Journal of Theoretical and Applied Information Technology
15th October 2016. Vol.92. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTEGRATION OF SCIENTIFIC EXPERIMENTAL DATA THROUGH ONTOLOGY APPROACH: A REVIEW

[1]NUR ADILA AZRAM, [2]RODZIAH ATAN

[1]Halal Products Research Institute, Universiti Putra Malaysia, Malaysia

[2]Assoc. Prof., Department of Software Engineering and Information System, Universiti Putra Malaysia, Malaysia

E-mail: [1]nuradila.azram@yahoo.com, [2]rodziah@upm.edu.my

ABSTRACT

Data integration in scientific experiments is important to the scientists in many research domains. This is because many experimental data involved multidiscipline areas and run in different machines or instruments which results in data stored in different ends and human intervention is required in forming a chain of data analysis. Ontology is one of the approaches that have been used in data integration in many domain areas. This paper described and reviewed ontology in data integration effort. Furthermore, the state of research for ontology-based integration of scientific experiment data also covered in this paper.

Keywords: *Data Integration, Ontology, Scientific Experiment, Scientific Research Data And Ontology-Based Data Integration*

1. INTRODUCTION

Data integration is a process where multiple data from different sources are combined through a single access point. It also can be defined as the problem of gathering related information from disparate sources and presenting it in a unified schema and semantic heterogeneity [1]. Data integration has become essential to multidiscipline domain of research areas such as biomedical, medical and epidemiology integrated.

Scientific researchers need an effective system or platform to not only manage their data, results and experiments but also to share and search the data. However, with multidiscipline domain involved, it is challenging for experiment data collection, analysis, management and sharing due to information infrastructure [2].

One of the approaches used for data integration is ontology. Ontology can be defined as representation of knowledge for a particular subject or domain which is written with standardized and structured syntax [3]. Using ontology would ease in identify more complex relationships in data, greater interoperability and more efficient using computer reasoning.

The coverage of the literatures and selection of reviews in this paper are based on the general knowledge information of ontology as data integration approach and the application of ontology approach in scientific experimental data integration.

So in this paper, we give an overview of the use of ontology for data integration consist of ontology architecture types, ontology components and ontology engineering comprising of ontology languages and tools. We will also review on state of the research of ontology-based approach for integration of scientific experimental data from domains which deals with these such as medical, biology, biomedical and epidemiology. The objectives of this review are to understand the concept of ontology approach for data integration as well as identify existing ontology-based approach for scientific experimental data integration.

2. PREVIOUS REVIEW ON ONTOLOGY-BASED DATA INTEGRATION

From literature that we have done, we found some paper that done a review on ontology-based data integration approaches.

Paper [4] reviewed on existing approaches of ontology-based integration of heterogeneous information sources. They analyze about 25 approaches such as SIMS, OBSERVER and KRAFT using four main criteria which are use of ontologies, ontology representation, use of mappings and ontology engineering. They evaluated and compared the languages used to represent ontologies, the use of mappings between ontologies and also evaluated ontologies connection with information sources, They conclude that there is a need to investigate mappings on a theoretical and an empirical basis as well as a need to develop a more general methodology that includes an analysis of the integration task and supports the process of defining the role of ontologies.

Reviewed by [5] described ontology-based data integration for seven systems (SIMS, OBSERVER, KRAFT etc) and three proposals which solves the problems of semantic heterogeneity. They use a framework (DESMET method) in order to compare the different approaches. The framework was divided into three main features in which each features consists of sub-features. The three features are architecture (information sources and architecture type), semantic heterogeneity (ontology use and representation language) and query resolution (understandability, query plan and optimization). Based on their comparison, they found some elements in common as well as original aspects of the systems. With their analysis, they hope it would help ontology-based data integration community in comparing different aspects of systems as a reference for further research. They also conclude that other several aspects need to be analyzing such as comparison of optimization techniques applied to the query plans.

Even though the reviewed paper mentioned above discussed on ontology-based data integration, it is for heterogeneous information sources and semantic heterogeneity which is different from what we want to reviewed. In this paper, we reviewed on the concepts of ontology as well as ontology-based data integration for scientific experimental data.

3. ONTOLOGY FOR DATA INTEGRATION

Data integration using ontology is defined as an explicit specification of a conceptualization [6] in which conceptualization refers to an abstract model of how people commonly think about a real thing in the world and explicit specification means that concepts and relationships of an abstract model receive explicit names and definitions [7].

Ontology can be use for data integration as it provides a vocabulary to represent and communicate domain knowledge along with a set of relationships containing the vocabulary's terms at a conceptual level. Thus, it can explicitly describe the semantics of data in information sources and to solve heterogeneity problems.

Using ontology gives many advantages. It provides high-level knowledge management capabilities and also supports consistent management. It also can facilitate interoperability by supporting communication and cooperation between systems developed at different sites.

Generally, there are three main architecture approaches that can be used in ontology-based data integration to describe the data source semantics and to make to content explicit. Section 3.1 gives an overview of these three architectures.

3.1 Ontology Architectures

Ontology has been used in data integration because they provide an explicit and machine-understandable conceptualization of a domain [8]. The three main ontology architectures are single ontology approaches, multiple single approaches and hybrid ontology approaches. Table 1 gives a brief overview of the ontology architectures.

Table 1: Overview Of The Ontology Architectures

Architecture Type	Overview
Single Ontology approaches	Use a domain ontology providing a shared vocabulary for the specification of the semantic and relate all data sources to one global ontology (see Figure 1).
Multiple Ontology approaches	Each data source is described by its own domain or application-specific ontology (see Figure 2).
Hybrid Ontology approaches	Similar to multiple ontology approaches but the source ontologies are built upon one global shared vocabulary to make it comparable to each other (see Figure 3).

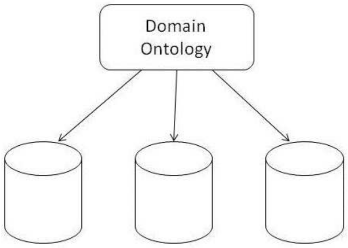

Figure 1: Single Ontology Approach

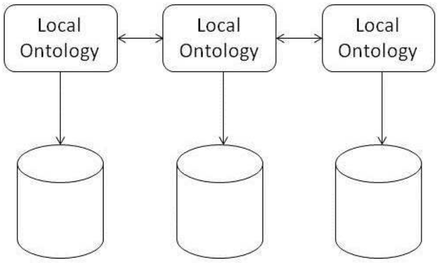

Figure 2: Multiple Ontology Approach

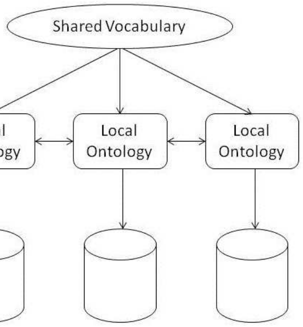

Figure 3: Hybrid Ontology Approach

Each of the ontology architecture has their own advantages and disadvantages in applying them. Table 2 lists the advantages and disadvantages of the architectures.

Table 2: Advantages And Disadvantages Of Each Ontology Architectures

Architecture Type	Advantage(s)	Disadvantage(s)
Single ontology approach	Can be applied for integration where all data sources to be integrated provides nearly the same view on a domain.	Vulnerable to changes in the data sources which can affect the conceptualization of the domain represented in the ontology.
Multiple ontology approach	No common and minimal ontology commitment around and ontology is needed.	Lack of common vocabulary makes it difficult to compare different source ontologies. Need an additional representation formalism defining the inter-ontology mapping.
Hybrid ontology approach	New sources can be easily added without modification needed. Supports acquisition and evolution of ontologies. Make the source ontologies comparable with the usage of shared vocabulary.	Exisiting ontologies cannot be reused easily and have to be redeveloped from scratch.

3.2 Ontology Components

Ontologies consist of several common components such as individuals (instances), classes (concepts), attributes, relations and axioms. The following paragraph gives the definitions of the components.

Individuals (instances) are the `things' represented by a concept. **Concepts (classes)** are a set or class of entities that can be divide into two categories which are *primitive concepts* (those which only have necessary conditions (in terms of their properties) for membership of the class) and *defined concepts* (those whose description is both necessary and sufficient for a thing to be a member of the class). **Attributes** are characteristics or features that classes can have.

Relations describe the interactions between concepts and properties. There are also two types of relations which are *taxonomies* that organise concepts into sub- super-concept tree structures and *associative* relationships that represent the functions, processes a concept has or is involved in, and other properties of the concept within a domain. **Axioms** are used to declare and restrict values for classes or instances.

3.3 Ontology Engineering

Ontology engineering is a research methodology which gives the design rationale of a knowledge base, kernel conceptualization of the world of interest, semantic constraints of concepts together with sophisticated theories and technologies enabling accumulation of knowledge which is dispensable for knowledge processing in the real world [9]. In other word, it refers to the activities of ontology development process, the ontology life cycle, the methods and methodologies for building ontology, and the tool suites and languages that support them.

It aims at making explicit the knowledge contained within software applications, and within enterprises and business procedures for a particular domain and offers a direction towards solving the inter-operability problems brought about by semantic obstacles [10]. Some of the methodologies to create single ontology from scratch include Uschold and King's method, Grüninger and Fox's methodology, KACTUS approach and METHONTOLOGY.

Ontology Languages are formal languages used to construct ontologies. It must describe meaning in a machine-readable way so that an ontology language needs not only to include the ability to specify vocabulary but also the means to formally define it in such a way that it will work for automated reasoning [11]. Ontology languages are usually declarative languages, are almost always generalizations of frame languages, and are commonly based on either first-order logic or on description logic [12].

There are two categories of ontology language. The first category is traditional ontology languages that are based on first-order predicate logic, frame-based languages; description logic (DL) based language and other languages. The second category is Web-based ontology languages, which are used to facilitate interchange on the Internet, and ontology languages, which are web standards compatible [13]. Table 3 shows the example of ontology languages according to their types.

Table 3: Example Of Ontology Languages According To Category And Logic Type

Categories of Ontology Language	Logic Type	Example(s)
Traditional ontology languages	First-order predicate logic	KIF, CycL, Common Logic
	Frame-based languages	Ontolingua, F-logic and OCML, OKBC, KM
	Description logic (DL) based languages	LOOM, KL-ONE, RACER
Web based ontology languages	-	OWL, RDF, RDFS, OIL, DAML+OIL

Ontology editors are tools or applications designed to aid in the creation or manipulation of ontology and often express ontology in one of many ontology languages. Some of the criteria in choosing the right ontology editors are the degree to which the editor abstracts from the actual ontology representation language used for determination and the visual navigation possibilities within the knowledge model, built-in inference engines and information extraction facilities, and the support of meta-ontologies [14]. Another criterion is the ability to import and export unfamiliar knowledge representation languages for ontology matching.

Table 4 gives an example of some of the ontology editors with descriptions.

Table 4: Examples Of Ontology Editor With Descriptions

Ontology Editor	Description
HOZO [14]	Java-based graphical editor especially created to produce heavy-weight and well planned ontologies.
NeOn Toolkit [14]	Eclipse-based, open source, OWL support, several import mechanisms, support for reuse and management of networked ontologies and visualization.
Protégé [15]	Open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-base applications with ontologies.
OntoStudio [15]	An Ontology Engineering Environment that are based on IBM Eclipse framework which support the development and maintenance of ontologies by using graphical means and also based on client/server architecture.
Swoop [15]	An open-source, Web-based OWL ontology editor and browser that contains OWL validation, offers various OWL presentation syntax views, has reasoning support (OWL Inference Engine), and provides a Multiple Ontology environment.

4. SCIENTIFIC EXPERIMENT DATA INTEGRATION

In scientific research areas, often there is a need to exchange valuable data or information between different researchers or research domain [16]. In all areas of science there is even more data and information to understand and, in some fields, this increase in data and information has become a ‘deluge’ [17]. Hence, this result in the increase of dependence on computers to store, integrate and analyze data.

Integration for scientific research or experiment data is not an easy thing to do because most scientific research areas involved multidiscipline domain. Different domains involved means variety of computing platforms, data storage environments, structures and models used. There is a need to have a ways or solutions in making data integration for scientific research easier to be done. Ontology is seen to be one of the best approaches to integrate these scientific data. As most characteristics feature in science is experiment-based, the development of ontology of experiments is a fundamental step in formalizing the integration of science [18].

We have reviewed several literatures that use ontology approach to integrate scientific data. The reviewed literature are selected based on its intended purpose which are generally used for any scientific data integration or specifically used for specialize domain. However, from our review, we identified that currently not many general-purpose ontology for scientific experiments proposed. Two general-purpose ontology for scientific experiments that were identified from literature are Basic Formal Ontology (BFO) and EXPO which will be explains in Section 4.1.

Nevertheless, several ontologies exist for specialized experimental research domains such as in biology, medical, biomedical and epidemiology. We will review several specialized ontology for each stated domains in Section 4.2.

4.1 General-purpose Ontology of Scientific Experiment

Basic Formal Ontology (BFO) is a strict and small upper-level ontology developed to support integration of data obtained through scientific research. It does not contain its own representations of physical, chemical, biological, psychological, or other types of entities which would properly fall within the domains of the special sciences [17]. BFO defines framework that will help to ensure consistency and non-redundancy of the ontologies created in it terms. It can be classified into three fundamental divisions. The first division is between continuants (entities that persist, endure, or continue to exist through time) and occurrents (events or happenings in which continuants participate). The second division is between dependent and independent entities. The third division is between instances and universals which furnishes formal specifications for the high-level formal universals (called 'categories' in what follows) which can be defined in terms of these three divisions, and also of a set of relations which link them [18]. Applications that have adopt BFO as a foundational ontology includes biomedical, security and defence.

Common ontology of scientific experiments (EXPO) was developed to formalize generic knowledge about scientific experimental design, methodology and results representation which would be practical and aimed for because all sciences follow the same experimental principles. EXPO is to abstract out the fundamental concepts in formalizing experiments that are domain independent [19]. The advantages of using EXPO would be it ensures consistency, clean updating and non-redundancy. The principle of designing EXPO is a combination of top-down bottom-up methodology in which Suggested Upper Merged Ontology (SUMO) was selected as the upper ontology as a reference. Validity of EXPO were tested in different scientific domains such as microbiological, particle physics and computer science to make sure that the classes in EXPO cover the essential concepts of scientific experiments.

4.2 Specialize Ontology of Scientific Experiment

As mentioned in Section 4, there are several ontologies specialized for specific scientific research domains. We review these specialized ontologies in this section to identify and gain understanding on specific domains that use ontology for scientific experiments data integration. Table 5 shows the example of domains with their specialized ontology currently existed.

Table 5: Example of domains with specialized ontology

Domain	Specialized Ontology (s)
Biology [20,22]	Preclinical Investigation of Bio Active Substances (PIBAS)
Biomedical [21]	Open Biomedical Ontologies (OBO), Ontology for Biomedical Investigations (OBI)
Medical [16,22]	Ontology-based UMLS Integration Project (OUIP)
Epidemiology [23, 24]	Epidemiology Ontology (EPO), Network of Epidemiology-Related Ontologies (NERO)

From the example in Table 5, we give a details overview for each of the specialized ontology that has been stated.

PIBAS for biology domain is ontology for modeling complex experimental structure of the Research Center (RC) for testing of active substances. It is designed to support laboratory staff to quickly reference and use complex experiment structure. It is suggested as a universal mean for fast and easy representing of required various semantic structures [20]. PIBAS was created with Protégé platform and represented in RDF/XML file. SPARQL was used as the ontology query language and visualization of the ontology is done using the InfoVis Toolkit library.

OBO for biomedical domain is an ontology library that contains interoperable reference ontologies and provides a set of principles for ontology development. It comprised more than 70 biomedical ontologies. Its role as an ontology information resource is supported by the NIH Roadmap National Center for Biomedical Ontology (NCBO) through its BioPortal [21].

OBI is an integrated ontology to serve the coordinated representation of designs, protocols, instrumentation, materials, processes, data and types of analysis in biology and medicine domain [22]. It addresses the need for controlled vocabularies to support integration of experimental data. OBI uses the OWL-DL Web Ontology Language.

OUIP for medical domain is a hybrid ontology that was constructed to describe the meta-structure of the code information that was stored in the Mid-America Heart Institute (MAHI) Data Repository. It represents clearly a shared understanding of the important concepts in the MAHI cardiovascular domain [16] and provides transparent access to heterogeneous data sources. OUIP was created using Protégé-2000 and stored in RDF file format.

EPO is an ontology designed to support the semantic annotation of epidemiology resources, data integration, information retrieval and knowledge discovery activities in epidemiology domain. It follows the OBO Foundry guidelines and uses the BFO as an upper ontology. Dictionary of Epidemiology (DoE) which is a well-established reference that captures the categorization commonly used in epidemiology is used in the creation of the EPO. It currently covers three main areas which are transmission mode, epidemiological parameters and demographic parameters [23]. It was created using Protégé 4.1 and encoded in Web Ontology Language–Description Logic of the W3 Consortium (OWL-DL). EPO has also been integrated in NERO.

NERO is a compilation of ontologies that support the semantic annotations of epidemiology domain resources. It currently includes thirteen external ontologies and resources such as Disease Ontology, Symptom Ontology and Vaccine Ontology [23]. It contributes to the preservation of epidemiological resources by allows the full scale of semantic web technologies to be used to search resources and enables performing simple but powerful queries. Its ontologies selected to ensure availability and longevity, and also the meaning of concepts is guaranteed to remain unchanged [24]. NERO is integrated in the Epidemic Marketplace (EM) which is a platform for epidemic research that enables and encourages epidemiological data sharing, enabling the community to perform data intensive research [25].

5. RESEARCH ISSUES IN SCIENTIFIC EXPERIMENTAL DATA INTEGRATION

From our review, we found that there is not much general-purpose ontology for scientific research or experimental data but there are many specialized ontology for specific domains existed. The lack of general-purpose ontology for scientific data because scientific experiments usually involved multidiscipline domains which results in scientific data varies in terms of data formats and types. So it is difficult to establish a general-purpose ontology that would suits any domains.

With many specialized ontology for specific domains existed, we can say that some specialize ontology related to one another because most of the ontology was for domains that involved with multidiscipline areas. So, there are certain specialize ontology that used or integrate ontology from other domain which suited and can be used. However, it would be difficult and time consuming to identify which specialize ontology would suit which domains.

6. CONCLUSION

In this paper, we presented an overview and stated of the research for ontology-based integration of scientific experiment data. We give overview on the use of ontology in data integration as well as related topics to ontology such as architecture types and components. We also review various ontologies for scientific research data.

The goal of this evaluation is to gain knowledge and understanding on the concept of ontology-based approach for data integration as well as the elements involved in creating ontology. We also want to identify existing ontology-approach for integration of scientific experimental data so that we can study and use it as a reference to establish new ontology that would suit any domains which involved with scientific experiment data.

We conclude that there is a need to have ontology for integration of scientific research data that are broad-spectrum and not domain specific so that it can be utilized and used by various research areas.

REFRENCES:

[1] S. Ae Chun and B. MacKeller, "Social Health Data Integration using Semantic Web", In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012, pp. 392-397.

[2] F. Wang, P. Liu, J. Pearson, F. Azar, and G. Madlmayr, "Experiment Management with Metadata-based Integration for Collaborative Scientific Research", In *Proceedings of the 22nd International Conference on Data Engineering*, 2006.

[3] P. Alexander, "The Importance of Ontologies", *In The MMI Guides: Navigating the World of Marine Metadata*. http://marinemetadata.org/guides/vocabs/ont/importance. Accessed April 16, 2016.

[4] H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hubner, "Ontology-Based Integration of Information – A Survey on Existing Approaches", *In: Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing*, 2001, pp. 108-117.

[5] A. Buccella, A. Cechich and N. R. Brisaboa, "Ontology-Based Data Integration Methods: A Framework for Comparison", *JCS&T*, Vol 3, No 2, 2003, pp. 62-68.

[6] T. Gruber, "A translation approach to portable ontology specifications", *Knowledge Acquisition 1993*, Vol. 5, No. 2, 1993, pp. 199–220.

[7] U. Visser, "Intelligent Information Integration for the Semantic Web", *Lecture Notes in Artificial Intelligence,* Vol. 3159, 2004.

[8] I. F. Cruz and H. Xiao, "The Role of Ontologies in Data Integration", *Journal of Engineering Intelligent Systems*, Vol. 13, No. 4, 2005, pp. 245–252.

[9] R. Mizoguchi, "Tutorial on ontological engineering part 1: Introduction to ontological engineering", *New Generation Computing*, Vol. 21, No. 4, 2003, pp. 365–384.

[10] L. Pouchard, N. Ivezic and C. Schlenoff, "Ontology Engineering for Distributed Collaboration in Manufacturing", In *Proceedings of the AIS2000 conference*, March 2000.

[11] J.R.G. Pulido, M.A.G. Ruiz, R. Herrera, E. Cabello, S. Legrand and D. Elliman, "Ontology languages for the semantic web: a never completely updated review", *Knowledge-Based Systems*, Vol. 19, 2006, pp. 489–497.

[12]https://en.wikipedia.org/wiki/Ontology_language. Accessed April 16,2016.

[13] X. Su and L. Ilebrekke, "A Comparative Study of Ontology Languages and Tools", In *Proceeding of the 14th Conference on Advanced Information Systems Engineering (CAiSE'02)*, Vol. 2348, 2002, pp. 761-765.

[14]https://en.wikipedia.org/wiki/Ontology_%28information_science%29#Editor. Accessed April 16, 2016.

[15] E. Alatrish, "Comparison Some of Ontology Editors", *Management Information Systems*, Vol. 8, No. 2, 2013, pp. 18-24.

[16] Q. Chong, A. Marwadi, K. Supekar and Y. Lee, "Ontology based metadata management in medical domains", *Journal of Research Practice in Information Technology*, Vol. 35, No. 2, 2003, pp. 139–154.

[17] B. Smith, "On Classifying Material Entities in Basic Formal Ontology, Interdisciplinary Ontology", *Proceedings of the Third Interdisciplinary Ontology Meeting*, Tokyo: Keio University Press, 2012, pp. 1-13.

[18] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector and C. Rosse, "Relations in Biomedical Ontologies", *Genome Biology*, Vol. 6, No. 5, R46, 2005.

[19] L. N. Soldatova and R. D. King, "An Ontology of Scientific Experiments". *Journal of the Royal Society Interface*, Vol 3, 2006, pp. 795-803.

[20] V. Cvjetković, M. Đokić, B. Arsić and M. Ćurčić, "The ontology supported intelligent system for experiment search in the scientific research center", *Kragujevac Journal of Science*, Vol. 36, 2014, pp. 95-110.

[21] D. L. Rubin, S. E. Lewis, C.J. Mungall, et al., "National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge", *OMICS 2006 Summer,*Vol. 10, No. 2, 2006, pp. 185–198.

[22] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al., "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration", *Nature Biotechnology*, Vol. 25, 2007, pp. 1251–1255.

[23] C. Pesquita, J. D. Ferreira, F. M. Couto and M.J. Silva, "The pidemiology ontology: an ontology for the semantic annotation of epidemiological resources", *Journal of Biomedical Semantics*, Vol. 5, No. 4, 2014.

[24] J. D. Ferreira, C. Pesquita, F. M. Couto and M. J. Silva, “Digital preservation of epidemic resources: coupling metadata and ontologies”.

[25] L. Lyon, A. Ball, M. Duke, and M. Day, “Developing a Community Capability Model Framework for data-intensive research”, In *iPres 2012-9th International Conference on Preservation of Digital Objects*, 2012, pp. 9-16.