International Journal on Advanced Science Engineering Information Technology

Vol.6 (2016) No. 6
ISSN: 2088-5334

Building an Event Ontology for Historical Domain to Support Semantic Document Retrieval

Fatihah Ramli [#,*] and Shahrul Azman Mohd Noah [*]

[#] *Department of Information Systems, Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia*
E-mail: rfatihah@unimas.my

** Knowledge Technology Research Group, Faculty Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
E-mail: shahrul@ukm.edu.my

***Abstract*— In the past years, there has been increasing concern on ontology for its ability to explain data semantics in the usual manner independent of the data source characteristics, providing a schema that allows interchanging data between heterogeneous information systems and users. The ontology development in some areas is not expected due to a large amount of information, particularly in history, leading its semantic impossible. Several works have been designed to improve the technological aspects of ontology, such as the representation of language and inference mechanisms, and less attention has been paid to practical results development of application methods. This paper presents a discussion on the experience and processes during ontology building in history: historical documents retrieval based on the event.**

***Keywords*— ontology; history; event; historical document; historical domain; semantic retrieval**

I. INTRODUCTION

Ontology has received recognition from the academy and industry in various fields [1]. The definitions of ontology vary according to the fields and applications. In information science, ontology can be defined as a dictionary of terms formulated in a canonical syntax and with commonly accepted definitions designed to yield a lexical or taxonomical framework for knowledge representation, which can be shared by different information systems communities[2], [3]. Thus, ontology is said to be a representation of the things that exist within a particular domain of reality such as medicine, geography, finance, or history. The development of the ontology for these specific domains is meant to support the implementation of intelligent applications such as decision support systems[4], recommender systems[5] and semantic search[5],[6].

One of the domains receiving great attention recently is history[7], [8], which may be due to increasingly available digitised historical documents and artefacts to the public. History can be referred to as a period of time after writing was invented. It is a field of research that uses narrative to examine and analyse the sequence of events, and it sometimes attempts to investigate the patterns of cause and effect that determine events objectively. Research on managing historical documents involves finding, using and correlating the documents in order to communicate an understanding of past events. Historical documents can be defined as those that keep the information related with time instant at which the documents were published at the same time that is still useful in the future [9]. According to Elena [10], [11], within the context of the historical archive, historians employ their knowledge, experience and intuition to decide on the information that they need to find and study; and attempt to locate sources that contain the information. The results from Elena [10] obviously stated that historians need historical sources repositories and building tools to enable them to access comprehensive information rapidly. Among the most important information for them is the event. Questions such as: *When did the specific event occur?*, *What are the relations among events?*, *Who were involved?* and *List the chronological of specific events*. An obvious way to retrieve such information from large repositories is via information retrieval (IR) systems, or commercially known as search engines. IR is a field concerned with the structure, analysis, organisation, storage, search, and retrieval of information" [12]. According to [13], the need of IR research areas led to the creation of semantic web. However,

conventional IR systems are unable to support these specific requirements due to the simple bag-of-words document representation. One of the ways to support the requirements is to semantically enhance the document representation using ontology. As a result, the development of ontology is a crucial aspect of supporting semantic retrieval and organisation of domain-specific documents [14]. Therefore, this paper describes the event ontology for a historical domain and its development to support the semantic retrieval and organisation of historical documents.

There are several works done on how to develop ontologies methodologically. For instance, among the proposed methodologies are Gruninger and Fox [15], 101 Method [16] and METHONTOLOGY [17], [18]. These methodologies were successfully used to define ontologies in the different domain[19], [20]. They presented different intermediate representations in their works. The purpose of intermediate representations is to organise knowledge domain in the conceptualization phase. For our work, we used the 101 and METHONTOLOGY methods as approaches for building the ontology. We develop our own ontology based on the historical domain to have a further understanding of the use of ontologies and the process for building it from existing ontologies in a specific domain. In this work, the event ontology describes the historical domain of the battles and operations in the Vietnam War. The ontology objective is to facilitate and support the query and retrieval of historical documents based on event query from battles and operations in the Vietnam War. This paper is organised as follows: section 2 discussed the methods carried out to build the event ontology, section 3 presents the ontology evaluation using TopBraid Composer and finally, section 4 presents our conclusions.

II. Material and Mehod

The objective of this section is to discuss the process of developing an event ontology for the historical domain that describes the semantics of the domain.

A. Method Selection

A survey by Wache et al. [21] suggested that methods for ontology development can be grouped into two. The first group is considered as experienced-based methods such as the method proposed by [15], which was based on TOVE project and the Enterprise Model by [18]. The second group is structured methods, which are usually based on software or system development methods such as the evolutive prototype models relating to the METHONTOLOGY [17]. Gomez Perez [17] proposed a set of activities based on its life cycle and prototype refinement. Another example is the 101 Method [16], which proposed an iterative approach for ontology development. There is no single and widely accepted standard method to develop ontologies[1]. Normally, the first group is applied when the requirements are clear from the beginning; whereas the second group is employed when the objectives are not clear. Apart from that, both groups can be merged depending on the ontology users and goals. Ontology users are people who provide annotations and reviews for ontology development [22]. Building an ontology is a difficult task. For this work, ontology development consists of two phases, which are specification and conceptualization. The specification phase focuses on identifying and obtaining informal knowledge about the domain. Meanwhile, the goal of the conceptualization phase is to organise and structure the knowledge into concepts using external representation, which is independent of language implementation and the environment. In order to define the ontology for the historical domain, we followed the 101 Method as a guide to creating our first ontology and used METHONTOLOGY to perform the analysis in the conceptualisation pro-cess.

B. Specification: Ontology Goal and Scope

As mentioned in section 2.1, we chose the second group of the method to develop the ontology since our objectives were not clear from the start. The first step of the specification phase was defining the ontology goal and scope as illustrated in 101 Method[16] and METHONTOLOGY[17]. The scope limited to what should and should not be included. This was important to minimise the number of data and concepts that would be analysed for domain specific, especially for the complexity of historical semantics. The main focus in this ontology was to consider the event concepts with the sub-event concepts, as well as the related event concepts for historical documents. According to[23], [24], the event has fundamental types, which include temporal intervals, spaces and places, participation in events, influence, purpose and causality, parts and composition. In general, the fundamental types are categorised in terms of four *Wh*s: *What happened?*, *Where did it happen?*, *When did it happen?* and *Who was involved?* [23]. Meanwhile, Danzer [25] also mentioned about World History that focused on basic concepts of people, space and time. Thus, to complete our proposed ontology, we modelled other concepts such as location, date, people and cause as the fundamental types of event. In our case, an example of motivating scenarios is as shown in Table I, represented based on the template provided by [18]. In this scenario, the user requests information about the sub-event and the related event for a specific event. For example, the user's query is: Find sub-event, start date and end date for Battle of Ap Bau Bang II. The result will display documents on the subevent of Battle of Ap Bau Bang II as well as information about the start date and end date. The user will get the specific document that is related to Battle of Ap Bau Bang II. Thus, for this issue, we created a new ontology to support semantic document retrieval that enables a user to retrieve other documents simultaneously. Table I shows an example of the above scenario.

Competency questions were then constructed from the motivating scenarios in order to build the event ontology for the historical domain. It helps to verify whether or not sufficient information is available in order to achieve the goals and scope of ontology.

Table II shows the examples of possible competency questions in the historical domain of the battles and operations in the Vietnam War. These are informal questions that the ontology must be able to answer and will be used to check the ontology is fit to its purpose.

TABLE I
AN EXAMPLE OF SCENARIO DESCRIPTION

Scenario: 1
Name: Parts and composition
Actors: Users request for some other events that happened in one event.
Description: The scenario proposed here is a user who requests for some other events that happened in one event. For example, a user wants to get information about inter-relation event of "Battle of Ap Bau Bang II". The user does not have knowledge about any other battle and he might not be able to provide specific words to search engine (e.g. when searching for Battle of Ap Bau Bang II user will not specify Operation Junction City as well because both events are sub-event). Therefore, a user tends to get a lot of unstructured results. This will cause user frustrated and confused. This task is important for historical documents retrieval because is necessary to know and to get the documents accurately on the fast track.
Possible terms: event-Battle of Ap Bau Bang II (or another event from Battle of War II)

TABLE II
SAMPLE OF COMPETENCY QUESTIONS

No	Competency questions
1	Find sub-event, start date and end date for Battle of Ap Bau Bang II.
2	Find related event and person involved in Battle of Hamburger Hill.
3	Find related event and location for Operation Apache Snow.
4	Find sub-event and belligerent involved in Battle of Saigon 1968.

TABLE III
BASIC TERMS OF THE HISTORICAL DOMAIN

Items	Basic terms
1	event
2	sub event
3	related event
4	location
5	person
6	date
7	time
8	cause
9	unit
10	belligerent

C. Specification: Domain Description

We started with the domain analysis task where historical documents were studied and revised. We gathered all past research about ontology and historical documents that used a similar method as ours. We also explored if there is an existing ontology as a guideline for developing our first ontology. For our work, the analysis task assisted in formulating the enumerated terms into the conceptualization form for event ontology. For example, the existing ontology has numerous enumerated terms that we could consider and follow as a guide for creating our first ontology. Furthermore, experts also met to confirm the truth of the issues that were raised. The issues have been assimilated by experts on history and ontology who bring information to support these tasks. For instance, historians confirmed about historical information such as whether the event was an important element in history, while ontology engineer brought technical knowledge support for this task. In the end, we had to define the intermediate representations of knowledge acquisition. Table III illustrated the basic terms of the historical domain.

D. Conceptualization: Consider Reusing Existing Ontology

In this step was to consider reusing existing ontology developed or built by others. Available resources had to be checked whether they could be improved and expanded for our particular domain and task. The process of reusing existing ontology might be needed if our system had to interact with other applications committed for a specific ontology. The characteristics of ontologies itself encourage for a shared knowledge conceptualization. Therefore, reusing ontological sources increases application interoperability both on the syntactic and the semantic levels [26]. For our work, ontology reuse was very helpful especially in terms of time constraint to developing new ontology from scratch especially in adapting and updating the necessary module in a new ontology. The ontology development was executed semi-automatically and formalised by the domain experts and ontology developers. We reused the existing Simple News and Press Ontologies (SNaP) ontology and expanded it based on our vocabulary as shown in Fig. 1.

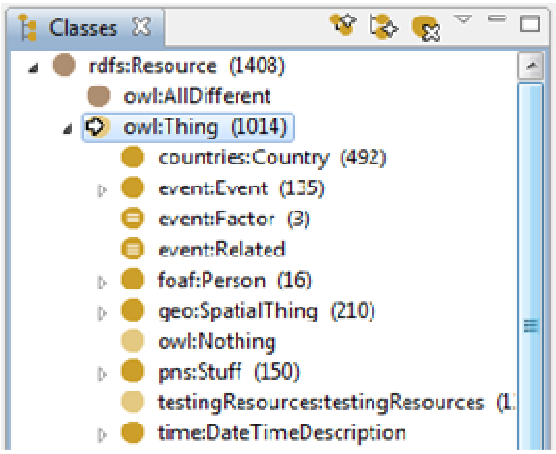

Fig. 1 Classes based on historical domain displayed using TopBraid Composer

SNaP ontology comprises of several ontologies, which describe assets (text, images, video) and the events and entities (people, places, organisations, abstract concepts etc.) that appear in a news content. Although it is meant for news document, it was found to be suitable in our case as it contains detailed representation about the event as well as documents (i.e. assets). The event ontology inherits fully from the public domain event Ontology. The object property of subEventOf is a rdfs:subPropertyOf event:sub_event with the addition of transitivity. Events are considered as compound entities in our domain (i.e. they are rich entities made through the relations with other entities, namely people, organisations, locations and things both tangible and intangible). Fig. 1 shows all the classes that were customised using TopBraid Composer.

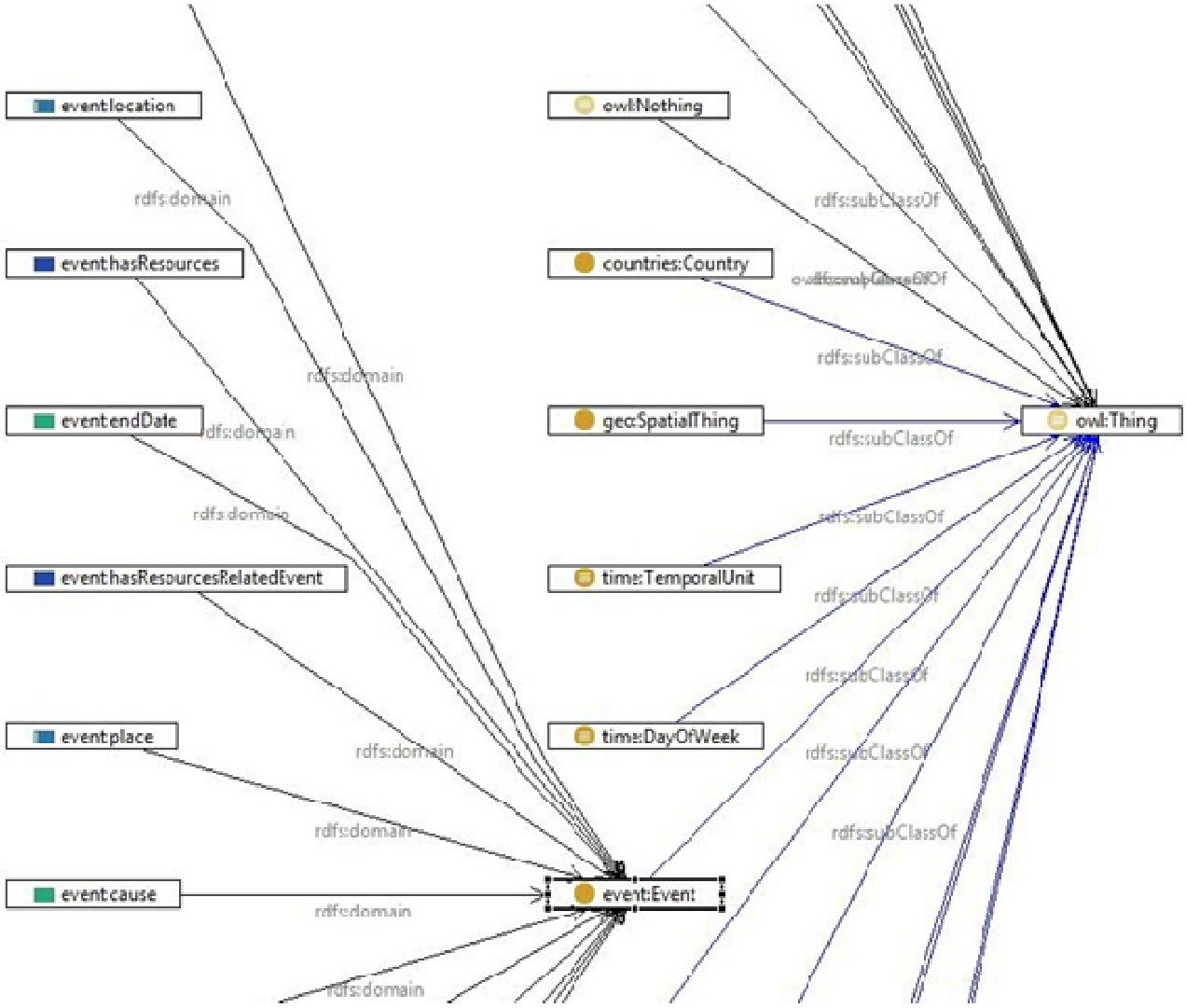

Fig. 2 Part of the domain ontology taxonomy

We have imported SNAP ontology into TopBraid Composer and started customising it based on our vocabulary i.e.: historical domain. Among the basic classes that were matched to our domain were event, factor, person, spatial thing (location) and time and date. Then, we expanded the ontology by adding some classes like country and stuff. The country class was added to know the country involved in each war, whereas stuff class includes both tangible and intangible entities to assign people involved in a war with their country and organisation. Figure 1 shows all the classes that were customized using TopBraid Composer.

E. Conceptualization: Define The Classes And The Class Hierarchy

In this step, the terms from the previous step were listed in the form of hierarchical taxonomy (see Fig. 2). The class hierarchy could be developed in three possible ways: top-down, bottom-up or a combination of both [16]. We chose the combination development process to define the few top level concept and few specific concept. For our work, the first step was to define a top level concept such as event, spatial thing, person, countries, date and factor. First, the event concept categorised all events in the battles and operations in the Vietnam War. The property `subEventOf` defined the sub-event for each instance of the event. Second, the `SpatialThing` concept described the places for each instance of the event connected by the location property. The third concept is the person, which defined the property of `involvedPerson` in order to know the person involved in each event. Fourth, the country concept was about countries engaged in the war in the same time. The `involvedCountry` property helped to create a relation between county and event concepts. The fifth concept was the date, which determined the `startDate` and `endDate` of each event. Finally, the factor concept defined the cause of each event. After that, we linked them to the middle-level concepts such as location, commander and others. Then, we expanded all possible classes that can produce relations between them.

F. Conceptualization: Define the Properties of Classes.

A property is a directed binary relation that defines the characteristic of the class. The class alone cannot provide

enough information to answer the competency questions listed in Table II. Therefore each class must have properties to provide detailed information for answering the competency questions. For instance, each type of event has instances such as the Battle of Ap Bau Bang II. Every instance of the event has properties such as `subEventOf`, `startDate`, `endDate` and others. Table IV consists of the core basic terms of concepts and properties for the event ontology. Meanwhile, Table V shows the relationship between concepts and properties. All the properties description generated from the TopBraid Composer (TBC) are listed in Table VI.

TABLE IV
CORE BASIC TERMS OF CONCEPTS AND PROPERTIES FOR EVENT ONTOLOGY

Concepts	**Properties**
Event	subEventOf
SpatialThing	location
Person	involvedPerson
Country	involvedCountry
Date	startDate & endDate
Factor	cause

TABLE V
RELATION BETWEEN CONCEPTS AND PROPERTIES

Concept Name	**Instance Name**	**Property**	**Value**
Event	Battle of Ap Bau Bang II	subEventOf	Operation Junction City
		startDate	19-3-1967
		endDate	20-3-1967

TABLE VI
DESCRIPTION OF PROPERTIES FOR EVENT ONTOLOGY

Properties	**Description**	**Function**
involvedCountry	Property defining a relation between belligerent and event	Country that involved in Battle of Ap Bau Bang II
involvedGroup	Property defining a relation between group and event	Group that involved in Battle of Ap Bau Bang II
involvedPerson	Property defining a relation between commander and event	Person that involved in Battle of Ap Bau Bang II
Location	Property defining a relation between location and event	Location for Battle of Ap Bau Bang II
producedIn	Property defining a parent-child relationship between events	Battle of An Lao is related event of Battle of Ap Bau Bang II
subEventOf	Transitive Property defining a parent-child relationship between events	Operation Junction City is a sub-event of Battle of Ap Bau Bang II
notablyAssiociatedWith	Property that notably associates stuff together	Giap Van Cuong is represented, Viet Cong
Cause	Property defining a relation between cause and event	Cause for Battle of Ap Bau Bang II
startDate	Property defining Start date and event	Start date for Battle of Ap Bau Bang II
endDate	Property defining End date and event	End date for Battle of Ap Bau Bang II

III. RESULTS AND DISCUSSION

A. *Implementation the Event Ontology with Topbraid Composer*

In order to implement the event-driven historical ontology, TopBraid Composer (TBC) tool was used. Particularly, we developed the event ontology in OWL and verified the accuracy and correctness of the information using SPARQL. SPARQL is the standardised query language for RDF. An SPARQL query consists of a set of triples where the subject, predicate and/or object can consist of variables. It also supports extensible value testing and constraining queries by source RDF graph. We produced several questions to ensure that all concepts and inference were created, and the information produced was correct and accurate. These questions were related to the competency questions listed in Table II. Examples of question used to ascertain the capacity of the ontology to answer all the competency questions are as follows:

Result 1: Find sub-event, start date and end date for Battle of Ap Bau Bang II (see Fig. 3).

```
SELECT ?subEvent ?startDate ?endDate
  WHERE {
     event:BattleofApBauBangII
     pne:subEventOf ?subEvent .
     event:BattleofApBauBangII
     event:startDate ?startDate .
     event:BattleofApBauBangII
     event:endDate ?endDate .
 }
```

[subEvent]	startDate	endDate
event:Operatio...	1967-03-19	1967-03-20

Fig. 3 Result for Competency Question 1

Result 2: Find related event and person involved in Battle of Hamburger Hill (see Fig. 4).

```
SELECT ?relatedEvent ?person
  WHERE {
       event:BattleofHamburgerHill
       event:producedIn ?relatedEvent .
      event:BattleofHamburgerHill
      event:involvedPerson ?person .
  }
```

[relatedEvent]	person
event:OperationApacheS...	pns:Ma_Vinh_Lan

Fig. 4 Result for Competency Question 2

Result 3: Find related event and location for Operation Apache Snow (see Fig. 5).

```
SELECT ?relatedEvent ?Places
  WHERE {
    event:OperationApacheSnow
    event:producedIn ?relatedEvent .
    event:OperationApacheSnow
    event:location ?Places .
   }
```

[relatedEvent]	Places
event:OperationDelaware	event:AShauValleyRepu...

Fig. 5 Result for Competency Question 3

Result 4: Find sub-event and belligerent involved in Battle of Saigon 1968 (see Fig. 6).

```
SELECT ?subEvent ?individual ?group
  WHERE {
    event:BattleofSaigon1968
    pne:subEventOf ?subEvent .
    event:BattleofSaigon1968
    event:involvedCountry ?individual .
    event:BattleofSaigon1968
    event:involvedGroup ?group .
    }
```

[subEvent]	individual	group
event:TetOffe...	pns:South_Vi...	pns:National_...

Fig. 6 Result for Competency Question 4

B. Discussion

In order to develop the ontology presented in this paper, the methodology outlined in 101 methods and METHONTOLOGY has been followed. According to METHONTOLOGY framework, our methodology was divided into three phases: Specification, Conceptualization and Implementation. This framework provided the idea of support activities: Knowledge Acquisition and Validation/Verification.

The most important task in the methodology is the definition of basic terms during the specification phase. All the knowledge acquired during the specification phase and it is the basis of conceptualization. This conceptualization has to be agreed on by domain experts.

Another important aspect to consider in developing ontology is validation process to check the accuracy and correctness of information. This provides for more abstract constraints as inferred knowledge from the ontology (e.g. subclass relations, transitive properties) is used to check whether the contents of a model are semantically correct or not. The required constraints can be specified as SPARQL queries.

IV. Conclusions

Building domain ontologies are difficult, particularly when the domain experts have a little background on knowledge engineering techniques and lack the skills of domain conceptualization. In this paper, the main conclusion that we showed was how domain ontology can be developed using the method proposed by Noy and McGuinness [16] and Uschold and Gruninger [18]. This approach was used to build the event ontology for historical documents. Both approaches had much guidance in defining the scope and identifying the basic terms for specification and conceptualization process for this new ontology. In the coming years, with increased development and availability of ontology, individual will take the challenge to develop ontologies especially domain expert in particular areas and make these ontologies available to the public. The contribution of this paper is the ontology development process of event ontology which was improved and expanded from SNAP ontology based on 101 method guide [16] and METHONTOLOGY [18].

Our future works include using the develop ontology for supporting semantic document retrieval of historical documents. In this case, concepts of the ontology will be mapped with the textual content of the Vietnam wars documents. We expect to achieve promising outcome whereby the historical documents can be retrieved based on the available events and complex queries can be supported.

References

[1] Brusa, G., M.L. Caliusco, and O. Chiotti. "Building Ontology in Public Administration: A Case Study", in SEBIZ. 2006. Citeseer.

[2] Smith, C.A., "Information retrieval in medicine: The electronic medical record as a new domain", in Proceedings of the American society for information science and technology, 2006. 43(1): p. 1-30.

[3] Staab, S. and R. Studer, Handbook on ontologies2013: Springer Science & Business Media.

[4] Ramli, R., S. Noah, and M. Yusof, Ontological-Based Model for Human Resource Decision Support System (HRDSS), in On the Move to Meaningful Internet Systems: OTM 2010 Workshops, R. Meersman, T. Dillon, and P. Herrero, Editors. 2010, Springer Berlin Heidelberg. p. 585-594.

[5] IJntema, W., et al., "Ontology-based news recommendation", in Proceedings of the 2010 EDBT/ICDT Workshops2010, ACM: Lausanne, Switzerland. p. 1-6.

[6] Noah, S., et al., Ontology-Driven Semantic Digital Library, in Information Retrieval Technology, P.-J. Cheng, et al., Editors. 2010, Springer Berlin Heidelberg. p. 141-150.

[7] Corda, I., "Ontology-based representation and reasoning about the history of science", 2007, The University of Leeds.

[8] Ide, N. and D. Woolner, Historical ontologies. Words and Intelligence II, 2007: p. 137-152.

[9] Cabo, M. and R. Llavori, A retrieval language for historical documents, in Database and Expert Systems Applications, G. Quirchmayr, E. Schweighofer, and T.M. Bench-Capon, Editors. 1998, Springer Berlin Heidelberg. p. 216-225.

[10] Elena, T., et al., "Historical research in archives: user methodology and supporting tools", International Journal on Digital Libraries, 2010. 11(1): p. 25-36.

[11] Meroño-Peñuela, A., et al., "Semantic technologies for historical research: A survey", Semantic Web, 2014. 6(6): p. 539-564.

[12] Salton, G., Automatic Information Organization and Retrieval1968: McGraw Hill Text.

[13] Katifori, A., et al., Effectiveness of Visualization for Information Retrieval through Ontologies with Entity Evolution. 2016.

[14] Fernández, M., et al., "Semantically enhanced Information Retrieval: An ontology-based approach", Web Semantics: Science, Services and Agents on the World Wide Web, 2011. 9(4): p. 434-452.

[15] Grüninger, M. and M.S. Fox, Methodology for the Design and Evaluation of Ontologies. 1995.
[16] Noy, N.F. and D.L. McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology. 2001.
[17] Gómez-Pérez, A., O. Corcho, and M. Fernandez-Lopez, Ontological engineering: with examples from the areas of knowledge management, e-Commerce and the Semantic Web. (advanced information and knowledge processing). 2004.
[18] Uschold, M. and M. Gruninger, "Ontologies: Principles, methods and applications", The knowledge engineering review, 1996. 11(02): p. 93-136.
[19] Brusa, G., M.L. Caliusco, and O. Chiotti. "A process for building a domain ontology: an experience in developing a government budgetary ontology", in Proceedings of the second Australasian workshop on Advances in ontologies-Volume 72. 2006. Australian Computer Society, Inc.
[20] Corcho, O., et al., "Building legal ontologies with METHONTOLOGY and WebODE", in Law and the semantic web2005, Springer. p. 142-157.
[21] Wache, H., et al. "Ontology-based integration of information a survey of existing approaches", in IJCAI-01 workshop: ontologies and information sharing. 2001. Citeseer.
[22] Noy, N.F., R.V. Guha, and M.A. Musen. "User ratings of ontologies: Who will rate the raters?" in AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors. 2005.
[23] Shaw, R., R. Troncy, and L. Hardman, "Lode: Linking open descriptions of events", The Semantic Web, 2009: p. 153-167.
[24] Tzompanaki, K. and M. Doerr, "Fundamental Categories and Relationships for intuitive querying CIDOC-CRM based repositories", 2012, ICS-FORTH/TR-429.
[25] Danzer, G.A., Windows to World History: Introducing World History. 1987.
[26] Bontas, E.P., M. Mochol, and R. Tolksdorf. "Case studies on ontology reuse", in Proceedings of the IKNOW05 International Conference on Knowledge Management. 2005.