6S: Adding a Semantic Model to 5S Framework

Heba Mahmoud Neama
Mathematics and Computer Science,
faculty of Science Alexandria University
Alexandria, Egypt
Email: heba_noima [AT] yahoo.com

Yasser. F. Hassan
Mathematics and Computer Science,
faculty of Science Alexandria University
Alexandria, Egypt

Mohamed Kholef
Computing and Information
Technology, Arab Academy for Science
and Technology.
Alexandria, Egypt

Abstract—**in this paper we proposed new model for digital library which is an extension for the 5S Model to include the semantic web layer in the structure of digital library to be 6S model for digital library. We also discuss the important role of semantic web in digital library and how are semantic web technologies affect the information retrieval accuracy. We represent the semantic layer in 6S Model by adding ontology to a digital library that satisfies the 5S model and enhance ontology by updating it automatically. This ontology is used in books classification and retrieval according to concepts in ontology.**

Keywords; semantic web; digital library; ontology; Hierarchical classification; Naive Bayes classifier

I. INTRODUCTION

Digital Libraries are systems specifically designed to assist users in information seeking activities. As a result, libraries face new challenges, competitors, demands, and expectations [1]. Libraries are redesigning services and information products to add value to their services and to satisfy the changing information needs of the user community. Traditional libraries are still handling largely printed materials that are expensive and bulky. Information seekers are no longer satisfied with only printed materials. They want to supplement the printed information with more dynamic electronic resources [2]. In Section 2 there is brief description about semantic web. In section 3 there is brief definition about 5S Model of digital library. In section 4 we introduce the 6S proposed model for digital library. Section 5 is the ontology implementation and update ontology method. Section 6 is case study of applying algorithm to update ontology in digital library and Finally Section 7 is the Conclusion with future work.

II. SEMANTIC WEB

Semantic Web is a technology which adds well-defined documents on the Web for computers as well as people to understand the meaning of the documents more easily, and to automate the works such as information searches, interpretation, and integration. The ontologies, which are an essential component of the semantic Web, define the common words and concepts used to describe and represent an area of knowledge [3].

A semantic information search based on the ontology can provide the inferred and associated information between data [3]. The use of ontologies in the context of digital libraries could be interesting in order to incorporate new functionalities by describing the relationships between elements. The concept of ontology introduced by the Semantic Web is a promising path to extend digital library formalisms with meaningful annotations [4].

The Semantic Web Stack, also known as Semantic Web Cake or Semantic Web Layer Cake, illustrates the architecture of the Semantic Web.

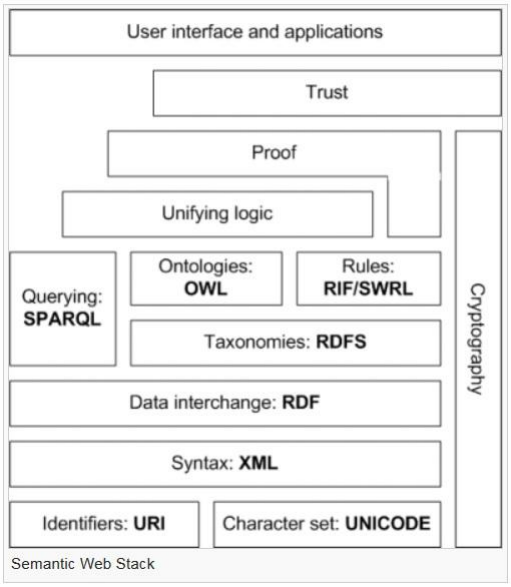

Figure 1. Semantic web stack

Each layer exploits and uses capabilities of the layers below. It shows how technologies that are standardized for Semantic Web are organized to make the Semantic Web possible [5].

III. 5S FORMAL MODEL FOR DIGITAL LIBRARY

5S provides a formal framework to capture the complexities of digital libraries. The definitions in [6] unambiguously specify many key characteristics and behaviors of digital libraries. This also enables automatic mapping from 5S constructs to actual implementations as well as the study of qualitative properties of these constructs (e.g., completeness, consistency). In this section, we summarize the 5S theory from

[6]. Here we take a minimalist approach, i.e., we describe briefly, according to our analysis, the minimum set of concepts required for a system to be considered a digital library [7].

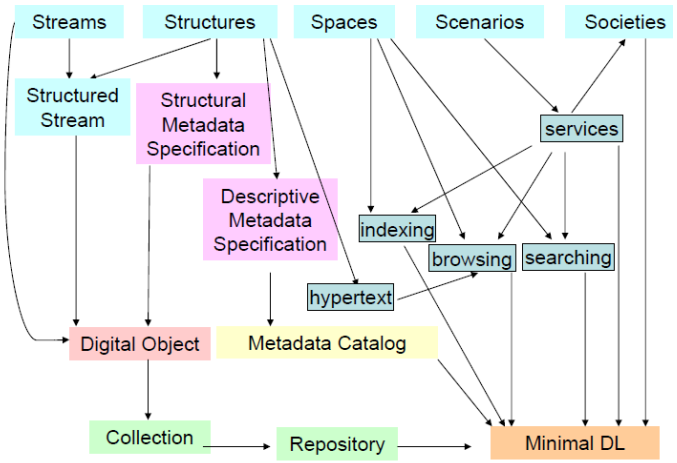

Figure 2. Digital library in 5S Framework

- Streams are sequences of arbitrary types (e.g., bits, characters, pixels, frames) and may be static or dynamic (such as audio and video). Streams describe properties of digital library content such as encoding and language for textual material or particular forms of multimedia data.
- Structure specifies the way in which parts of a whole are arranged or organized. In DLs, structures can represent hypertexts, taxonomies, system connections, user relationships, and containment– to cite a few.
- Space is a set of objects together with operations on those objects that obey certain constraints. Spaces define logical and presentational views of several DL components, and can be of type measurable, measure, probability, topological, metric, or vector space.
- Scenario is a sequence of events that also can have a number of parameters. Events represent changes in computational states; parameters represent specific variables defining a state and their respective values. Scenarios detail the behavior of DL services.
- Society is "a set of entities and the relationships between them" and can include both human users of a system as well as automatic software entities which have a certain role in system operation. These 5Ss, along with fundamental set theoretic definitions, are used to define other DL constructs such as digital objects, metadata specification, collection, repository, and services [7].

IV. 6S Proposed Model for Digital Library Streams, Structures, Spaces, Scenarios, Societies and Semantics

Due to the important role of semantic web technologies in enhancing the digital library functionalities we propose new model which is an extension of the 5S formal model of digital library to include semantic web technologies.

In figure 3 Semantic web technologies represents important layer where 6S layers depends on semantic to gain meaningful data from digital library while searching, browsing and indexing. In the following section we will describe the 6th S in the 6S formal model.

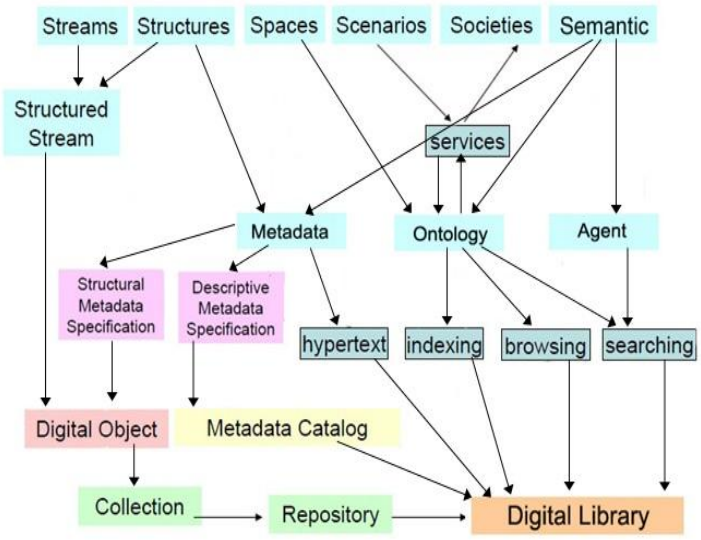

Figure 3. Digital library in 6S Framework

6-Semantic: is the semantic layer in the 6S proposed model for digital library. The semantic layer provides the components concerned with the creation, enhancement, maintenance, mediation, and querying of ontological information that is linked to the data stored in digital library.

For example, semantic search is used in searching in digital library. Ontology is used to define the concepts of digital library and the relationship between them. This increases the quality of data stored in the digital library and results in enhancing the Digital library functionality of classification, browsing and information retrieval.

TABLE I. 6S MODEL EXAMPLES AND OBJECTIVES

Ss	Examples	Objectives
Streams	Text; video; audio; image	Describes properties of the digital library content such as encoding and language for textual material or particular forms of multimedia data
Structures	Collection; catalog; hypertext; document; metadata	Specifies organizational aspects of the digital library content(e.g., structured stream or protocol), profiles, logs, services
Spaces	Measure, measurable topological, vector, probabilistic	Defines logical and presentational views of several digital library components; host and user locations; GIS
Scenarios	Searching, Browsing, recommending	Details the behavior of digital library services, workflows, life cycle, preservation
Societies	Service mangers, learner, teacher, etc.	Defines managers, responsible for running digital library services; and relationships among them
Semantic	Semantic search, ontology based navigation and classification, components concerned with the enhancement, maintenance, mediation, and querying of ontological information	Defines the different semantic web technologies components on which the digital library depends to make enhancements such as semantic search, explicit metadata, ontology based information retrieval, navigation and classification, logic and inference which help to uncover unexpected relationships, enhancement in usability and interoperability.

A. *Why 6th S: Semantic layer in the 6S model?*

5S Model doesn't represent explicitly the semantic web technologies in digital library despite of their important role. Many digital libraries are using semantic web technologies in their creation and implementation of their functionalities like BRICKS and DELOS digital library. The Semantic Web enables data interoperability, allowing data to be shared and reused across heterogeneous applications and communities [13]. The Semantic Web is mainly based on the Resource Description Framework (RDF) by which is possible to define relations among different data, creating semantic networks [14] and integrate information based on different metadata, e.g.: resources, user profiles, bookmarks, taxonomies, high quality semantics, highly and meaningfully connected information provide interoperability with other systems (not only digital libraries) on either metadata or communication level or both, RDF as common Delivering more robust, user friendly and adaptable search and browsing interfaces empowered by semantics (legacy, formal, and social annotations)

V. ONTOLOGY IMPLEMENTATION

Here the Subject ontology in Figure 4 created in digital library is arranged in a tree hierarchy, at each level of the tree there are classes represent subjects of data points, to classify data point as a member of existing class we will start with the root node. Every data point belongs to the root.

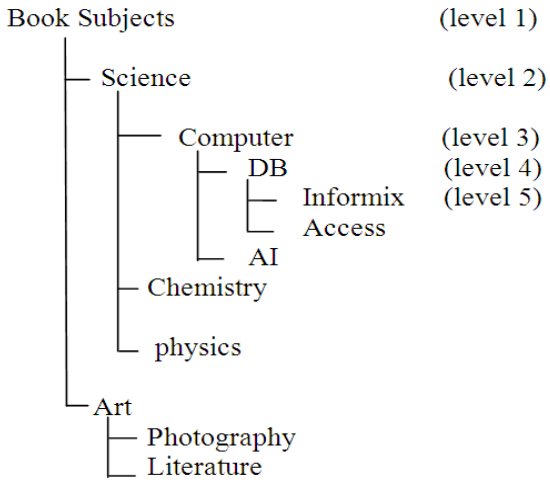

Figure 4. Subject ontology

A. *Example of Ontology Constraints*

Examples of class constraints are suggested by ontology as shown in figure 5 are as follows:

- The Subclass-Superclass constraint: if a data point is member of "Database", then it should also be member of "Computer".
- The "Mutual Exclusion" constraint states that: if a data point belongs to "Computer" class, then it should not belong to "Physics" class

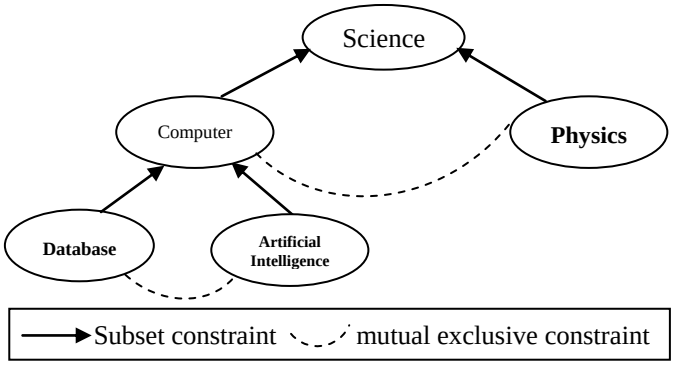

Figure 5. Constraints Model

B. *Algorithm used for updating Ontology*

The ontology in figure 4 can be updated using the hierarchical algorithm, which provides the ability to classify any data point to existing class or create new class and update the constraints according to the new added class.

We used the naïve bayes classifier formula to determine the probability distribution of data point over all classes, Naive bayes is able to compare not only single words, as in most current approaches, but substrings of an arbitrary length.

C. Naïve Bayes

Your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Naïve Bayes classifier is useful in our case study of the tree classifier. We begin with a set of training examples with each example document assigned to one of a fixed set of possible classes, C = {C1, C2, C3,... CJ}. Naïve Bayes classifier uses this training data to generate a probabilistic model of each class; and then, given a new document to classify, it uses the class models and Bayes' rule to estimate the likelihood with which each class generated the new document. The document is then assigned to the most likely classes [10].

Hence, given a document d= {d1, d2, d3,...,dL}, we use Bayes theorem to estimate the probability of a class Cj:

$$P(C_j \mid d) = \frac{P(Cj)P(d \mid Cj)}{P(d)} \quad (1)$$

To combine multiple pieces of evidence it is that, if two different key words w1 and w2 the probability calculation will start with the following equation [10]:

$$P(C_1|W_1 \wedge W_2) = \frac{P(C1)\ P(W1 \wedge W2|C1)}{P\ (W1 \wedge W2)} \quad (2)$$

When two features are conditionally independent, we can calculate their co-occurrence as a simple multiplication [11]

$$P(C_1|W_1 \wedge W_2) = \frac{P(W1|C1)P(W2|C1)\ P(C1)}{P(W1 \wedge W2)} \quad (3)$$

Russell and Norvig explain that, we can eliminate the term P(W1∧W2) with normalization, which uses the conditional probabilities and the assumption of conditional independence to calculate this term[11].

$$P(C_1|W_1 \wedge W_2) = \frac{P(W1 \wedge W2|C1)\ P(C1)}{P\ (W1 \wedge W2)} \quad (4)$$

$$P(C_2|W_1 \wedge W_2) = \frac{P(W1 \wedge W2|C2)\ P(C2)}{P\ (W1 \wedge W2)} \quad (5)$$

The two equations sum to 1, since the word W1 is certainly either related to C1 or C2 and then multiplies the whole equation by the common denominator and the resultant equation is

$$P(C_1|W_1 \wedge W_2) = \frac{P(W1|C1)\ P(W2|C1)\ P(C1)}{P(W1|C1)\ P(W2|C1)\ P(C1)+P(W1|C2)P(W2|C2)\ P(C2)} \quad (6)$$

D. Hierarchical algorithm

Based on the ontology in figure 4, given a non-classified data point d and a set of classified data points D associated to classes C, each class has set of constraints consi, It is required to classify d in certain class whether by adding class or creating new class with its constrains $cons_{i+n}$

Function Update_Ont_Algorithm ($D^{\cdot}$,C, d, $Cons_i$):C_d, $Cons_{i+m}$,
Input:
D set of labeled data points,
C set of classes,
d unclassified data point,
$Cons_i$ set of constraints;
Output:
C_d class label of d,
$Cons_{i+m}$ constraints on new class;

P_{new} probability of creating a new class
h = Number of levels of ontology

for J=1 to h do
for K=1 to $|C_j|$ do
($|C_j|$ is the number of classes in one level J)
Using Naive Bayes classifiers Find $P(C_{jk}|d)$

$$P(C_{jk}|d) = \frac{P(W_1|C_{jk})\ P(W_2|C_{jk})P(W_n|C_{jk})\ P(C_{jk})}{P(W_1|C_{jk})P(W_2|C_{jk})P(W_n|C_{jk})P(C_{jk}) + P(W_1|C_{j(k+1)})\ P(W_2|C_{j(k+1)})P(W_n|C_{j(k+1)})P(C_{j(k+1)})}$$

$$P(C_{i(k+1)}|d) = \frac{P(W_1|C_{j(k+1)})\ P(W_2|C_{j(k+1)})P(W_n|C_{j(k+1)})\ P(C_{j(k+1)})}{P(W_1|C_{jk})P(W_2|C_{jk})P(W_n|C_{jk})P(C_{jk}) + P(W_1|C_{j(k+1)})\ P(W_2|C_{j(k+1)})P(W_n|C_{j(k+1)})P(C_{j(k+1)})}$$

C_d = **DataPointClassify**($P(C_{jk}|d)$,$P(C_{j(k+1)}|d)$, h)
$Cons_i$ = **Update_Cons**({d∪D},{C_d∪C},$Cons_i$)
end for
$Cons_{i+m}$= $Cons_i$
end for
end function

function DataPointClassify (P1,P2, h):C_d
Input:
P1=$P(C_{jk}|d)$: probability of first class given a datapoint d,
P2=$P(C_{j(k+1)}|d)$: probability of second class given a datapoint d
h: heigth of ontology
Output:
C_d classification of d to certain class

for L = 2 to h **do**
if d has seed label at level L **then**
class(d, levelL) = seed label;

else
Classofdatapoint = children(label(d, $level_{l-1}$))
if Classofdatapoint is not empty **then**
if (maxProb(P1,P2) / minProb(P1,P2)) < 2
then Assign d to C_{dnew}
Set parent(C_{dnew}) = root class at level L-1
Else
C_d=Class(maxProb(P1,P2))
Assign d to maxProb(P1,P2)
end if
end if
end if
end for
end function

Function Update_Cons (C, D,$Cons^{old}$):$Cons^{new}$
Input:
D: Datapoints;
C: Class assignments all datapoints in D;
$Cons^{old}$: Old constraints on the existing set of classes.
Output:
$Cons^{new}$: Updated set of class constraints,
$Cons^{new} = Cons^{old} + C_{dNewCons}$;

Each new class created is added to a single parent at the time of creation.
Add each parent, child relationship as a constraint in $Cons^{new}$.
end function

VI. . CASE STUDY

We created Subject Ontology of 5 levels (figure 4) hierarchy classes and super classes to associate subject of documents to concept in ontology. In this way user can view subjects organized in a classification scheme and can browse over this scheme to retrieve documents. This digital library is an open source Java project which is helping user to organize and retrieve documents in PDF format. It is software of Personal Information Management (PIM) that works with technologies of Semantic Web [12].
User can insert/edit information on documents like author, title, description, subject and so on. This information is stored as RDF (Resource Description Framework) and use standard properties like those defined in Dublin Core metadata set [12].

A. Digital library Stored Data

These are samples of data stored in digital library beside the other metadata of books.

TABLE II. DIGITAL LIBRARY SAMPLE DATA

Labeled Data points (D)	Classes (C)	Unlabeled Data Points (d)	Classes (C_d)
Book1: Introduction to Microsoft Access 2003 Book2: Microsoft-Access Tutorial	Access	Book11:Oracle Automatic Storage Management Administrator's Guide Book12:Oracle Database Administrator's Guide Book13:Oracle Database Backup and Recovery User's Guide	Oracle
Book6: INFORMIX-4GL Reference Manual Book7: IBM Informix Administrator's Guide	Informix		

The experiments were performed with a collection of 200 documents that includes master thesis, conference papers and books from the computer science domain. The 200 document are divided into 55 documents related to access subclass, 65 books related to Informix subclass and 80 books of oracle subject. In our case study we will compare between keyword based searching, ontology based searching and searching after updating ontology.

To evaluate the information retrieval systems precision and recall were used to quantitatively measure the performance of information retrieval. Precision is the ratio of relevant retrieved documents to the number of retrieved documents and recall is the ratio of relevant retrieved documents to the all relevant documents.

All documents in digital library are members of the root class, for example books which are members of computer and physics and chemistry classes are also members of the science class. Class constraints are mutually exclusive in the same level of hierarchal ontology, for example books which are members of Informix subclass must not be member of any other subclass like oracle and access subclasses.

A. Retrieving books Using Keyword Search

Documents in the form of PDF are stored into digital library and their metadata extracted once they are uploaded. Metadata like document title, author, keywords and description can be used in the keyword search. In the following four tasks study of retrieve all documents of certain subject in digital library using keyword search.

- Task 1. Find documents about Oracle.
- Task 2. Find documents about access.
- Task 3. Find documents about Informix.
- Task 4. Find documents about physics.

To evaluate the searching results by calculating the precision and recall for each task as following:
A=No of relevant records retrieved.
B=No of relevant records not retrieved.
C=No of irrelevant records retrieved.

Precision= A/A+C
Recall= A/A+B
For Task 1 the precision and recall for keyword search based are computed as following: Precision=63/63+10 = 86.3% and Recall= 63/63+17= 78.7%
In Task 2: 45 documents were retrieved from 55 documents of Access subject. The precision and recall for keyword search based are computed as following: Precision=42/42+3 = 93.3% and Recall= 42/42+13= 76.4%

B. Retrieving Documents Using Ontology based Search

Here the previous 4 tasks will be performed using Ontology created about documents subjects. This ontology based searching is done by querying the RDF of documents to list all documents of subject equal to searching text as following:
Task1: 74 documents were retrieved when searching for oracle subject using ontology based searching. The precision and recall For Task1 will be computed as following:
Precision=65/65+9 = 87.8% and Recall= 65/65+15= 81.25%
The results of the case study performed on the system indicate improvements in precision and recall as shown in Table 5 the values of precision and recall for the 4 tasks. From this comparison it's clear that values are enhanced when using ontology-based search techniques as Ontology use RDF schema for mapping the meaning of each word searching for in digital library.

TABLE III. RESULTS OF SEARCHING USING KEYWORDS AND ONTOLOGY:

	Keyword Search		Ontology-based Search	
Tasks	***Precision***	***Recall***	***Precision***	***Recall***
Task 1	86.3%	78.75%	87.2%	81.25%
Task 2	93.3%	76.4%	94%	85.4%
Task 3	96.6%	89.2%	96.8%	91.3%
Task 4	88.8%	80%	93.3%	84%

C. Searching after updating ontology:

We need to retrieve book of certain subject and this subject is not exist in ontology so we will try to classify books to certain class in ontology or create new class according to book subject. In the following section we will apply algorithm to update ontology.

D. Applying Hierarchical Algorithm

Starting from level 4 in subject ontology figure 4 assuming that Book7 from table3 is related to database super-class, we need to classify this unlabeled data point to one of the children which are (access and Informix) or creating new class in ontology for it.
Book7: Oracle Automatic Storage Management Administrator's Guide P(ACC)=46%
P(INF)=54%

TABLE IV. PROBABILITY DISTRIBUTION OVER CLASSES

Word	N1	N2	W/Acc	W/Inf
Oracle	20	25	0.36	0.39
Automatic	18	30	0.33	0.46
Storage	15	14	0.27	0.22
Management	27	20	0.49	0.3
Administrators	20	18	0.36	0.28
Guide	30	35	0.55	0.54

N1: number of books where the word exists in first class (access)
N2: number of books where the word exists in second class (Informix).
W/Access: number of access books where the word exists divided by the number of all books in access
W/Informix: number of Informix books where the word exists divided by the number of all Informix books.
7: Find P(Ci|X) for all classes (Access and Informix)
P(Access|Xu)= 0.6
P(Informix|Xu)=0.4
From the function of Consistent Assignment and since the Pcand of each class is nearly uniform then Create a new class Cnew="Oracle" at level L and assign the Xu to this new class
Set parent(Cnew)=class choice at level L-1 (Database)
From the function of update constraints we update the constraints of the newly created class "oracle" as following:
1- Oracle is subclass of Database class (subset-constraints) and
2- Oracle class members cannot be member of any other class on the same level (mutually exclusive constraints)

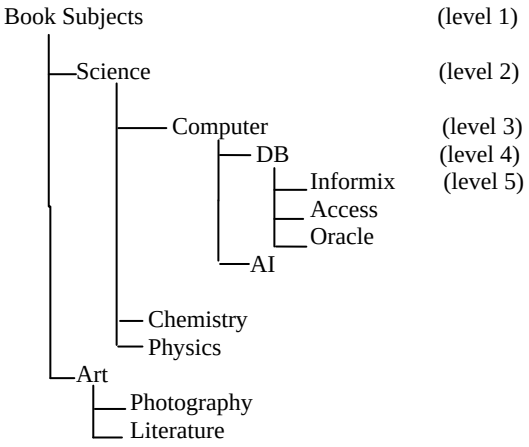

Figure 6. Updated ontology with new class Added

Searching After updating ontology shows enhancements in the values of precision and recall. From 80 books we retrieved 76 books 72 relevant to subject and 9 irrelevant to it.
Precision= (62/76+10)*100 = 90.7%
Recall= (62/62+18)*100= 81.5%
Updating ontology shows much better results of high precision and recall values than of keyword search and searching with incomplete ontology.

VII. CONCLUSION AND FUTURE WORK

Semantic Web technologies are valuable add-ons for digital libraries. In this paper we proved that using ontology which is one of the semantic web technologies in digital library structure to define concepts and relationship between entities; it organize and gives meaningful metadata about digital library content and finally it improves information retrieval and books classifications in digital library. Key word searching was not effective with low precision and recall values. Searching in digital library using updated ontology results in the best values of precision and recall. The proposed Hierarchy algorithm is using the books keywords and using naïve bayes classifier to automatically classify books into the created subject ontology whether by creating new class or assign it to any existing class. Digital Library Structure should be modified to include the semantic web layer and this leads us to build new digital library model 6S as extension of the 5S model to use the semantic web technologies in digital libraries.

Future work will consist of evaluating the implementation and approach more carefully, validating the 6S digital library model with a number of quality aware case studies and using large digital library resources and different types of resources not only PDF files contents. Also future work should consist of measuring different semantic techniques with digital library to increase the quality of digital libraries.

REFERENCES

[1] H. Suleman, Open digital libraries, in, Citeseer, 2002.

[2] M. Trivedi, Digital libraries: functionality, usability, and accessibility, Library Philosophy and Practice (e-journal), (2010) 381.

[3] H.-S. Hwang, K.-S. Park, C.-S. Kim, Ontology-based information search in the real world using web services, in: Computational Science and Its Applications-ICCSA 2006, Springer, 2006, pp. 125-133.

[4] J.M. Gómez-Berbís, R. Colomo-Palacios, Á. García-Crespo, CallimachusDL: using semantics to enhance search and retrieval in a digital library, in: Emerging Technologies and Information Systems for the Knowledge Society, Springer, 2008, pp. 540-548.

[5] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Scientific american, 284 (2001) 28-37.

[6] M.A. Gonçalves, E.A. Fox, L.T. Watson, N.A. Kipp, Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries, ACM Transactions on Information Systems (TOIS), 22 (2004) 270-312.

[7] U. Murthy, D. Gorton, R. Torres, M. Gonçalves, E. Fox, L. Delcambre, Extending the 5S digital library (DL) framework: From a minimal dl towards a dl reference model, in: Proceedings of the 1st Workshop on Digital Library Foundations, ACM IEEE Joint Conference on Digital Libraries, 2007, pp. 25-30.

[8] M.A. Gonçalves, E.A. Fox, L.T. Watson, Towards a digital library theory: a formal digital library ontology, International Journal on Digital Libraries, 8 (2008) 91-114

[9] B. Dalvi, W.W. Cohen, J. Callan, Classifying entities into an incomplete ontology, in: Proceedings of the 2013 workshop on Automated knowledge base construction, ACM, 2013, pp. 31-36

[10] R.M. Pampapathi, B. Mirkin, M. Levene, A suffix tree approach to email filtering, arXiv preprint cs/0503030, (2005).

[11] S. Sinclair, Adapting Bayesian statistical spam filters to the server side, Journal of Computing Sciences in Colleges, 19 (2004) 344-346.

[12] Digital library Example URL: http://spdl.sourceforge.net/index.htm

[13] M. Trivedi, Digital libraries: functionality, usability, and accessibility, Library Philosophy and Practice (e-journal) (2010), 381.

[14] H.-S. Hwang, K.-S. Park, C.-S. Kim, Ontology-based information search in the real world using web services, in: Computational Science and Its Applications-ICCSA 2006, Springer, (2006), pp. 125-133.