A SURVEY ON AN ONTOLOGY DEVELOPMENT ON DATA INTEGRATION

[1]N.Mahesh Raj, [2]Dr. R. Jegadeesan and [3]Satya Teja

[1,2]Assistant Professor-CSE, [2]Associate Professor-CSE

[1,2,3]Jyothishmathi Institute of Technology and Science, Karimnagar, India

Abstract

Implementation of data integration in the current days still has many issues to be solved. Heterogeneity of data with non-standardization data, data conflicts between various data sources, data with different representation and as well as semantic aspects problems are among challenging research areas. Semantic data integration using ontology approach is considered as an appropriate solution to deal with semantic aspects problem in data integration. However, most methodologies for ontology development are developed to cover specific purpose and thus not suitable for common data integration implementation. This research offers an improved methods for ontology development on data integration to deal with semantic aspects problem. There are three main parts in this research, the first part is to review, compare and critically analyse the existing methodologies for ontology development. The second part is to create custom ontology development phases for specific purpose in the data integration implementation. The third part is to implement and evaluate OntoDI. This research is also a continuation and improvement of the previous work about ontology development methods on agent system. Furthermore, the ultimate goal of this research is customization, improvement and simplification of the existing ontology development phases for specific purpose on the data integration implementation.

Keywords: Data integration, Methods, Ontology development, Semantic issues, Semantic approach.

1.Introduction

The implementation of data integration still leaves many problems to be solved.Sharing and integrating data from loosely coupled, heterogeneity of data representation and mapping data on different data source are among serious problemson data integration [1-6].Moreover, a big data that most likely includes the heterogeneity of data producesdata conflicts issues especially on semantic aspects between different data representation and sources [7, 8]. This phenomenon to be more common and to be the main challenges in the data integration implementation in the last few years [7, 9-15].

Semantic aspects problem is related to the meaning of every word between terms in a special context or system [7, 16]. There are two possibilities of data problem on semantic aspects [17]. The first problem is about data that have different names with the same meaning. For example, between two data sources with different applications in education domain, they store data about students. In the one data source, student’s data saved by pupil name and in the other data source student’s data stored by learner name. This condition produces semantic data conflict between learner and pupil, because in these two data sources are store the same data about student information. The second possible problem on semantic aspect is about data that has the same name with different meaning. For example, inside education domain between two data sources with different applications, they store about students (undergraduate and postgraduate students) data. In the one data source, undergraduate data saved by student name and in the other data source postgraduate data stored by student name also. Semantic technology is the solution for this problem using ontology approach to make semantics relationship between these two semantic aspects.

The methodologies for ontology development have been growing up in recent years. Every ontology development methods that has been proposed is based on specific goal and domain area to implement the ontology result [18-20]. In this research also discussed about review and comparison activity to analyse the existing ontology development methodologies. It is expected to obtaina brief summary of existing ontology development methodologies.

The aim of this research is to produce an improved methods phases for ontology development specific on data integration domain area (OntoDI). The development of OntoDI is based on review, comparison and analysis activity in the section two and an improvement of ontology development methods from our previous work. The ultimate goal of the development OntoDI is the customization, improvement and simplification of the existing ontology development phases for specific purpose on the data integration implementation.

2.Existing methodologies for ontology development

Several methodologies for ontology development have been developed since late eighties [21-25]. The first objective in this research is to review, compare and analyse existing methodologies for ontology development based on five criteria’s. The firstcriteria is the name of methods and the year when the methodologies are developed. Thesecond criteria is the name of the developerthat creates methods. Thethird criteria is the purpose of the methods development. The fourth criteria is about methodscategories. Thefinal criteria is about methods steps.

There are three categories of development methods, the first category is the methodologies that consider about collaborative and distributed construction (CoDi), the second category is the methodologies that do not consider about collaboration and distributed construction (NoCoDi) and the third category is the methodologies that can be reengineered (Reeng) [19]. Based on Badr et al, there are four methodologies in the first category, seven methodologies in the second category and no methods has been developed in the third category [19]. However, this research is to update the number of methodologies with the latest methodologies, improve some methodologies categorization, and compare with more detail and with different perspective of the existing methodologies.

Table 1 shows sixteen existing methodologies for ontology development that has been reviewed based on five categories.We list sixteen existing methodologies from the oldest until the latestontology development methods. In the table 1 also compare sixteen methodologies based on five criteria's. There are two criteria's to be main concern on the comparison table, they are category and methods steps criteria.

The first concern is about methodologies category. Our contribution in this section is that we do improvement about methodologies criteria for NeOn and ENTERPRISE methodologies. In the previous research [19], Badr et al grouping the ENTERPRISE methods into NoCoDi criteria and NeOn into CoDi criteria. However, based on literature review and analysis process in this research, we found that inside ENTERPRISE development steps there is integration process that's mean this process consider about collaborative and distributed construction, so we put ENTERPRISE methods into NoCoDi and CoDi category. Whereas, in the NeOn methods we found that inside NeOn development steps there is reusing and reengineering ontological resources process, this mean that this methods also enter into reengineering methodologies category.

Another contribution of this section is about improvement the number of methodologies review from eleven methodologies to be sixteen methodologies, by including the newest methodologies method. There are three methodological review that were lost in the previous research [19], they are Unified, ontology integration and semi-automatic creation ontologies methods. Furthermore, this research alsosupplement two methodologies as a newest methodologies in the recent years, they are CoMOn and OmMAS methods.

A review and analysis of literature that show in the table 1, there are no methodologies for ontology development specific for data integration implementation. This research is also a continuation and improvement of the previous work about ontology development methods on agent system [18].

The ultimate goal of this research is to create a new methods for ontology development for specific purpose in the data integration implementation. This research is also to customize, improve and simplify the existing ontology development phases to make better ontology development in the future.

3. Methods For Ontology Development On Data Integration (ONTODI)

In the current days there are many problems in the implementation of data integration related to the semantic aspects problem [9-12, 14, 15, 35, 36]. Ontology knowledge is become the one of the solution to solve semantic aspects problem in the data integration implementation. Since 1989 researchers has been developed methodologies to develop ontology knowledge [17-30, 32-34, 37-45]. However, are the existing methodologies for ontology development suitable for data integration implementation? This research is to answer that question to develop new methods for ontology development that specific for data integration implementation.

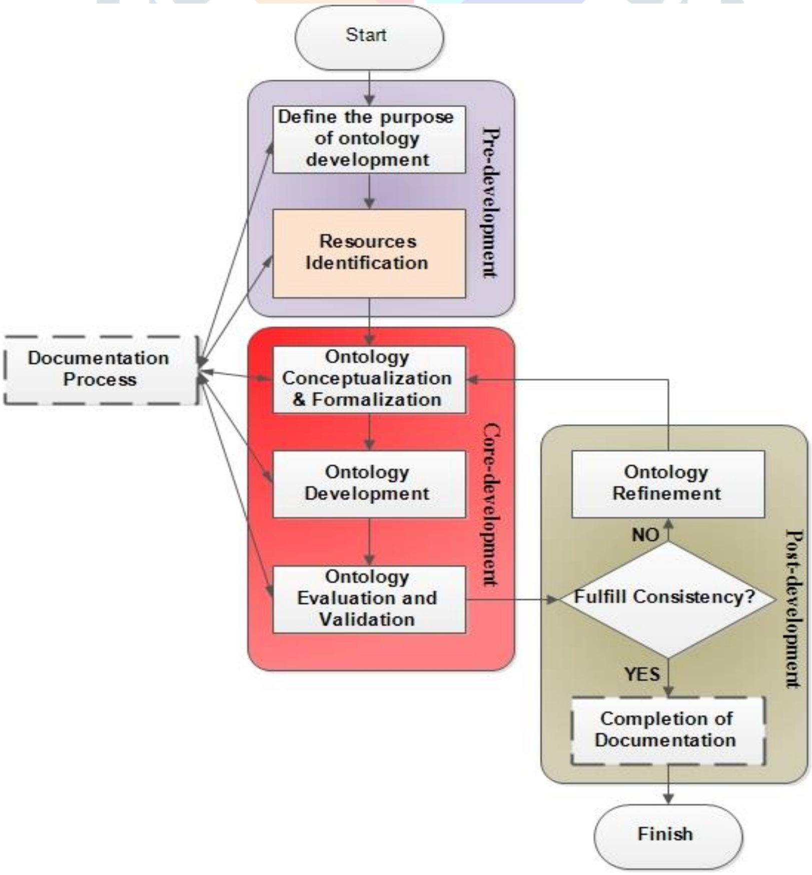

Fig. 1. Methods for Ontology Development on Data Integration (ONTODI).

There are a lot of diversity of steps to develop ontology in the specific implementation domain. This aspect is related with the specific goals to produce ontology. However, in several research has been discussed the common steps of the ontology development methods [17, 19, 20]. The most common steps in the ontology development are specification, conceptualization, formalization, implementation, evaluation and documentation.

The specification process is related to the purpose of the ontology development, why this ontology need to develop and what is the domain of this ontology development. Before we go to conceptualization process, we need to add more process related with the data integration domain. Identification of the data resources need to be done before we go to conceptualization process. After finishing the process of identification data resources, then the next process named conceptualization is formed. The conceptualization process is related to the organization and structuring domain knowledge. The process after is transforming the conceptual model into formal model in the formalization process. The core of the ontology development are on the implementation and evaluation process. These two processes are related and have looping iteration to refinement the ontology. The last process is the documentation process to finishing documentation file about ontology development process.

Figure 1 shows detail about methods for ontology development on data integration domain (OntoDI). OntoDI has three main parts for the ontology development process, they are pre-development, core-development and post-development part. Every part has detail activities, the pre-development has defined the purpose of ontology development and resources identification process. In the core-development has three processes, they are ontology conceptualization & formalization, ontology development and ontology evaluation and validation. Post-development part is to handle ontology refinement and completion of documentation. In the OntoDI, documentation process is related with all processes in the OntoDI phases, because we argue that since we initiate the first process, it require to create documentation until the ontology development process finish.

4. An Experiment and Discussion of Ontology Development On Data Integration(OntoDI)

This section describe implementation of OntoDI in the specific domain for data integration and discuss in detail about OntoDI. The main purpose of the OntoDI is to develop the ontology knowledge to handle semantic aspects problem to support the implementation of data integration. There are two parts in this section, the first part is the implementation of OntoDI to develop ontology knowledge that will be used in the data integration implementation. In the second part is to deliberate OntoDI thoroughly the implementation result to analyse the performance of OntoDI.

4.1. Implementation of OntoDI in the education domain

The OntoDI has three main parts of the ontology development and every part contains several phases on ontology development. The first is pre-development part, in this part contains two phases in order to define the purpose of ontology development and to identify the data on every data source. The second part is core development of ontology knowledge. In this part contains three phases, they are conceptualization and formalization of the ontology knowledge, development of ontology knowledge using specific tools and evaluation and validation of ontology knowledge. The third part related to the post development that contains two activities named ontology refinement and completion of documentation. The documentation process in the OntoDI is start from the beginning phase and has relationship with all phases in the OntoDI.

4.1.1. Define the purposes of ontology development

The experiment in this research is correlated to data integration implementation in the electronic learning system domain. Therefore, the purpose of the ontology development in this research is to produce learning knowledge to share and integrate different learning information between different systems. This ontology knowledge also believed capable on solving semantic aspects problem that happen on the sharing and integration process in the education area.

4.1.2. Resources identification

The second phase on OntoDI is to identify and select the specific data that require an integration. There are many sources in different systems on education domain. However, this research only focus on Student Evaluation System (SES) and Grading System (GS).There are four data are selected from SES, there are student, student2, questions and mark. Furthermore, in the GS there are three selected data,they are student, student_undergraduate and grade.

From these two sources there are semantic aspect problems happen between them. The first semantic problem is between mark and grade. Among these two data, it contains same data regarding student mark but it stored with different name. That's mean this condition has semantic aspect of different name with the same meaning. The second semantic problem is between student in SES and student in GS. These two data sources has the same name but contains different student information, in the student on SGS is contains data about student undergraduate information and in the student on GS is contains data about postgraduate student information. This is the semantic problem related to same name but has different meaning.

4.1.3. Ontology conceptualization and formalization

Conceptualization phase is the process to generate and reform all terms and relationships,then do the formalization process to produce meaningful models at the knowledge level. All possibility tables and fields name in the database system will be represented into classes and subclasses term in the ontology knowledge. Furthermore, in the formalization phase, every class or subclass term will be given a semantic relationships between them. Table 2 portray all relationships that used in the ontology knowledge. Table 3 show all possibility terms on SGS and GS to be candidate of classes and subclasses for ontology knowledge.

4.1.4. Ontology development

Ontology development phase is the process to develop ontology knowledge for specific domain and purpose using certain tool or application. The ontology development in this research is using Protégé tool as the one of recommendation tool to develop ontology knowledge. The main reasons to use Protégé as a tool to develop the ontology knowledge is because of Protégé is a free tools and has a reasoner feature to evaluate and validate the ontology knowledge.

The result from the ontology development using Protégé is Web Ontology Language (OWL) syntax that can be used in programming language such as JAVA, programming language. Protégé also provide others options to save ontology knowledge into RDF/XML file format, OWL/XML format, OWL Functional Syntax, KRSS2 Syntax, OBO Format and Manchester OWL Syntax. Figure 3 shows the ontology knowledge in the diagram view which is exported by OntoGraf feature in the protégé.

4.1.5. Ontology evaluation and validation

Evaluation and validation phase is the process to verify the consistency level acceptance of the ontology knowledge. Consistency level is about semantic terms and relationships that used in the ontology to verify and validate weather the ontology sill has any inconsistency or all semantic terms and relationships already in the consistency level acceptance. In this research the evaluation and validation of the ontology is using reasoner in the protégé application. There are several reasoner standard in the protégé tool, such as FaCT++, HermiT and Pellet. Figure 4 show the evaluation and validation process using FaCT++ on the protégé.

4.1.6. Ontology refinement

The refinement process need to do when the evaluation and validation process using reasoner get some errors message. Ontology refinement phase is the iterative process to edit and improve the ontology knowledge to get better ontology result and to fulfil the evaluation and validation process for the consistency level acceptance.

4.1.7. Completion of documentation

The documentation process in the OntoDI is conducted from the beginning when the first phase on OntoDI started. On the last phase on OntoDI the completion of documentation of ontology knowledge need to be done and to be final version of documentation file. Documentation file will help the ontology client/user to customize and improve the ontology to adjust with any changes in the future.

4.2. Discussion of OntoDI

The development of ontology knowledge using OntoDI has been completed to be implemented in education domain.We claim that using the OntoDI to develop ontology knowledge give simpler phases, complete steps, and clear documentation for the ontology client. The crucial phase on OntoDI for data integration implementation is on the resources identification phase. In this phase, OntoDI is trying to identify and select a specific data or information that want to integrate. Furthermore, in the next phase, the ontology development started with the conceptualization and formalization of all data that related with the resource and generate all terms into classes and subclasses on ontology perspective. After generating process, then formalize it using semantic relationships.One of the advantage of the OntoDI is on the documentation process, because the ontology developer also create documentation file from the beginning phase of OntoDI and completing it on the last phase.

5. Conclusions and Future Works

Ontology become popular in the recent years, because there are a lot of semantic aspect problem in many system implementation domain. Especially for data integration implementation, ontology is become the one of the solution to solve semantic aspect problem on the implementation of data integration. There are a lot of ontology development methodologies have been produced, however based on study literature there is no existing ontology development methodologies for data integration implementation. This research has successfully create new methods for ontology development on data integration (OntoDI). It is also applied in the education domain to see OntoDI performance in more details. The ultimate goal of the OntoDI is to make customization, improvement and simplification from existing methodologies to get better ontology development result for data integration area.

In the future work we will examine OntoDI with other real case study implementation and tried to implement the ontology result on the data integration implementation. In the future work also we will do critical evaluation to improve OntoDI for the better ontology development in the future.

References

1. Fagin, R.; Kolaitis, P.; and Popa, L. (2005). Data exchange: getting to the core. ACM Transactions on Database Systems, 30(1), 174-210.

2. Calvanese, D.; Giacomo, G.; Lenzerini, M.; and Rosati, R. (2004). Logical Foundations of Peer-To-Peer Data Integration. *Proceedings of the 23rd ACM SIGMOD Symposium on Principles of Database Systems, PODS 2004*. 241-251.

3. Zheng, L.; and Terpenny, J. (2013). A hybrid ontology approach for integration of obsolescence information. *Journal of Computers & Industrial Engineering*,65(3), 485-499.

4. Jegadeesan,R.; Sankar Ram,N. "Energy Consumption Power Aware Data Delivery in Wireless Network", Circuits and Systems, Scientific Research Publisher,2016 (Annexure-I updated Journal 2016)

5. Sandborn, P.; Terpenny, J.; Rai, R.; Nelson, R.; Zheng, L.; and Schafer, C. (2011). Knowledge representation and design for managing product obsolescence. *Proceedings of NSF civil, mechanical and manufacturing innovation grantees conference*. Atlanta, Georgia.

6. Ke, S.; Feng, G.; Qing, X.; and Guoyan, X. (2014). Integration framework with semantic aspect of heterogeneous system based on ontology and ESB. *Proceedings of the Control and Decision Conference*.4143-4148.

7. Ekaputra, F.J.; Serrai, E.; Winkler, D.; and Biffl, S. (2014). A semantic framework for data integration and communication in project consortia. *Proceedings of the Data and Software Engineering (ICODSE), 2014 International Conference*. 1-6.

8. Bansal, S.K. (2014). Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration. *Proceedings of the Big Data (BigData Congress), 2014 IEEE International Congress*. 522-529.

9. Jegadeesan, R.; Sankar Ram,N. "Energy-Efficient Wireless Network Communication with Priority Packet Based QoS Scheduling", Asian Journal of Information Technology(AJIT) 15(8): 1396-1404,2016 ISSN: 1682-3915,Medwell Journal,2016 (Annexure-I updated Journal 2016)

10. Vavliakis, K.N.; Grollios, T.K.; and Mitkas, P.A. (2013). RDOTE – Publishing Relational Databases into the Semantic Web.*Journal of Systems and Software*, 86(1), 89-99.

11. Sonsilphong, S.; and Arch-int, N. (2013). Semantic Interoperability for Data Integration Framework using Semantic Web Services and Rule-based Inference: A case study in healthcare domain.*Journal of Convergence Information Technology(JCIT)*, 8(3), 150-159.

12. Nguyen, T.H.; Prinz, A.; Friisø, T.; Nossum, R.; and Tyapin, I. (2013). A framework for data integration of offshore wind farms. *Journal of Renewable Energy*, 60(0), 150-161.

13. Wiesner, A.; Morbach, J.; and Marquardt, W. (2011). Information integration in chemical process engineering based on semantic technologies. *Journal of Computers & Chemical Engineering*. 35(4), 692-708.

14. Yunianta, A.; Barukab, O.M.; Yusof, N.; Dengen, N.; Haviluddin, H.; and Othman, M.S. (2017). Semantic data mapping technology to solve semantic data problem on heterogeneity aspect.*International Journal of Advances in Intelligent Informatics*, 3(3), 161-172.

15. Jaziri, W.; and Gargouri, F. (2010). *Ontology Theory, Management and Design: An Overview and Future Directions. Ontology Theory, Management and Design: Advanced Tools and Models*. pp. 27-77. IGI Global.

16. Yunianta, A.; Barukah, O.M.; Yusof, N.; Musdholifah, A.; Jayadiyanti, H.; Dengen, N.; Haviluddin; and Othman, M.S. (2017). The ontology-based methodology phases to develop multi-agent system (OmMAS). *Proceedings of the 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. 1-6.

17. Jegadeesan,R.; Sankar Ram "Defending Wireless Sensor Network using Randomized Routing "International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 9, September 2015 ISSN: 2277 128X Page | 934-938

18. Badr, K.B.A.; Badr, A.B.A.; and Ahmad, M.N. (2013). *Phases in Ontology Building Methodologies: A Recent Review in Ontology-Based Applications for Enterprise Systems and Knowledge Management*. 100-123. IGI Global.

19. Abdullah, N.S.; Sadiq, S.; and Indulska, M. (2013). A Study of Ontology Construction: The Case of a Compliance Management Ontology. *Ontology-Based Applications for Enterprise Systems and Knowledge Management.* pp. 276-291. IGI Global.

20. Uschold, M.; and King, M. (1995). Towards a methodology for building ontologies. *Paper presented at the Workshop on Basic Ontological Issues in Knowledge Sharing, held in conduction with IJCAI-95*. New York, NY.

21. Schreiber, G.; Wielinga, B.; and Jansweijer, W. (1995). The KACTUS view on the 'O'word. *IJCAI workshop on basic ontological issues in knowledge sharing*. 159–168.

22. Gruninger, M.; and Fox, M.S. (1995). Methodology for the design and evaluation of ontologies.*Workshop on Basic Ontological Issues in Knowledgen Sharing: International Joint Conference on Artificial Inteligence (IJCAI95).* New York, NY.

23. Jegadeesan,R.; Karpagam, Sankar Ram , "Defending Wireless Network using Randomized Routing Process" International journal of Emerging Research in management and Technology ISSN: 2278-9359 (Volume-3, Issue-3) . March 2014

24. Euzenat, J. (1995). Building Consensual Knowledge Bases: Context and Architecture.*Proceedings of the KB\&KS '95 Conference*. 143-155.

25. Jegadeesan,R.; Sankar Ram,Tharani (September-October, 2013) "Enhancing File Security by Integrating Steganography Technique in Linux Kernel" Global journal of Engineering,Design & Technology G.J. E.D.T., Vol. 2(5): Page No:9-14 ISSN: 2319 – 7293

26. Lenat, D.B.; and Guha, R.V. (1989). *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc.

27. Uschold, M. (1996). Building ontologies: Towards a unified methodology. *Proceedings of the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems*. London, UK.

28. Fernandez, M.; Gomez-Perez, A.; and Juristo, N. (1997). Methontology: From ontological art towards ontological engineering. *Proceedings of the AAAI 1997 Spring Symposium Series*. Menlo Park, CA.

29. Swartout, B.; Patil, R.; Knight, K.; and Russ, T. (1997). Towards distributed use of large-scale ontologies. *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop*. Banff, Canada.

30. Richard, B.V.; Fensel, D.; Decker, S.; and PÉRez, A.G. (1999). (KA)2: building ontologies for the Internet: a mid-term report.*International Journal of Human-Computer Studies*, 51(3), 687-712.

31. Pinto, H.S.; and Martins, J.P. (2001). A methodology for ontology integration. *Proceedings of the 1st international conference on Knowledge capture*. Victoria, British Columbia, Canada, 131-138.

32. Staab, S.; Studer, R.; Schnurr, H.-P.; and Sure, Y. (2001). Knowledge Processes and Ontologies. *Journal of IEEE Intelligent Systems*, 16(1), 26-34.

33. Pinto, H.S.; Staab, S.; and Tempich, C. (2004). DILIGENT: towards a fine-grained methodology for distributed, loosely-controlled and evolving Engineering of ontologies. *Proceedings of the 16th European Conference on Artificial Intelligence*. Valencia, Spain, 393-397.

34. Paredes-Moreno, A.; Martínez-López, F.J.; and Schwartz, D.G. (2010). A methodology for the semi-automatic creation of data-driven detailed business ontologies. *Journal of Information Systems*, 35(7), 758-773.