T_{ji},\ j\ = 1, 2, .., \ i-1,\ i+1, .., \ n$$

Therefore; for each (R_i) in (**n**) representations of the data: (n−1) + (n−1) = 2 * (n−1).

The principle of this approach: for (**n**) representations there exists n * (n−1) translation mechanisms.

We suppose that valid for (n-1) representations and we prove by the recursion for (n) representations.

For (n−1), the representation number is (n−1) * (n−2)

For "n^{th}" the representation number is 2 * (n−1)
So, the number of representation this sum of:
(n−1) * (n−2) + 2 * (n−1)
As a result, to represent (n) there must exist [n * (n−1)] translation mechanisms.

b) An approach of data translation by architectures EAI, ESB

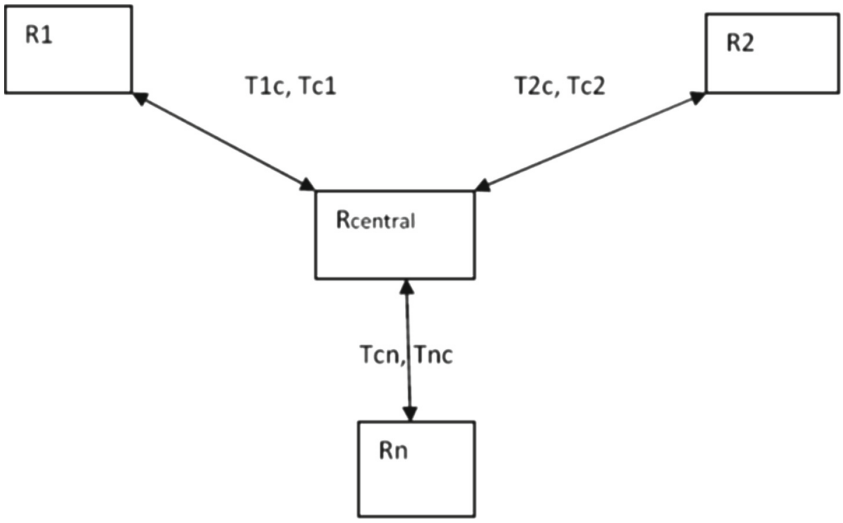

Fig. 2. Data translation by central mediation EAI, ESB

In this approach of the data representation of the EAI (Enterprise Application Integration), ESB (Enterprise Service Bus) architectures they are necessary to translate the entire source representation to a central representation and then translate this central representation to the target representation scheme (Fig. 2).

Consequently, for each representation scheme (R_i) there are two translation mechanisms from and to the central representation scheme ($R_{central}$): T_{ic}, T_{ci}.

The principle of this approach is that for (n) diagrams of representation there must be (2 * n) translation mechanisms.

We suppose that valid for (n-1) representations and we prove by the recursion for (n) representations.

For (n−1), the representation number is 2 * (n−1)
For n «n^{th}» the number of representation is 2
So, the number of representation this sum of:
2 * (n−1) +2 = 2 * (n−1 + 1) = 2 * n
Finally, to represent n there must exist [2 * n] translation mechanisms.

4.3 Management of the Semantic Heterogeneousness

The semantic heterogeneousness means resources of the data highly varied and more or less structured information sources (databases, XML documents, texts); whereas, in homogeneous sources from the viewpoint of their level of constitution but nevertheless coherent. The semantic heterogeneousness describes the difference in the meaning of the data between the various sources of the data [6, 14] (Fig. 3).

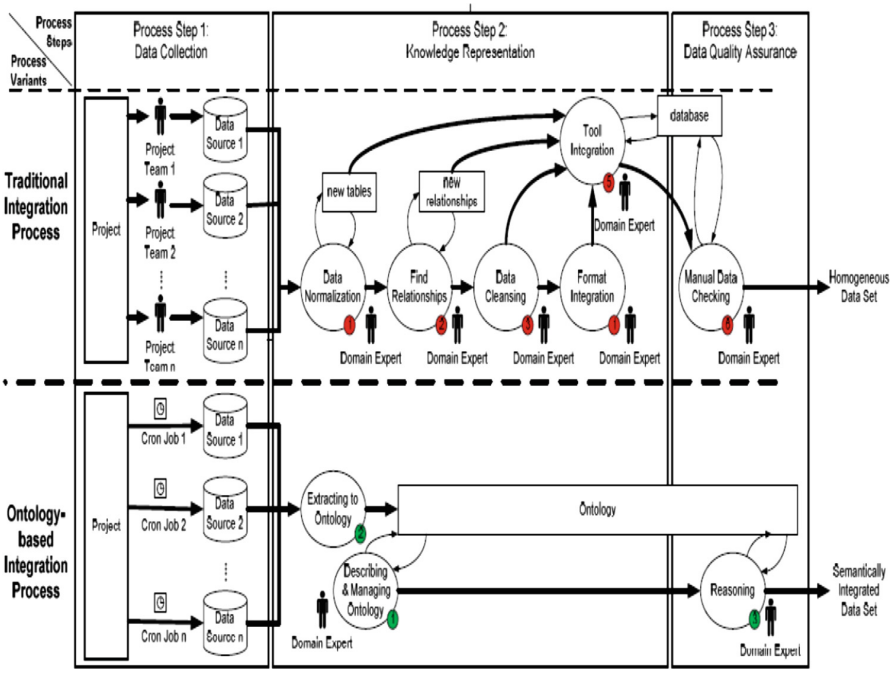

Fig. 3. Comparison of traditional and ontology integration processes [27].

The Extraction of integration rules is based on the heterogeneity analysis of the different systems. Therefore the quality of the heterogeneity description decides directly the quality of integration rules. In order to describe the heterogeneity, semantics of the source systems need to be described first. Machine readable semantic models are preferred compared to only human understandable semantic models because machine readability provides the possibility of utilizing automated reasoning and thereby the possibility of automation of the integration rules extraction.

In the ontology-based approach, the knowledge representation is done by designing and implementing an initial ontology that captures all data models and requirements in advance. The designer defines the classes, attributes and relations of ontology that will be populated automatically by using tool during the data collection step. In the ontology, the designer should also define restrictions, rules and axioms that are used for data quality assurance to check the syntax and the constraint of the data [26].

In this context, Semantic web is an integrative vision of many problems of heterogeneous data and ontological knowledge and of indexing and the communication of information. However, ontologies as information server interrogation interfaces and a tool for combining data from heterogeneous data sources and where core semantic Web applications and as well as rules and axioms that cover a rich data management.

According to the authors [15–17, 23, 25]; they are based on the three approaches:

- The acquisition: identify all the different sources of the knowledge and the concepts and its logical relations in a specific field.
- The modeling: to indicate the set of data or knowledge that fixes the linguistic meaning of the concepts in a specific domain.
- The representation: define the set of concepts linked by specific relations.

Finally, heterogeneity managing approaches are limited in terms of modeling ontological for each domain of knowledge.

5 Ontology Models and Data Integration

Ontology models are used to facilitate the integration and the exchange of heterogeneous data. In association with the syntactic translation and the semantic management almost one-to-one object, class, type entity and the converter to the attribute, property, type value and contribute to the accurate modeling of the properties, and classes for the semantic integration of information [18, 22].

In this context, Ontologies have been extensively used in data integration systems because they provide an explicit and machine-understandable conceptualization of a domain.

5.1 Modeling Ontological Knowledge

Knowledge modeling involves depositing the entities of different identical concepts, retaining the concepts and the relationships, in determining their domain, and indicating the set of data or knowledge that sets the linguistic meaning of the concepts for each type of semantic ontology and logical link.

The works of the authors [19, 20, 24] proposed ontologies ranking approaches before semantic integration:

- The regional ontology is seen as a tree of semantic concepts and its relationships. Its interpretation is constrained by the differential principles associated with the elements constituted of the tree: root, ancestor, branch, leaf…
- The referential ontology describes a set of referential (or formal) concepts that are characterized by a term/ designation whose semantics is defined by an extension of objects.
- The computational ontology deals with computational concepts that are characterized by the operations that can be applied to them to generate inferences. The global ontology provides a conceptual view over the schematically-heterogeneous source schemas.

5.2 Semantic Integration

The semantic integration focuses on the intended meaning of the concepts, and to establish the semantic relationships between the concepts of the modeled ontologies. This process of semantic integration based on two ontological approaches: [2, 21].

- A posteriori: to perform in a manual or semi-automatic manner and to establish the correspondence between the basic concepts of the ontologies.

  For [n] ontologies one has to create [n * (n−1)] correspondence.

- A priori: allows to automatically integrating each ontology source [Oi] of the semantic relations as a subset of the global ontology.

  For each [Oi] of the global ontology of [n] elements one has to create [(n−1)] correspondence.

5.3 A Semantic Integration Procedure

In this procedure we have proceeded to an approach of our semantic integration by: Proposed algorithm, schema simulation of syntactic and semantic representations, table of semantic integration results of heterogeneous data, and discussion.

(A) -Algorithm "Sematic Integration"

BEGIN

- management of technical heterogeneity
  Web Services "connection of heterogeneous and distributed systems"
- management of syntactic heterogeneousness
  IF Architecture Type = Point-to-Point THEN
  Translation number = n * (n-1)
  ELSE Translation number = 2 * n
  END IF
- management of semantic heterogeneousness
  Creating, Modeling and Representing Semantic Data
- Semantic ontology modeling
  Classification of ontologies by type
  Definition of ontological relationships

BEGIN
IF type of information Automatic THEN number of matches = 2 * (n-1)
ELSE number of matches = n * (n-1)
END IF
END
END

(B) - Algorithm Simulation
Table 1

Table 1. Schema of syntactic and semantic representations

Number	Syntactic heterogeneousness		Semantic heterogeneousness	
	Point-to-point	Mediation	Manual or semi- automatic	Automatic
1	0	2	0	0
2	2	4	2	1
3	6	6	6	2
4	12	8	12	3
..				
N	N*(N-1)	2*N	N*(N-1)	(N-1)

(C) - Result

- Manual semantic integration by point-to-point representation MP = (N * (N−1)) ^{2}
- Manual semantic integration by representation Mediation MM = 2 * (N−1) * N2
- Automatic semantics integration by point-to-point representation AP = N * (N−1) ^{2}
- Automatic semantics integration by representation Mediation AM = 2 * N * (N−1) (Table 2).

Table 2. Result of heterogeneous semantic integration data

N	MP	MM	AP	AM
1	0	0	0	0
2	4	8	2	4
3	36	36	12	12
4	144	96	36	24
5	400	200	80	40
6	900	360	150	60
7	1764	588	252	84
8	3136	896	392	112
9	5184	1296	576	144
10	8100	1800	810	180

(D) – Discussion
The solution we propose offers a heterogeneity management algorithm that resembles all the necessary processes for the syntactic and semantic integration of heterogeneous information. In our algorithm, we have combined the two approaches: [6] and [21], integrating the identification and description of data that existing web services

capabilities into models for integration direct and effective interesting parts that implement the web service of the information environment.

The process begins with the management of technical heterogeneity, which is considered as a preparatory step of the integration phase, where we will decide whether it is possible or even recurring to continue composing the information environment towards an integrated environment. This step is very important to manage the heterogeneity of an integrated data environment, hence the clear and deep understanding of the translation mechanism and the management of syntactic and semantic heterogeneity.

In the same run, we calculate the two last phases of translation and the representation of the data. From these, we distinguish the types of architectures that exist in the system and that must be exchanged the translation mechanism to other representations. The integration of information environments is realized by the integration of their applications via web services interfaces. The final step is the integration of web data into a central control flow of the information system.

6 Conclusion

The integration of heterogeneous data is an important problem in the current and future times. To this end, the ontology of the semantic web has been proposed to solve this problem. An important evaluation of the web generations shows that it produces a semantic integration of heterogeneous data.

In this paper, we have shown the semantic integration process based on the management of heterogeneity and semantic ontology of knowledge and we discussed the main issues and the proposal solutions.

This proposal presents a limited syntactic integration approach towards semantic integration, calculations of the different models translation mechanism and the information representation.

Our future works concerns the semantic evolution problems towards knowledge integration. The advantages and disadvantages for each proposed approach, the queries optimization approaches of the heterogeneous web data, and to explore other ontologies models to develop the semantic integration process towards global integration.

References

1. Hasan, M.K., et al.: A community-driven approach for missing background knowledge in Semantic matching. Int. J. Eng. Sci. Technol. **2**(10), 5921–5928 (2010)
2. Nguyen Xuan, D.: Intégration de bases de données hétérogènes par articulation apriori d'ontologies application aux catalogues de composants industriels, thèse (2006)
3. Sneed, H.M.: Integrating legacy software into a service oriented architecture, discussion paper. In: Lehner, F., Nösekabel, H., Kleinschmidt, P. (eds.) Multikonferenz Wirtschaftsinformatik 2006, Band 2, XML4BPM Track, GITO-Verlag Berlin, pp. 345–360 (2006)
4. Bellatreche, L., et al.: An a priori approach for automatic integration of heterogeneous and autonomous databases. In: International Conference on Database and Expert Systems Applications (DEXA 2004), pp. 475–485, September 2004

5. What is heterogeneous? Conférence CIO-MIDMARK 20000. http://searchcio-midmarket.techtarget.com/definition/heterogeneous. Accessed 31 Jan 2019
6. Touzi, A.C.J.: Aide à la conception de Système d'Information Collaboratif support de l'interopérabilité des entreprises, thèse Doctorat Ecole des Mines (2007)
7. Wiederhold, G.: Mediators in the architecture of future information systems. IEEE Comput. **25**(3), 38–49 (1992)
8. Hacid, M.-S., et al.: L'intégration de sources de données. http://leo.saclay.inria.fr//publifiles/gemo/GemoReport-416.pdf. Accessed 30 Jan 2019
9. David, R., et al.: Virtual integration for improved system design. In: The First Analytic Virtual Integration of Cyber-Physical Systems Workshop, pp. 57–64, Avicps (2010)
10. Zhao, X., et al.: Web services in distributed information systems: availability, performance and composition. Int. J. Distrib. Syst. Technol. **1**(1), 1–16 (2010)
11. Remus, D.B.: Transparent high availability for database systems. In: Minhas, U.F. et al. (eds.) Proceedings of the VLDB Endowment the 37th International Conference on Very Large Data Bases, Seattle, Washington, 29th August 3rd September 2011, vol. 4, no. 11 (2011)
12. La Virtualisation: machine virtuelle ou hyperviseur New Technologies System Virtualisation. http://www.ntsysv.com/index.php/la-virtualisation-machine-virtuelle-ou-hyperviseur. Accessed 17 Jan 2019
13. Abrouk, L., DiJorio, L., Fiot, C., Hérin, D., Teisseire, M.: «Enrichissement d'ontologie basé sur les motifs séquentiels». 23èmes Journées Bases de Données Avancées, BDA 2007, Marseille, 23–26 October 2007
14. Laublet, P., Reynaud, C.: Ontologies et Gestion de l'Hétérogénéité Sémantique conférence GDR, INRI, 3 juillet, Grenoble, France, vol. 13, p. 5 (2007)
15. Ouksel, A.M., Jurca, O., Podnar, I., Aberer, K.: Efficient probabilistic subsumption checking for content-based publish/subscribe systems. In: Middleware, pp. 121–140 (2006)
16. Nagiba, A.M., et al.: AlSIGHTED: a framework for semantic integration of heterogeneous sensor data on the Internet of things. Proc. Comput. Sci. 83 529–536 (2016). The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)
17. Bachimont, B.: Engagement Sémantique et Engagement Ontologique: Conception et Réalisation D'ontologies En Ingénierie Des Connaissances, chap. 19, pp. 305–324. Eyrolles (2000)
18. Gruber, T.: A translation approach to portable ontology specifications Knowl. Acquisition **5** (2), 199–220 (1993)
19. Chandrasekaran, B., Josephson, J.R., Benjamins, V.R.: What are ontologies and why do we need them? IEEE Intell. Syst. **14**(1), 20–26 (1999)
20. Brisson, L.: Mesures d'intérêt subjectif et représentation des connaissances. Rapport technique, Laboratoire I3S, Université Sophia Antipolis, Nice France, Octobre 2004. http://www.i3s.unice.fr/mh/RR/2004/RR-04.35-L.BRISSON.pdf. Accessed 10 Jan 2019
21. Mellal, N.: Réalisation de l'interopérabilité sémantique des systèmes, basée sur les ontologies et les flux d'information, thèse (2007)
22. Wang, J., et al.: Integrating heterogeneous data source using ontology. J. Softw. **4**(8), 843–850 (2009)
23. Noy, N.F., et al.: Semantic integration: a survey of ontology-based approaches. ACM SIGMOD Rec. **33**(4), 65–70 (2004)
24. Shi, L., et al.: SBVR as a semantic hub for integration of heterogeneous systems - a case study and experience report -. Statsbygg, Pb. 8106 Dep, 0032 Oslo, Norway. http://ceur-ws.org/Vol-1004/paper10.pdf
25. Cruz, I.F., Xiao, H.: ADVIS, the role of ontologies in data integration, Lab Department of Computer Science University of Illinois at Chicago, USA. www.cs.uic.edu/∼advis/publications/dataint/eis05j.pdf

26. Olaru, M.O.: Heterogeneous data warehouse analysis and dimensional integration Ph.D. dissertation, International Doctorate School in Information and Communication Technologies XXVI Cycle. www.dbgroup.unimo.it/tesi/Tesi_Phd/phdOlaru.pdf
27. Biffl, S., et al.: Semantic integration of heterogeneous data sources for monitoring frequent-release software projects. In: 2010 International Conference on Complex, Intelligent and Software Intensive Systems, pp. 360–367 in p. 365. IEEE Computer Society (2010)
28. Macura, M., et al.: Integration of data from heterogeneous sources using ETL technology. Comput. Sci. **15**(2), 109–132 (2014)