Modeling, Extracting and Visualizing an Organization's Knowledge with Topic Maps

David RAMAMONJISOA

Faculty of Software and Information Sciences, Iwate Prefectural University
152-52 Sugo Takizawa Iwate 020-0193, Japan

and

Tomoyuki TAN

Faculty of Software and Information Sciences, Iwate Prefectural University
152-52 Sugo Takizawa Iwate 020-0193, Japan

ABSTRACT

This paper presents the results of our organization (university) modeling and knowledge extraction from the university homepages and a pamphlet. The ISO/IEC 13250 standard topic maps are used during the modeling. Tasks concerning topics, topic types, associations, roles and occurrences, definitions and extraction from the existing documents (in HTML or PDF file) were the primary focus. Although ontology and database schema exist in the university domain, the extraction of topics within documents is very difficult when applying those metadata to our specific university. We model the university from our understanding and methodology. The result is a metamodel, which is visualized with the tool. It can be concluded from this experiment that the student easily learns topic maps rather than other metadata modeling such as RDF or OWL. Constrained metamodel languages are easily manipulated rather than using unlimited online ontology.

Keywords: Semantic Web, Topic map, Knowledge Extraction and Visualization, university organization modeling

1. INTRODUCTION

Our university has a homepage and pamphlet maintained by the webmaster and faculty committee. These information resources are displayed as text or html. Students or academic staff use them daily, and have found some pages difficult to find without following the menu from the top page. We wanted to find an alternative way to access the information with the help of the existing metadata modeling and information visualization tools. We wanted also to extract knowledge from the web pages and pamphlet document files as autonomously as possible. With these goals in mind, we started manually building the metadata of the web pages and the pamphlet. We extracted several concepts or subject topics and their associations. Graphs were then produced to represent the documents. Stated in more technical terms, we built a knowledge base of the web pages and the pamphlet. The concepts and their associations formed the conceptual graphs or topic maps. After this basic sketch of the knowledge extraction, we decided to put them in a formal representation so that the computer could process the text data to produce graphs or maps. XML, RDF, and OWL were considered initially, but then we changed to XML Topic Maps (XTM) because XTM is an ISO/IEC standard. Moreover, XTM is easy to learn and use compared to the RDF or OWL language. After all, it was enough to help us model the required information resources.

This paper is organized as follows. Section 2 presents the problem and purpose of our research, while Section 3 describes the two semantic technology languages. Section 4 presents our modeling process. Finally, the results and scope for future research are discussed in Section 5.

2. PROBLEM AND PURPOSE

We aim to model the knowledge within our university web pages and pamphlet. Various kinds of information are included within those documents and accessing it is always a difficult task for students or academic staff. For example, some students want a list of the compulsory lectures during the first year, or a list of lectures in a specific course within a faculty or department during a year or semester. Students may also want to know about the research on a specific laboratory that he or she might utilize for a graduate thesis. Each lecture has a syllabus that is used during the class by the enrolled students. Each of these documents is in text format. For a mid-size university, those documents can have 100 or 200 pages. An alternative way to access the content of those documents quickly is the indexing of topics and the use of a local search engine service. This service already exists, but it does not provide a visual representation of how to access the data.

We developed graphical representations of the documents whereby content is accessed through the interaction of topics within the graphs or maps. The ultimate goal of our system is to build the maps autonomously and match any visual query to the knowledge and relevant documents or paragraphs.

The semantic web offers a solution to our problems. However, the technology itself is too complex for beginners. The semantic mapping of documents should not go beyond one layer or it becomes confusing. The metametamodel or higher orders of metamodel are not important for beginners. An illustration of a clear presentation of metadata and their real sources is depicted in Figure 1 below [8]. It concerns the knowledge representation of the book “Howards End” by E.M. Forster, who was a member of the Bloomsbury Group, along with Virginia Wolf; Margaret

Schlegel and Henry Wilcox are characters in the book who were played by Emma Thompson and Anthony Hopkins in the movie; Margaret Schlegel is married to Henry Wilcox in the story. At the bottom layer, there are resources in URIs (Universal Resource Identifier) difficult to remember and understand for human users. At the middle layer, there are topics and their associations, which are easily understood. At the top, the actual subjects are depicted by their photographic likenesses, which we perceive as the reality. These subjects are called *published subjects*, because they provide a mechanism whereby computers and humans can know when they are talking about the same thing. In other words, the subjects of discourse have *identities*. The bottom and the middle layers in Figure 1 contain respectively the *subject identifiers* and the *subject indicators*. Humans are familiar with the subject indicators to manage knowledge. They are the exact words in the documents encountered when we read or think. Subject identifiers are artificially built for a machine to process the documents and express as an address starting with the characters http://... or www…. The subject identifiers are *addressable subjects* and the subject indicators are *non-addressable* ones. These are the important distinctions in information and knowledge management that are taken into account during the course of our modeling and system implementation process.

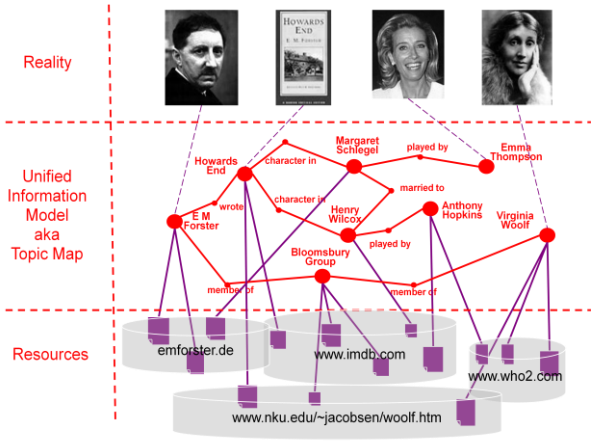


Figure 1: Knowledge representation of the book "Howards End".

3. TOPIC MAPS/TMCL OR RDF/OWL

We investigated different kinds of meta-modeling languages. We mean by models (including metamodels) the abstraction of the data which hides certain details while illuminating others things inside the data. The modeling is based on the semantic technology known as the Semantic Web. The Semantic Web languages are XML, RDF, OWL, Topic Maps, TMCL, Rules ML and so on. We focused on RDF/OWL and XTM. Figure 2 compares the two modeling languages. XTM is an ISO standard and RDF is a W3C standard. Topic maps have a query language called TMQL and RDF has SPARQL. Topic maps have an upper layer TMCL and RDF has RDFS and OWL.

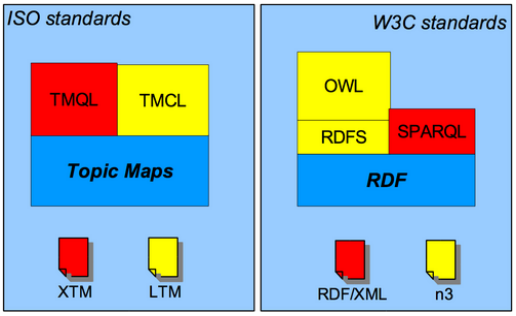
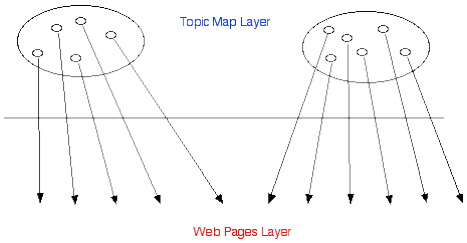


Figure 2: XTM and RDF layers comparison

3.1 Topic Maps

Topics are similar to the index terms, but are related to the knowledge in the documents rather than the list of all terms within documents. Figure 3 features two layers. At the top is the topic map layer and at the bottom is the resources layer. Topics have references to objects and concepts in the resources.



Topics have references to objects and concepts.

Figure 3: Topic map and documents layers

Topic maps are built from the **topics**, **topic types**, their **associations** and **roles**, and finally **occurrences [1]**.

Topic is anything whatsoever. That is, a subject or an idea.

Topic types are the classes of zero or more similar topics. Topic is an instance and topic type is a class. Example: Japan or Italy is an instance (topic) of the topic type "country". Student and teacher are instances of the "academic person" class.

Topics have three kinds of characteristics: names, occurrences, and roles in associations. Topic names include explicit names, base names, variants, display name, and sort name.

Association describes relationship between topics. Association type describes a relationship between topic types. Example: two topic types "city" and "country" have an association type: "is_capital_of" or "is in". Association roles: the role that each topic plays in the association. Association role is also a topic (in the sense that it is a type of the topic in an association).

Occurrence is a resource deemed relevant to the topic in some way. There are two kinds of Occurrence, the first is an *external occurrence* (document) to the topic map or URI used to pinpoint them and the second is an *internal occurrence*. The Occurrence type is the type of occurrence (example: book, article, or web page ...).

In our application described in Figure 4, our university entities are modeled as topics. Entity examples are faculties and the university itself. The university-faculty-rel association is the relation between the university and the faculty. It has an association name

"is_an_academic_organization_within". University is a type of organization and faculty is a type of faculty. Topics are the instances of those topic types identified with their respective names (example, "Software and Info science", "nursing", etc). Occurrences are the resources such as the homepage (URI) of each topic.

Scope specifies the extent of validity for a topic characteristic (base name, occurrence, or association). For example, we can scope the topic map to a certain language like Japanese only or English only.

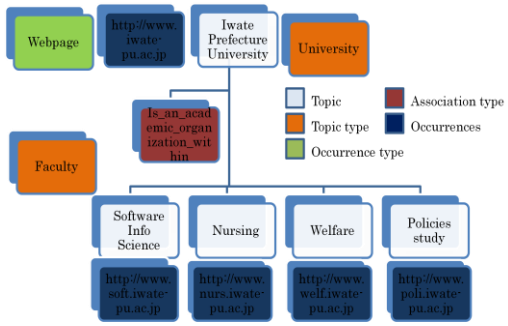


Figure 4: Part of the topic map model of our university

The details of the modeling process are described in the next section.

3.2 RDF and OWL

The **Resource Description Framework (RDF)** is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model [9][10].

The RDF data model is similar to classic conceptual modeling approaches such as Entity-Relationship or Class diagrams, as it is based upon the idea of making **statements** about resources (in particular Web resources: aka subject identifiers) in the form of subject-predicate-object expressions. These expressions are known as *triples* in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

A collection of RDF statements intrinsically represents a labeled, directed multi-graph. An example of an RDF statement is as follows.

http://www.example.org/index.html has a creator
http://www.example.org/staffid/85740

Figure 5: an RDF statement graph

The predicate "has a creator" is converted into URI http://purl.org/dc/elements/1.1/creator. It is clear that this form of knowledge representation is difficult for beginners to understand.

We can also write in the opposite direction as the below statement shows.

http://www.example.org/staffid/85740 has created
http://www.example.org/index.html

This is not a conventional RDF triple because the subject is not the resource as we mean, but syntactically and semantically, it is a correct statement. An RDF graph is isomorphic so we can write anything about anything without constraint. The confusion arises when we are trying to find the URIs of these "anything". We have to realize that all the elements in the RDF triple are resources. A resource must have an URI. The nodes and link of the graph in Figure 5 are made of resource URIs.

3.3 Comparison

In comparison to the topic maps, any resource in RDF can be considered a topic in Topic maps. The reference of the real object or abstract thing in the world is the topic and the resource URI in RDF. As we know, metadata is also data. Hence, it is difficult for beginners to grasp.

In a higher level of meaning, the occurrence in the topic map is equivalent to the resource in RDF. RDF graphs and Topic maps are similar when each uses the subject indicators to represent the nodes and edges. Associations in a topic map are similar to predicates in RDF in certain contexts. Association in a topic map can connect more than two topics, which is not possible in RDF. The clarity of the visual graph obtained from both languages was also compared. We realize that RDF/OWL is designed for machine only and topic map is for human and machine. Lars Garshol explains in a more in-depth fashion the similarities and differences in his web pages [3].

The ontology of topic maps consists of topic type, topic name, occurrence type, association type, and association role type. The strength of Topic Maps is the opportunity to define arbitrary those types as ontology [6] and to define their constraints using a standardized Topic Maps Constraint Language (TMCL) [2].

TMCL and OWL allow the user to model the restrictions between two roles in the association type (these are called domain and range of a property in OWL).

4. MODELING PROCESS

We have followed the guidelines for authoring Topic maps suggested by Anita Altenburger [4] to model our university. There are three steps:

- Step 1: Define the *theme* that should be covered.
- Step 2: Collect as many *topics* as possible which are relevant for the theme, together with other external information resources, such as web sites (so-called *occurrences*).
- Step 3: And finally, consider the relationships between the collected topics (so-called *associations*).

Steve Pepper and Lars Garshols refine this process [6]. Following the third step, they propose to populate the ontology by discovering the topics, associations, and occurrences from the data.

Then finally, an application that uses the resulting topic map can be developed to share the knowledge.

The result of this process is the list described in the following table (table 1). This part of the model is obtained after several trials and errors to retain only the essential topics. Instances are any individual terms related to these types. There are no guidelines to define roles in the topic maps. Some topics have natural roles, such as how a person topic has a role professor or associate professor. However, topic such as faculty or department cannot have any role other than their meaning. This is redundant but useful for the clarity of the model. The topic map visualization allows us to justify this necessity of adding role to each topic type. This is also valid for the association types. The model is enriched when those entities are there, as can be seen in Figure 6.

Table 1: List of topics, associations, roles types

Topic Types	Association Types	Role Types
Organization	Organization-Faculty-Rel	University
Faculty	Faculty-Department-Rel	Faculty
Department	Department-course-Rel	Department
Course	Course-Lecture-Rel	Course
Lecture	Person-Lecture-Rel	Lecture or Course Lecture
Laboratory	Course-Laboratory-Rel	Laboratory
Person	Laboratory-Person-Rel	Professor,Associate Professor, Lecturer, Associate Researcher or Teacher Responsible
Syllabus	Lecture-Syllabus-Rel	Syllabus
Research_Area	Person-Research_Area-Rel	Research_Area

We populate the instances of the topic maps with the information available on the documents (web pages and pamphlet). Instances are subject names that can be found on the document as university name, faculty names, course names, lecture names, person names, and so forth. We remember that those instances are also topics, so they can have associations to each other according to their topic type. Organization of the board committees (president, vice-presidents, faculty deans, department deans, and office managers) is then added. Implicit knowledge is not obtained from this stage. For example, a university and department association is implicit because a department is a subset of a faculty and a faculty is a subset of a university. A method to implement this implicit knowledge is by using the description logic and a reasoner. Some common knowledge about the university such as address, phone number, office building, room number, library, sports facilities, and admission office information, are omitted because they are irrelevant to our model. A simple XML database repository would instead suffice to represent them.

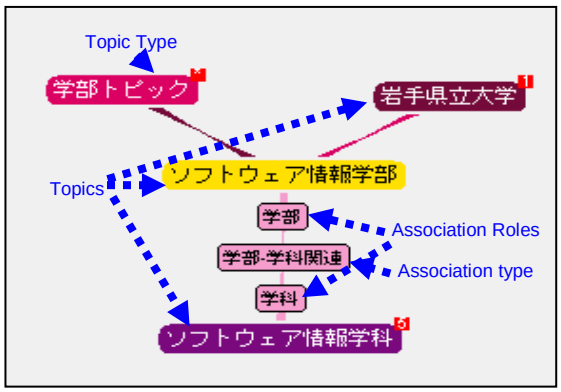


Figure 6: Example of topic map model visualization

We then modeled our faculty's internal organization. Our faculty has four courses and one department. Each course has four to six laboratories. Each laboratory has members composed of teaching staff and some students. Each member of the teaching staff delivers lectures and engages in research activity. Each lecture has a syllabus. A syllabus contains many topics about the lecture such as data types (affiliation, title, objective, description, course Website) and object types (assignment, resource, course code, teaching staff, grading, specific schedule, prerequisites, textbook, exam, general schedule). The syllabi databases are stored in PDF files. The template for the syllabi can be easily converted into a topic map by extracting all of the attributes. We took the algorithm and data structures from the work of X. Yu et al. on syllabi automatic conversion into metadata [5]. This meant using the taxonomy of the syllabi and then applying a regular expression search to extract the name entity from syllabi and populating the RDF databases from selected syllabus documents. Syllabi are written in several formats so they also formed a system of classification based on SVM and Naïve Bayes.

Our model is not made in exactly the same way. Our syllabi database is in Japanese and PDF format. The NLP system is different. We built a new taxonomy adapted to our data. Some attribute-value pairs are <topic types, instances> pairs, and others are <occurrence types, texts or data> pairs. For examples, title and teacher(s) are topics defined in the previous topic map, so they are topic types. They are linked with the association “Person-Lecture-Rel” (example in Figure 8 (c)). Credit is an integer data type. The lecture period has a value composed of (1S: 1st Year Spring, 1F: 1st Year Fall, and so on), and students from those years can only attend the class according to the curriculum rules.

An example of a syllabus and an XML data model features in Figure 7:

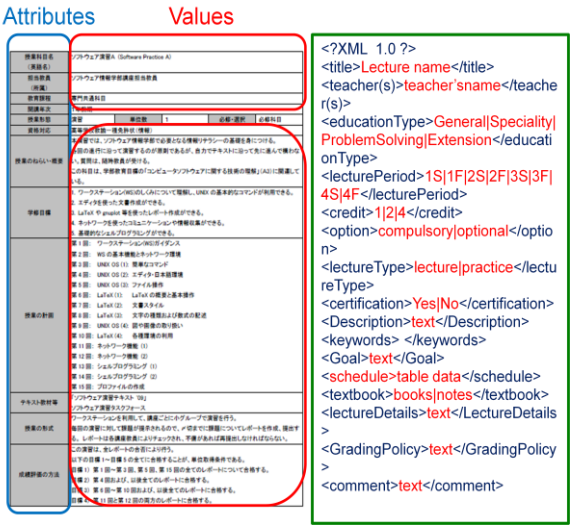


Figure 7: Example of syllabus and XML formatted document

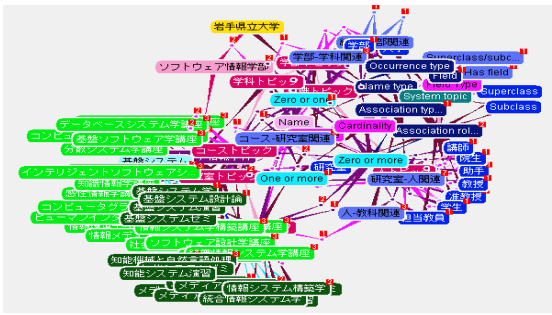
The syllabus topic map is then merged with the faculty and university topic maps. This merging allows us to link the teachers in the syllabus with the faculty person in the organization topic map and the lectures with the syllabus lecture titles.

It is clear that the extracting knowledge while reading the pamphlet or web pages is difficult. It is useful for the student to navigate the topic map of this data and discover some interesting graphs. For example, gathering the topics with credits greater than 4 during the Fall semester for sophomores will show all the syllabus responding to the criteria but also the topic's title, description option (compulsory or optional). Such requests or queries can also be expressed with the TMQL or SPARQL against the knowledge base. Figure 8 (b) shows the compulsory lectures (blue or dark topics) that students in this course must take in order to graduate.

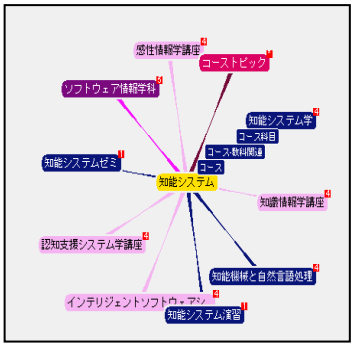
5. TOPIC MAPS NAVIGATION, VISUALIZATION AND QUERYING

By using the Ontopia's tool Vizigator part of the Ontopia Knowledge Suite (OKS) [8], we provide an interface to access the university through the topic maps. Vizigator transforms the topic maps in XML format into a visual graph by focusing only on the selected topic (see figure 8 (a), (b)). The selected topic is in the center and colored yellow. Linked topics have different colors. Associations and roles are represented with the same color. Another tool is the Ontopia's VizDesktop.

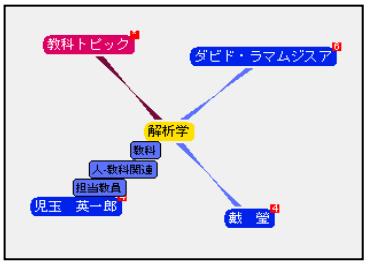
These figures lead us to previously untapped, interesting knowledge within our university. We concur with the authors of the book "beautiful visualization" [11] who argue, "beautiful visualization reflects the qualities of the data that they represent, explicitly revealing properties and relationships inherent and implicit in the source data." Tag clouds such as Wordle [12] are widely used for the text analysis and information visualization on the Web or academic institutions. Unfortunately, up to now, they are limited to the alphabet characters that we could not use for our needs. Concept maps relate research to topic maps that uses a graphical representation of relationships among concepts [13]. Digital interactive concept maps (CMaps) help students navigate complex knowledge domains, such as the content of a course or a curriculum [14]. The authors find value in visual navigation structures in regard to the organization and simplification of learning environments, primarily by appealing to visually oriented learners.



(a) Selected topic map of the university



(b) Course topic map including laboratories in pink (light color) and compulsory lectures for this course in blue (dark color)



(c) Lecture and the teacher(s) topic map

Figure 8: topic maps view.

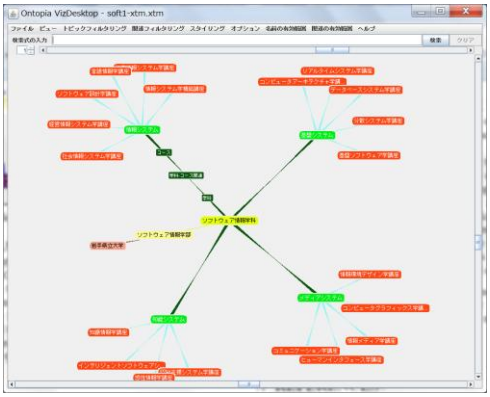


Figure 9: VizDesktop's snapshot includes the software faculty and its 4 courses and 20 laboratories in topic map.

Besides the visualization of those topics, queries can also be asked to match parts of the knowledge or topics with high accuracy by using the Topic Map Query Language (TMQL) similar to SPARQL or SQL syntax [15]. An example of such a query is explained at the end of Section 4.

6. CONCLUSION AND PERPECTIVES

In this paper, we presented the topic maps of our university built with knowledge management tools such as semantic technology modeling based on metadata extraction and organization. The main goals are (1) to visualize the topics within our organization and (2) access those topics in an efficient manner and (3) evaluate the semantic technology learning from beginners who are not native English speakers, and not using the alphabet when writing characters. We achieved those goals and we realized that learning semantic technology is better when metadata are using limited ontology (few topic types or classes without deep hierarchy structure by using the ***is_a*** relationship). At first, beginners need to focus on the modeling of the essential and necessary entities and their relationships according to the data rather than learning the language syntax or finding ways to fit them into the built-in vocabularies RDFS and predefined OWL classes. In the next step, an international version of the knowledge base will be developed by using or creating the published subject identifiers (PSIs) on the shared topic maps repository [16] [17]. For intermediate or advanced users, an ontology modeling with TMCL or OWL2 [18] is recommended. The merging, alignment, and mapping of multilingual topics or multiples ontology also remain a challenge.

The differences and similarities between Topic maps and RDF/OWL are subject to much debate. The topic map developers strongly emphasize human cognitions or uses, while RDF/OWL developers focus on intelligent agent platforms and Linked Open Data [19]. Given that they can coexist and will not merge, efforts have being made to ameliorate the Topic maps system and RDF vocabularies interoperability [7] for future reuse.

Knowledge modeling (extraction and visualization) is part of the learning mechanisms as it offers a metacognitive tool to map data and share models to a community by discovering new patterns hidden in the data by the modelers. By sharing models, modelers will obtain feedback and perform self-assessments. Work in progress seeks to automate the modeling that computers can learn and enable them to discover patterns.

In the future, the potential for an automatic conversion of the restrictions and statements in the data into metadata will be investigated. That is to say, a faculty with four courses and twenty laboratories referred to in the document should be autonomously described into OWL class restrictions or an association type cardinal setting in the topic map.

7. REFERENCES

[1] **ISO/IEC 13250 standard**, available, last accessed 07/12/2010 at http://www.isotopicmaps.org/TMRM/TMRM-7.0/tmrm7.pdf, 2007.

[2] **TM Constraint Language draft**, available, last accessed 07/12/2010 at http://www.isotopicmaps.org/tmcl/2010-03-25/, March, 2010.

[3] Lars Marius Garshol, **Comparing Topic Maps and RDF**, last accessed 07/12/2010 at http://www.garshol.priv.no/blog/92.html,

[4] A. Altenburger, **Authoring XTM Topic Maps, Part I**, http://topicmaps.it.bond.edu.au/docs/6?style=printable, 2000

[5] Xiaoyan Yu et al., **Using automatic metadata extraction to build a structured syllabus repository**, in Proceedings of the ICADL'07 (10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers) pp.337-346. LNCS publisher, 2007.

[6] Steve Pepper and Lars Grashol, **Lessons on applying Topic Maps**, last accessed 07/12/2010 at http://www.ontopia.net/topicmaps/materials/xmlconf.html, 2004.

[7] Steve Pepper, **Expressing Dublin Core in Topic Maps**, Lecture Notes in Computer Science, Volume 4999/2008, pp.186-197, 2008.

[8] Pamela Gennusa, **Now You See It!, Ontopia's Vizigator**, http://www.ontopia.net

[9] Frank Manola and Eric Miller, **RDF Primer**, http://www.w3.org/TR/rdf-primer/

[10] Graham Klyne and Jeremy Carroll, **Resource Description Framework (RDF): Concepts and Abstract Syntax**, http://www.w3.org/TR/rdf-concepts/

[11] Noah Iliinsky, **On Beauty**, chapter 1, in Beautiful Visualization, Julie Steele and Noah Iliinsky (Editors), O'Reilly Media Inc publisher, June 2010.

[12] Jonathan Feinberg, **Wordle**, chapter 3, in Beautiful Visualization, Julie Steele and Noah Iliinsky (Editors), O'Reilly Media Inc publisher, June 2010.

[13] **CMapTools**: Knowledge Modeling Kit by Institute for Human Machine Cognition, last accessed 07/12/2010 at http://cmap.ihmc.us.

[14] Miertschin, S., Willis, C., **Using Concept Maps to Navigate Complex Learning Environments**, in Proceeding of SIGITE '07 Proceedings of the 8th ACM SIGITE conference on Information technology education, pp. 175-183, 2007.

[15] **TM Query Language draft**, available, last accessed 07/12/2010 at http://www.isotopicmaps.org/tmql/

[16] Dmitry Bogachev and Steve Pepper, **Subject-centric computing**, last accessed 07/12/2010 at http://www.ontopedia.net/

[17] Steve Pepper, **Topic Maps**, Encyclopedia of Library and Information Sciences, Third Edition, Taylor and Francis (Publishers), DOI:10.108, 2010.

[18] **OWL 2 Web Ontology Language Primer**, available, last accessed http://www.w3.org/TR/2009/REC-owl2-primer-20091027/

[19] Michael Hausenblas, **Exploiting Linked Data to Build Web Applications**, in IEEE Internet Computing Magazine, pp.68-73, July 2009.