A Study on Retrieve and Archive of Web pages

Shien-Chiang Yu

Professor, Department of Information and Communications, Shih-Hsin University,
Taipei, Taiwan, Republic of China

Abstract

Web resources are recording the human societal information at that time, including visual design art, style, and included various kinds of resources in these web pages. The value of very having preserved. But web resources not only grow up quickly, but also disappear fast. Web resources will be probably unable to be utilized again because of factors such as server shutdown, revision, etc. at any time. Even preserved the pages, still face the fragile characteristic of digital information. Therefore, the preserved method must address methodological and practical issues to archive and manage digital preservation.

This study adopt two major research methods: content analysis and experiment. Through content analysis to explore the characteristics of web pages mode, the procedure of retrieve web content, structured data processing standards, and the relation of mapping between both. Experimental method implemented web tree down mining techniques.

In addition, to cope with the migration of HTML version, to avoid thereafter browser cannot parse the preserved web content today, this research also study long-term preservation issues, including standardization, in line with long-term application format, data extraction and restructuring and other factors. Based on these requirements, this study covers the solution of long-term preservation of web site archive. Using the way of Topic Maps syntax to reorganize unstructured HTML of original web pages into the semi-structured. Making Web content description can provide automated management, analysis and application.

Keyword:

Web archives, Breadth First Search, Topic Maps, Transformation

1. Introduction

The amounts of web resources have increased rapidly in recent years. Libraries and the information science professionals turn to cooperate in long-term preservation of digital resource to pass down human civilization, and many national-class libraries and archives around the world have launched web archives projects to ensure long-term preservation and provide continued access to digital resources (Kimpton & Ubois, 2006).

Observing the contents of the web resources, it could be noticed that most of them are written in unstructured html texts. Some of those html texts are static, and others are driven by computer programs. Therefore, to record the content and the way of production of the web resource is absolutely vital. The holistic plan of the description and management for the metadata format is necessary in order to preserve and to use the web resources in the future. Individual structure of web has to be analyzed, to deconstruct the attributes of the web content. The analysis of the web resource not only helps preserve the original relationship of the content, but also reconnects different web pages by the demand of the users.

HTML is a kind of procedural markup language and it is impossible to inference the meaning of the content. Because of that, it can’t be used for exact searching, automatic documents processing and content mining (Coombs, Renear & DeRose, 1987). This paper focuses on the execution project of the long-term preservation of the web resource. Try to make the reuse and exchange of web page resources through analyzing the formation and structure of the web resources. Using Topic Maps to connect with the semantic Web (Ahmed et al., 2001), and then try to analyze the structure of the web pages in order to make the reuse and the preservation of the raw pages.

2. Significance and Method of Web Archives

Confronted to different HTML version described pages, must consider about browser's support and long-term preservation issues, especial in standardization, data extraction, restructuring and other essential factors to be able to conform to future application. But the long-term preservation of digital information must consider technology paradigm shift, that is to say, digital information is easy invalid because of technical reasons. These technical factors of digital preservation include: inadequate media longevity, rapid hardware obsolescence and dependencies on particular software products (Day, 1998). Based on the fragility characters of digital preservation, the preservation of web resources must through research and experiment to develop strategic methods of archive and management. Wauth et al (2000) proposed five keys to long-term preservation of digital information:

(1) Encapsulation: wrapping the information within descriptive metadata;

(2) Self documentation n: can be able to understand and decode the data;

(3) Self sufficiency: minimum rely on system, data, or documents;

(4) Content documentation: Able to provide future users find or implement software to present information;

(5) Organization preservation: allows an organization to use actually.

According to these keys, this study must consider the multi-level description model with the appropriate metadata formats to meet the key requirements for encapsulation; and adopt standard markup syntax to achieve the purpose of self documentation and self sufficiency. Besides, developing a structured approach to conform the application target of content documentation and organization preservation.

The multi-level description method and metadata format will follow the study of scholar Wang (2007). To specific implement control level for retain the original resource content and data structure. This mode is different from subjectivity of subject content classification, therefore have objectivity and purpose rationality for organizational tools, and can express the background situation of resource generation, provide the use value of resource generation history. In the information level of web page, content will contain text, graphics, images, and many different elements of multi-media composition.

To achieve preserve all information of page content for future application, this study design following ways:

(1) Save entire content and structure of original web pages, and through artificial markup to record metadata for indexing and utilizing;

(2) Re-deconstruction original web page, and then adopt the XML syntax that was provided the characteristics of resource description, and retain the original linked relation. Re-construct web resources be able to render original content, but also extend application scope.

To archive web pages must cover "long-term preservation" and "do minimal harm" both principles. Compare these two principles, the second principle obviously can provide better preservation applicability.

Based on the difference characteristics and application purposes of data content to achieve long-term use of digital information, there are 5 major methods include: system preservation, refreshing, migration, emulation, standardization, encapsulation, redundancy, converting to paper or analog media, etc (OuYang, 2007). In order to conform the five keys to long-term preservation of digital information which was discussed previously, and refer to Lawrence et al. (2000) study conclusions and recommendations, the migration method is better to the quality requirements of risk management. Furthermore, analysis the characteristics of HTML format, this study adopt migration is the most suitable approach.

The research purpose is exploring the long-term preservation solution, and must follow standardization to implement. According to the perspective of the data structure to analyses content characteristics and composed patterns. This study divided resource preservation into two levels, one is the metadata described web pages, and the other is reconstruct web page into structured document. Besides, the research design on preservation of web pages, the major approach is used migration to achieve the purpose of preservation. Therefore, this study apply system analysis method to reconstruct web pages. If the content belongs to particular resource, such as JavaScript, Applet, Flash and other programs or objects, based on type can be described by Resource Description Framework (RDF), then through Topic Maps to link these resources all together, combined with the multi-level description metadata record to provide overall of web site effective preservation, management and utilization.

3. Topic Maps

3.1 The Origin and development of Topic Maps

To enrich Standard Generalized Markup Language (SGML, ISO/IEC 8879-1986) with functions of multimedia and hyperlinks, Goldfarb and Stev Newcomb tried to design an architectural form which could make hyperlink available to multimedia and documents at any time. That made Hypermedia/Time-based Structuring Language (HyTime) be introduced, and HyTime became an ISO and International Electrotechnical Commission (IEC) joint standard upon publication as ISO/IEC 10744 in 1992 (SGML SIGhyper, n.d.). Being inherited from SGML, the syntax rule of HyTime is very complicated, and that intrigued the Graphic Communication Association Research Institute (now known as IDEAlliance) in an activity called Conventions for the Application of HyTime (CApH) proposed a revised clause for identifying information objects that share a common topic (Pepper, 1999). The solutions developed are called "Topic Navigation Maps" which became an ISO/IEC standard as ISO/IEC 13250 in December of 1999. Topic Navigation Maps adopts HyTime as the definition syntax, and it is a kind of Document Type Definition (DTD) of SGML. It can define subset of various types of field and rename their name and attribute. According to that, it could describe various kinds of information concepts as topics which possess their own name, attribute, resource guide, etc. Besides, it could define the relations that topics bear to one another.

Moreover, to break through the restriction for application of SGML and HTML, World Wide Web Consortium (W3C) developed a new generation markup language-eXtensible Markup Language (XML) which could be used in web environment and can define interchange format of structured data file (Bray, Paoli & Sperberg-McQueen, 2001). Not composed by specific tags like HTML and supporting language neutral as well as platform neutral, XML allows users to define markup languages needed by themselves and could be used in various areas widely. So TopicMaps.Org established in 2000 adopted XML syntax to develop XML Topic Maps (XTM) 1.0 Specification in 2001 (Pepper & Moore, n.d.). In 2002, ISO/IEC 13250: Topic Maps containing 2 syntax structures which are HyTime and XTM are approved (Pepper, 2005).

3.2 Elements of Topic Maps

The XTM standard identifies the key of Topic Maps. The key concepts sum up as the "TAO" of Topic Maps s, from the initials of the constructs for representing find aids: topics, associations, occurrences, subject descriptor, and scope (Park & Hunting, 2002)(Newcomb, Biezunski & Bryan, n.d.)(Daconta & Smith, 2003):

(1)Topics

A topic is a representation of the subject; according the XTM standard, it acts as a resource that is a proxy for the subject. Each topic is implicitly an instance of a topic type-that is, the class of the topic.

(2) Occurrence

An *occurrence* is a resource specifying some information about a topic. The resource is either addressable (using a URI) or has a data value specified inline.

(3) Association

An association is the relationship between (one or more) topics and it is represented as <association> in XTM Structure.

(4) Subject Indicator

Originally named as subject descriptor in ISO/IEC 13250, a subject indicator is a resource that is intended by the Topic Maps author to provide a positive, unambiguous indication of the identity of a subject.

(5)Scope

Scope is a special topic and it defines a group or a specific range of related topics. The function of Scope is similar to name space: The base name of topic should be the only one in a certain scope.

3.3 Topic Map and RDF Comparison

RDF (Resource Descriptive Framework) is a common framework specified W3C that adopt XML to provide exchange and present information organization resources among application or agent (Sarukkai, 2002). But, Topic Maps just like W3C's Semantic Web belongs to the same class of metadata use to descript the relation among resources (Cregan, n.d.). As shown in Table 1. Compare Topic Maps and RDF these two standards. Whether used for the purpose or technical implementations are very similar: both can be described in XML, are also provided with constraint language and the query language. All the concepts related to the Topic Maps can be similarly expressed in RDF, but RDF focus more on the description of resources, while the Topic Maps is emphasized on the link between resources. If transferred the resources described in RDF into elements of Topic Maps, there will be some loss of semantic (Borghoff et al., 2005).

Table 1. Comparison of Topic Maps and RDF

		Topic Maps	RDF
Same	Constraint Language	Topic Maps Constraint Language (TMCL)	RDF Schema
	Query Language	Topic Maps Query Language (TMQL)	RQL(RDF Query Language)
	Description Language	with XML Topic Maps (XTM) syntax specification	W3C Recommend adopt XML as RDF description language, but is not mandatory
	Unique Identifier	Using "id" attribute as the unique identifier of subject, URI can also be used as a resource unique identifier	Using the URI as a resource unique identifier
Different	Responsible	ISO	W3C
	Utility(Garshol, n.d.)	created to support high-level indexing of sets of information resources to make the information in them findable	intended to support the vision of the semantic web through providing structured metadata about resources and a foundation for logical inference
	Scope	built-in *scope* element, support the function of scope	Does not have the function of scope

	Reference mode	Use *xlink:href* attribute indirect reference	Only direct reference mode using

4. Design the Web Archive Model

To sum up, the nature of Topic Maps is very simple. Through the application and relation of elements to simulate HTML. So, we can use Topic Maps to record the full resources of web pages. Besides, we can also link web pages of entire site. Through the establishment of these relations, we can reorganize unstructured web pages into structured document which have the coordinate concepts.

4.1. Preserve complete resources of web pages

Use the "topic" element as the basic unit of web page, the individual "topic" element through the description of subject type to group the relational topics, and then by "occurrence" element gather topics of event-related together. Adopt "association" element links the semantics between related events. And, utilize "scope" element to limit the effective range of name, resource indicator, associations. So we can preserve the entire resource of original web pages under a structured syntax.

4.2. Keep the link of each web pages within entire site

As show in Figure 1. The "association" element combine with "roleSpec" and "instnceOf" elements, will be able to represent clearly the up and down between the levels of topic relevance. Expressing such as the conception of thesaurus (Kowalski & Maybury, 2000). Therefore, Topic Maps can use these elements to keep the link of each web pages within entire site

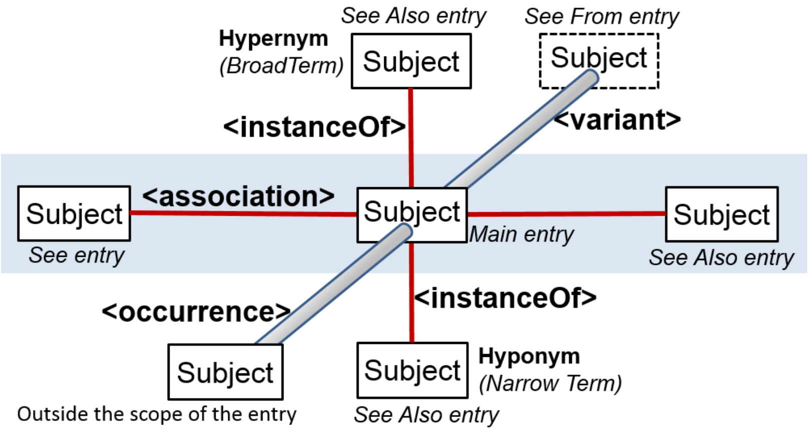

Figure. 1. Based on the Topic Maps indicate the authority control among subjects

4.3 Workflow and procedure

Analysis of the above procedures, this study design processes shown in Figure 2.

The preservation contains recording metadata descriptive data and keeping original resources of web page these two major operating items. The preserved method of original resources are implemented through the following parsing procedure:

First deconstruction HTML pages and content clustering (Zoning). Second, follow XTM syntax converted into the Topic Maps document. Then, combined with metadata records to link related external resources such as multi-media which include in the original web page... procedures. Finally, store the Topic Maps document and metadata record into the database.

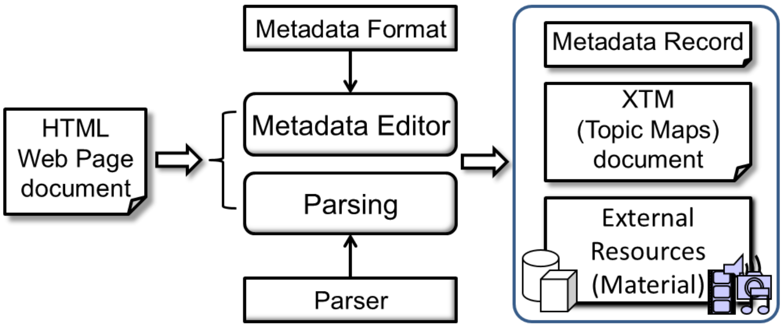

Figure 2. The workflow and procedure planning

4. The Benefits of This Study

Considering the application would be restricted by archived web resources themselves in the future, the authors try to make use of XTM to deconstruct HTML elements of web pages, describe and markup them, then apply XTM association and occurrence to keep original link and related reference of web resources. Several benefits of this study are listed below:

(1) Possess the dependability that the file is kept for a long time (stability)

(2) Offer the convenience made and rebuild again of the materials

(3) Realize the usability of file content mining

(4) Support the interoperation between the systems

(5) The scarce one is structural to solve HTML file

5. Design Practice

5.1 Retrieve pages

The initial procedure of web archives is use web crawler (may also be called a web spider) to systematically retrieve overall pages within the specified site. The web structure can be seen as a graph with multiple nodes. Each page is one node. The links in one page can be seen as a "directed edges" of this graph. Therefore, we can through traverse this graph (web site) to retrieve the contents of each page. There are two primary graph traversal algorithms: breadth-first search (BFS) and depth-first search

(DFS) (Skiena, 2008).

Because DFS may be in depth traverse too deeply cause crawler to fall into "black hole" (an infinite link loop), therefore this study adopted BFS mode to traverse web site. Home page as a beginning and assign each link a priority sequence, and then according to the order to retrieve and parse linked page.

5.2. Deconstruct web page

Single web page shown on user's browser usually are HTML documents. As Figure 2 shows, its content is contained within the range which <HTML> Tag declares. The HTML document could be separated in two parts: <HEAD> Tag providing identification information for application software, and <BODY> Tag providing content represented by the web page (Smiraglia, 2005).

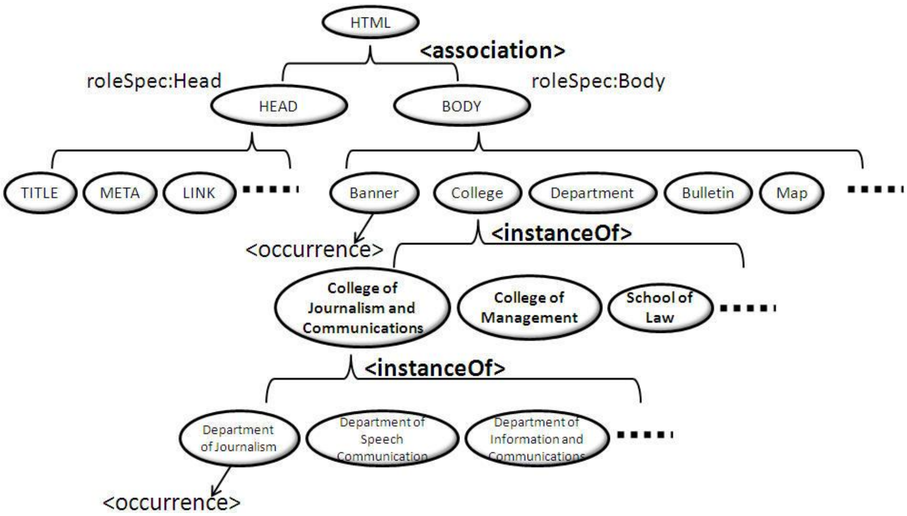

Figure 3. Deconstruction web page using Topic Maps example

Firstly, the two document range declared by <HEAD> and <BODY> are regarded as two different topics. If a single web resource is regarded as a Topic Maps, each Topic Maps could be combined via <mergeMap> Tag of XTM. If a web page is a combination of several web pages which associate with each other through <FRAME> Tag, the Topic Maps of single web page could be combined via <mergeMap> Tag, too.

<BODY> Tag declares the actual content of a web page resource. Distinguishing web page elements by their function, there are four kinds of information contained in <BODY> Tag: tag, text, hyperlink, and embedded object. The way in which HTML elements converted into XTM is described below:

(1) Tag: HTML Tags are usually used for dealing with the representation of data. Tags could be converted to independent Topics generally, and Tags with

attribute could be recorded one by one using the <parameters> Tag of XTM <variant> Tag. Besides, the id value of repeated Tags could be identified by adding serial number to the original Tag name.

(2) Text: Like CSS, Scripts (such as JavaScript, VBScript) and text data shown on web page context, etc. , this study describes the resources by their type using RDF standard, then links the relationship between resources through Topic Maps.

(3) Hyperlink: Tags with attribute *href* or *src* like <a>, <img>, <embed> ,etc. provide links to other web resources by hyperlinks, or make other multimedia objects embedded in web pages. XTM make hyperlinks available by using the syntax XLINK.

(4) Embedded object: they are programs or objects like Java Applet, Plug-in software (such as Flash), etc., and can be regarded as multimedia objects to store separately for the purpose of preservation management.

5.3. Construct Topic Maps

In this study, Using Shih-Hsin university English website (http://english.shu.edu.tw/) as a test. Try to deconstruct web page automatically into blocks. As shown in Table 4, the process of deconstruction is the same as parsing HTML. Identify each element within HTML file and complete parsing, and then one by one each block of HTML syntax content converted into XTM syntax.

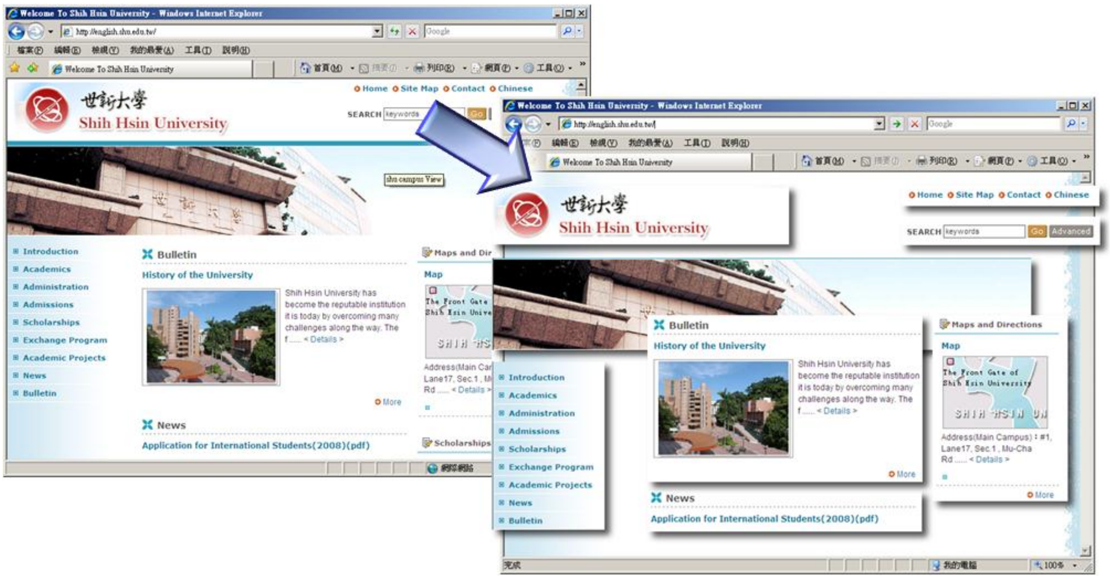

Figure 4. Deconstruction web page example

On the left of the web, taking Academics and Administrator these to menu items page as example, the HTML code shown in Table 2., menu item uses "href" attribute to link to specified page.

Table 2. HTML code of link to a specified page.

```
<td class="leftbg">

  <H3><A title="Academics    " href="Academics/Departments.htm">Academics</A></H3>

  <H3><A title=Administration" href="Administration/Offices.htm">Administration</A></H3>

……

</td>
```

According to the tag mapping and processing convert web page content. The conversion result listed in Table 3, including <head> element as well as web pages of Academics and Administrator partial content which was converted into Topic Maps document in XTM syntax described.

Table 3. Sample for converting web page (partial) into Topic Maps document of XTM

1. head element

```
<topic id="default.html">
     <baseName>
          <baseNameString>http://english.shu.edu.tw</baseNameString>
     </baseName>
  </topic>

  <topic id="head">
       <baseName>
            <baseNameString>Head of web page</baseNameString>
       </baseName>
       <instanceOf><topicRef xlink:href="#default.html"/></instanceOf>
  </topic>

  <topic id="title">
            <instanceOf><topicRef xlink:href="#head"/></instanceOf>
            <baseName>
                 <baseNameString>Welcome To Shih Hsin University</baseNameString>
            </baseName>
  </topic>

  <topic id="meta">
            <instanceOf><topicRef xlink:href="#head"/></instanceOf>
            <baseName>
                 <baseNameString>meta</baseNameString>
                 <variant>
                      <parameters id="http-equiv">Content-Type</parameters>
                      <parameters id="content">text/html; charset=utf-8</parameters>
                 </variant>
                 <variant>
                      <parameters id="name">Education</parameters>
                      <parameters id="content">Welcome To Shih Hsin University</parameters>
                 </variant>
```

```
        </baseName>
    </topic>
```

2. Academics information(Part)

```
<association id="unit">
   <instanceOf><topicRef xlink:href="#Academics"/></instanceOf>
   <member>
        <roleSpec>
            <topicRef xlink:href="#session1"/>
        </roleSpec>
        <topicRef xlink:href="#smgmt"/>
    </member>
   <member>
        <roleSpec>
            <topicRef xlink:href="#session1"/>
        </roleSpec>
        <topicRef xlink:href="#cjc"/>
    </member>
   <member>
        <roleSpec>
            <topicRef xlink:href="#session1"/>
        </roleSpec>
        <topicRef xlink:href="#law"/>
    </member>
</association>
```

3. Administrator information(Part)

```
<association id="dept">
        <instanceOf><topicRef xlink:href="#department"/></instanceOf>
        <member>
            <roleSpec>
                <topicRef xlink:href="#session2"/>
            </roleSpec>
            <topicRef xlink:href="#oga"/>
        </member>
           <member>
            <roleSpec>
                <topicRef xlink:href="#session2"/>
            </roleSpec>
            <topicRef xlink:href="#osa"/>
        </member>
           <member>
            <roleSpec>
                <topicRef xlink:href="#session2"/>
            </roleSpec>
            <topicRef xlink:href="#ord"/>
        </member>
        <member>
            <roleSpec>
                <topicRef xlink:href="#session2"/>
            </roleSpec>
            <topicRef xlink:href="#oga"/>
        </member>
        <member>
            <roleSpec>
                <topicRef xlink:href="#session2"/>
            </roleSpec>
            <topicRef xlink:href="#library"/>
        </member>
        <member>
            <roleSpec>
                <topicRef xlink:href="#session2"/>
            </roleSpec>
            <topicRef xlink:href="#po"/>
        </member>
        <member>
            <roleSpec>
                <topicRef xlink:href="#session2"/>
            </roleSpec>
            <topicRef xlink:href="#oa"/>
        </member>
    </association>
```

5.4 Process Description and Empirical Test

By analyzing the structure and content of the HTML file syntax features, This study sorts out the processing flow illustrated in Fig. 5, that offers archive application program to map the HTML syntax of each component to XTM syntax and convert the content of web pages to Topic Maps document.

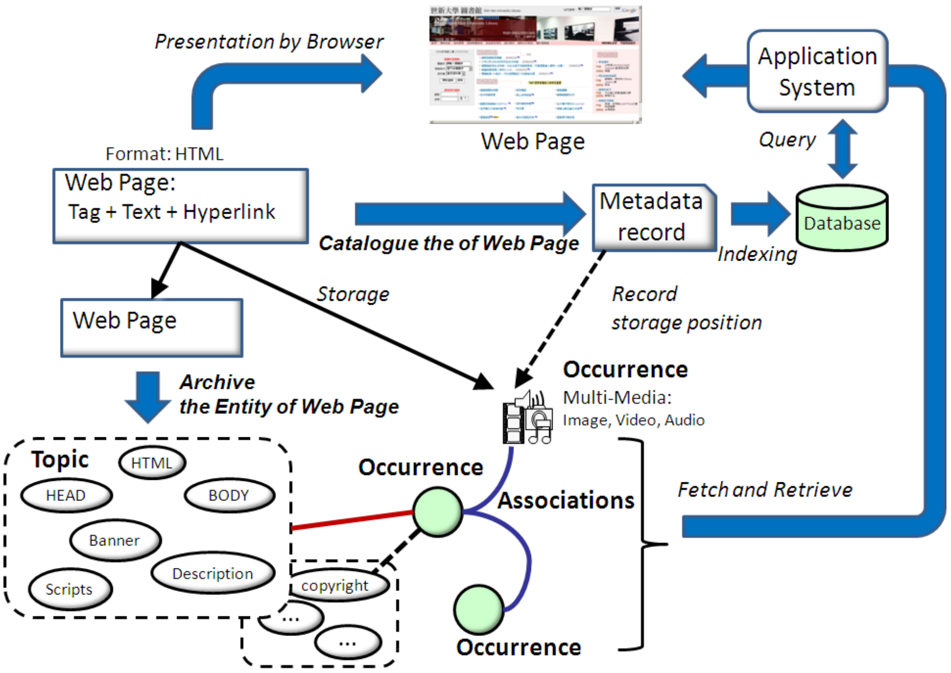

Figure 5. The Major process flow of this study

Being used for describing the web page contents, Topic Maps could be the guide of them, and it could also reflect the structure of web resources. Users can search for web resources needed through description of metadata, they could find out links between web resources and recompose them additionally via the structure displayed by Topic Maps. When searching for a specific web resource, users don’t have to search in a large database. Only if with the links provided by Topic Maps, users could find out related topics and events.

Additionally, this study use open-source tools developed by the Topic Maps for Java program (TM4J) (Ahmed, n.d.) to parse the web page shown in Fig.3 and get the converted Topic Maps document shown in Fig. 6. By representing the visualized environment of Topic Maps, TM4J can provide interactive function of interface and display topic concepts clearly and completely to help users understand the topic concept between web resources. This study is trying to achieve the purpose of reviewing the correctness of documents conversion through visualized interface.

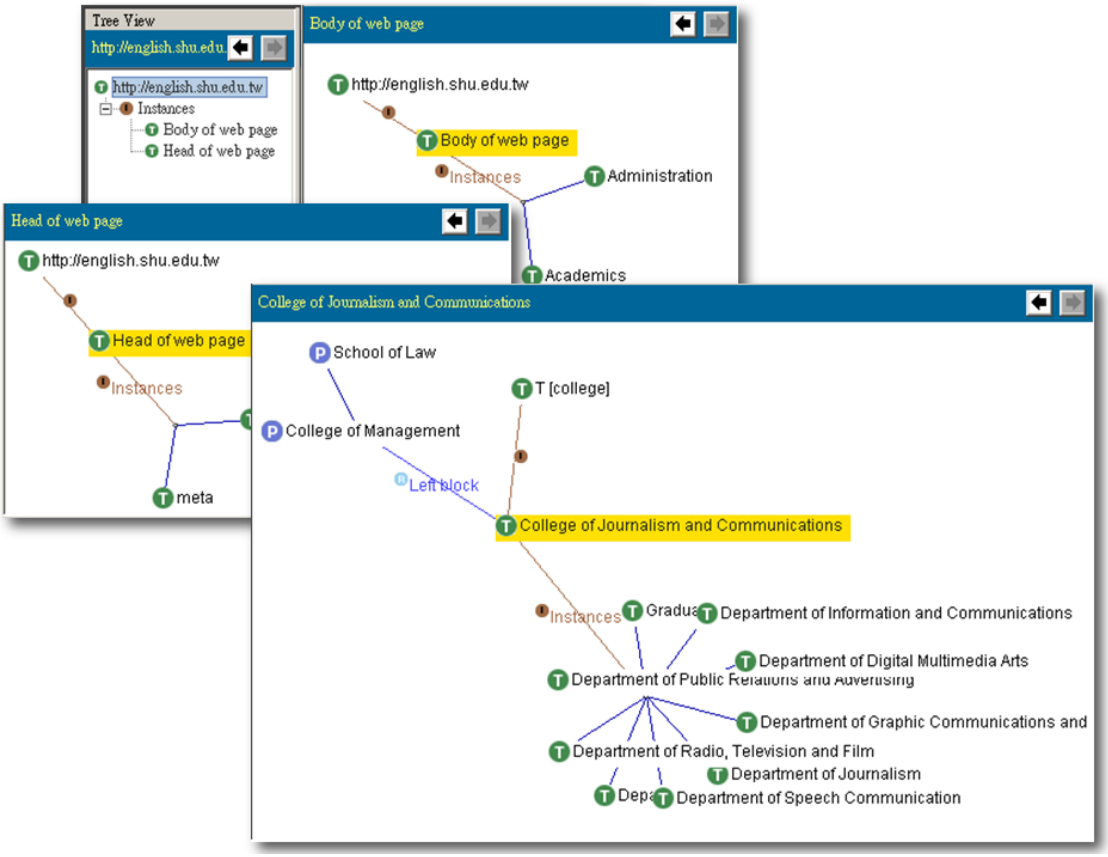
Tree View
http://english.shu.edu
http://english.shu.edu.tw
Instances
Body of web page
Head of web page
Body of web page
http://english.shu.edu.tw
Body of web page
Instances
Administration
Academics
Head of web page
http://english.shu.edu.tw
Head of web page
Instances
meta
College of Journalism and Communications
School of Law
T [college]
College of Management
Left block
College of Journalism and Communications
Instances
Gradua
Department of Information and Communications
Department of Digital Multimedia Arts
Department of Public Relations and Advertising
Department of Graphic Communications and
Department of Radio, Television and Film
Department of Journalism
Depa
Department of Speech Communication

Figure 6. Apply TM4J tools to review the converted Topic Maps document

5. Conclusion

Archive application program offers the function of converting web pages to Topic Maps documents, it also permits Topic Maps documents restoring to web pages. Moreover, considering, the structured XTM document converted from the unstructured HTML document could meet the new standard of web markup language to achieve the purpose of long-term preservation.

References

Ahmed, K. (2002). TM4J Developer's Guide. Retrieved October 14 , 2013, from http://tm4j.org/tm4j/docs/devguide/

Ahmed, K., & Rivers-Moore, D., & Lubell , J., & Watt, A., & Birbeck, M., & Cousins, J., & Rob, W., & Nic, M., & Ayers, D., & Wrightson, A. (2001). *Professional XML Meta Data*. UK: Wrox Press.

Borghoff, M. U., & Rödig, P., & Scheffczyk, J., & Schmitz, L. (2005). *Long-Term Preservation of Digital Documents*. New York: Springer.

Bray, T., & Paoli, J., & Sperberg-McQueen, C. (1998). Extensible Markup Language (XML) 1.0. Retrieved October 14, 2013, from http://www.w3.org/TR/1998/REC-xml-19980210

Coombs, H. J., & Renear, H. A., & DeR, j. S. (1987). Markup systems and the future of scholarly text processing. *Communications of the ACM, 30*(11), 933-947.

Cregan, A. (2005). Building Topic Maps in OWL-DL. Retrieved October 14, 2013, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.7781&rep=rep1&type=pdf

Daconta, C. M., & Obrst, J. L., & Smit, T. K. (2003). *The Semantic Web: a Guide to the Future of XML, Web Services, and Knowledge Management*. Indiana: Wiley.

OuYang, C. J. (2002). Investigation on Digital Information Preservation. *Archives quarterly, 1*(2), 36-47. (In Chinese)

Garshol, M. L. (2003). Living with topic maps and RDF. Retrieved October 14, 2013, from http://www.ontopia.net/topicmaps/materials/tmrdf.html

Kimpton, M., & Jeff, U. (2006). *Year-by-Year: From an Archive of the Internet to an Archive on the Internet", in Julien Masanès Editor, Web archiving* (in Julien Masanès Editor). Berlin: Springer.

Kowalski, J. G., & Maybury, T. M. (2000). *Information Storage and Retrieval Systems: theory and implementation*. Boston: Kluwer Academic Publishers.

Lawrence, W. G., & Kehoe, R. W., & Rieg, Y. O. (2000). *Risk Management of Digital Information: A File Format Investigation*. Washington, DC: Council on Library and Information Resources.

Newcomb, R. S., & Biezunski, M., & Bry, M. (1999). ISO/IEC 13250 Topic Maps: Information Technology-- Document Description and Processing Languages. Retrieved February 25, 2008, from http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0129.pdf

Park, J., & Hunting, S. (2002). *XML Topic Maps: Creating and Using Topic Maps for the Web*. Boston, MA: Addison-Wesley.

Pepper, S. (1999). Euler, Topic Maps, and Revolution, (Eds.), *Proceedings of XML Europe 99 Conference* (pp. 2). Alexandria, VA: GCA.

Pepper, S. (2002). The TAO of Topic Maps: finding the way in the age of infoglut. Retrieved October 14, 2013, from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=956E75187A164A99E503231655D62B12?doi=10.1.1.32.5473&rep=rep1&type=pdf

Pepper, S., & Moore, G. (2001). XML Topic Maps (XTM) 1.0. Retrieved October 14, 2013, from http://www.topicmaps.org/xtm/1.0/

Sarukkai, R. R. (2002). *Foundations of Web Technology*. Boston, MA: Kluwer.

SIGhyper, S. (1994). A Brief History of the Development of SMDL and HyTime. Retrieved October 14, 2013, from http://www.sgmlsource.com/history/hthist.htm

Skiena, S. S. (2008). *The Algorithm Design Manual* (2 Ed.). London: Springer.

Smiraglia, P. R. (2005). *Metadata: a cataloger's primer*. NY, Binghamton: Haworth.

Wang, L. (2007). Web Archives: The Concept and Application of Multi-level Description Model. *Journal of Educational Media & Library Sciences*, *44*(4), 455-471. (In Chinese)

Waugh, A., & Wilkinson, R., & Hills, B., & Dell'oro, J. (2000). Preserving Digital Information Forever, *Proceedings of the fifth ACM conference on Digital libraries* (pp. 175-184).