A Study on Web Archives Design: The Description and the Format Approach

Li-Chiao Wang; Executive Officer, Institute of Taiwan History, Academia Sinica, & PhD. Student, Department of Library and Information Science, National Taiwan University; Taipei, Taiwan : Primary author
Shien-Chiang Yu; Associate Professor, Department of Information & Communications, Shih-Hsin University; Taipei, Taiwan : Co-author
Wen-Ying Wang; Officer, Planning Division, National Archives Administration; Taipei, Taiwan : Co-author

Abstract

This paper focuses on the long-term preservation of the web resource. The authors try to explore web archives strategy from two viewpoints which are metadata format and the structure of contents. In the analysis of the metadata format, review several web archives programs in the world and analyze their rule of organizing information. Secondly, study the archival principle of provenance and its application in controlling levels of archives description. In addition, analyze multilevel structure of web resources to explore the possibility of developing a practical model for multilevel description of web resources. Finally, make use of Topic Maps to deconstruct web pages and build up the relational model of topics to achieve the goal of preservation and providing continues access to web resources.

1. Introduction

The amounts of web resources have increased rapidly in recent years. The library science and the information science professions 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 programs to ensure long-term preservation and provide continued access to digital resources.

Organization of information like arrangement and description are the foundation of making use of information. In the web world, how to describe various kinds of digital resources in an effective and consistent way is more important than ever. Subject-oriented web resources archiving, involving possible misleading because of the collector's decision, may make resources removed from the original context, their links with other related resources be cut off, and even important resources with academic research value be missed. To find out a way to archive web resources and describe their content as well as creation information in a detailed and accurate way is essential for preservation and sustained access.

Moreover, by deconstructing the structures of web pages like sharing, linking and frame, etc., various attribute of web resources content need to be extract from each other. It contributes to preserve the original relationship of web pages, and helps to link different web pages again as users' need. And Topic Maps, containing basic models and index functions of semantic web, could be used as the structure of representing knowledge. Besides, most of the web pages are created through unstructured html files which are a kind of procedural markup language and used for information presentation. It's impossible to inference the meaning of the contents and that makes web pages not allow for exact searching, automatic documents processing and content mining. This study focuses on not only long-term preservation of the web resource, but also tries to make web resources reusable for data mining or other application.

By analyzing the formation and structure of the web resource, this study tries to explore web archives strategy from two viewpoints which are metadata format and the structure of contents. In the analysis of the metadata format, the authors review several web archives programs in the world and analyze their rule of organizing information. Secondly, the authors study the archival principle of provenance and its application in controlling levels of archives description. In addition, the authors analyze multilevel structure of web resources to explore the possibility of developing a practical model for multilevel description of web resources. Finally, the authors make use of Topic Maps to deconstruct web pages and build up the relational model of topics to achieve the goal of preservation and providing continues access to web resources.

2. Multilevel Description of Web Archives

2.1 Preserving the Web and Archival Description

Because of the great changes in information technology, such as the development of internet and World Wide Web, the main media for storing human knowledge have been shifted from paper to digital form. Web is the main communication channel of digital resources, and it is also the largest assembling and distribution center of digital resources in the world. As digital resources grow and disappear so rapidly, the importance of preserving them has been noticed in the world and several web archive program have been carried out, such as the Pandora project of the National Library of Australia from 1996, the Internet Archives project of San Francisco, U.S.A., and the Minerva program of US Library of Congress from 2000.

For the purpose of long-term preservation and providing access to web resources, the core work is the arrangement and description of them. The Internet Archive program snapshots the whole web domain to collect large amount of web pages and doesn't process or describe them. It provides users to inquiry URL and browse web pages, but doesn't provide keyword or content searching. The Pandora program selects and collects web resources based on their content subject, describes the whole website or the

collection in the light of cataloguing standards of digital resources like AACR2 and MARC. It not only builds up an inquiry system, but also makes the catalogue of web resources created be integrated with the national bibliographical network for consistency inquiry. The Minerva project of LC also selects and collects web pages resources according to their subject, but it provides two-level arrangement and description of web resources. It takes AACR2 and MARC as the standard of cataloguing at the collection level, and adopts MODS standard as title description standard at the title level.

Like archival materials, web resources are organic and structural, and they grow up continuously and link with each other. Archival description represents the fonds, a complex body of materials, frequently in more than one form or medium, sharing a common provenance. The description involves a complex hierarchical and progressive analysis. It begins by describing the whole, proceeds to identify and describe sub-components of the whole, and sub-components of sub-components, and so on. Frequently, but by no means always, the description terminates with a description of individual items. The description emphasizes the intellectual structure and content of the material, rather than their physical characteristics. (Pitti, 1999) Different from subject-oriented organization of information, archival description based on provenance is objective and reasonable. Moreover, it is agreeable with the organic and growing nature of web resources. Namely, it's effective and feasible to describe accumulated web resources using archival description multi-level model.

Many archives serve the law, functioning as the institutional memory of specific corporate bodies. Government agencies, public institutions, and businesses have legal requirements pertaining to the keeping of records. Archives and manuscript libraries also remember on behalf of history, which is to say, they preserve a large portion of the raw material on which our historical understanding is based. Both legal and historical memory requires a high degree of user confidence in the authenticity and integrity of records and documents. The materials in archives and manuscript libraries are *evidence*, both legal and historical. In contrast to the published items collected by libraries, the identifiable object of interest in the archive is a complex body of interrelated, unique materials. The *fonds* coheres and is identifiable because all of its records or papers share a common *provenance*, derived from one source and context. (Pitti, 1999) As the development of archival science, the theoretical structure of archival arrangement and description has been formed concretely. Namely, the concept of archival arrangement and description is based on the principle of provenance, shown as respecting fonds and original order, and with the application of archival levels of control. The implementation of Principle of Provenance could be observed in archives levels of control, in accordance with respecting de Fonds and original order, processed from general to specificity, from summary to details.

2.2 The Application of Multilevel Description in Web archives

Arrangement and description based on the principle of provenance, different from subject-oriented organization of information, is objective, rational, and suitable for describing Web resources with organic and accumulating nature. Web resources are identified with Uniform Resource Locator (URL) which is combination of communication protocol (like http, ftp, and gopher), hostname, path, filename, such as http://www.ntu.edu.tw/info/.

The concepts of domain names and hierarchical attribute (RFC1034 and RFC1035) were introduced in 1984. The whole structure of Domain Name System (DNS) is like a tree, and it fulfill the goal of searching correctly via the only Root Server in the world. Domain names are administrated in a hierarchical way: the administrator can set up sub-domains for different departments' use; he can also authorize other department to administrate the sub-domain by itself only if the administrator has made declaration at the last upper layer.

According to the hierarchical structure, levels of control of web resources could be established and the arrangement and description of them could be created subsequently. Each level's definition of web resources is as follows:

I. The first level: the institutional or individual website

Considering the creating institution of web resources as a whole, the domain name of the institution could be regarded as the first level of arrangement and description. The domain name is the identification of organization, enterprise or individual in cyberspace and it could be used for identifying the fonds clearly.

II. The second level: sub-domain, primary content or service items under website

According to the service or function provided by website, there could be many different web series under the institution web domain. For example: Info.ntu.edu.tw, Information network for students of the National Taiwan university.

III. The third level: web pages with the same nature in a sub-domain or service item

To cope with Web resources' growth, it is convenient for preservation and management to gather related web pages together and arrange them in alphabetical or date order.

IV. The fourth level: specific webpage

Web pages are basic components of Web resources, existing in the form of text, video, sound as well as animation, and could be identified and retrieved via URL in Internet.

Levels of control of web resources established in accordance with domain names and URL structure are defined with the original structure of website built up by the creators. We could say that it is the realization of the archival principle of provenance and levels of control.

Both the Pandora program and the Minerva program of US Library of Congress take collections of websites as the primary descriptive level. Most collections are acquired and preserved in subject-oriented ways, and the description of them focuses on the contents of them. On the other hand, the multilevel description of web resources is in accordance with principle of provenance, the structure and hierarchical structure established by the creator to carry on multilevel arrangement and description.

Multilevel description model established in accordance with the principle of provenance goes through web resources gradually, from general to specificity, from summary to details, and matches up with accumulating as well as organic nature of web resource.

Fig. 1 shows the application structure of multilevel description model in organizing the information of web resources. Each arrangement level of control has its own need for information content to satisfy the requirement of management, preservation, and content searching.

The detail extent of information content describing in each level of control is in inverse proportion to its position in the hierarchy. Namely the description of the highest level is summary and that of the bottom level is detailed to content of single webpage. The description contents of each level are listed below:

I. The first level: the institutional or personal website

The first level content of description is based on the principle of provenance and with respect to the creators (institution or individual) of web resources. Different from the subject-oriented collection of the Pandora program and the Minerva program, It regards (institutional or personal) the whole website as a provenance, content and is an objective structure based on web domain. The description should cover information need of the website's creator, owner and collector, and the descriptive information should include the background information of website's creator like institutional history, individual biography, the summary of website content, the rule of retrieving website, preservation needs, and management of website organization, etc. It's the highest priority for web archives manager to collect the institutional or individual's and the whole content summary of the website.

II. The second level: sub-domain, primary content or service items under website

Sub-domain or service items of the institutional or personal website are usually the classification result made by website creator according to the subject content or service function. The description items of this level should contain unit name, summary of content, subject, dates covered by web resource, and organization of web resource, etc.

III. The third level: web pages with the same nature in a sub-domain or service item

Website creator gathers web pages with the same nature under the second sub-domain or service items for the purpose of convenient management and usage. The description items of the third level are unit name unit name, dates covered by web resource and the format of web resources.

IV. The fourth level: single web page

They are the most basic elements. They could be single web page identified, linked, and retrieved by URL, and their format could be text, image, sound, animation, etc. The description of single webpage can be detailed to the content of the web page, including full text.

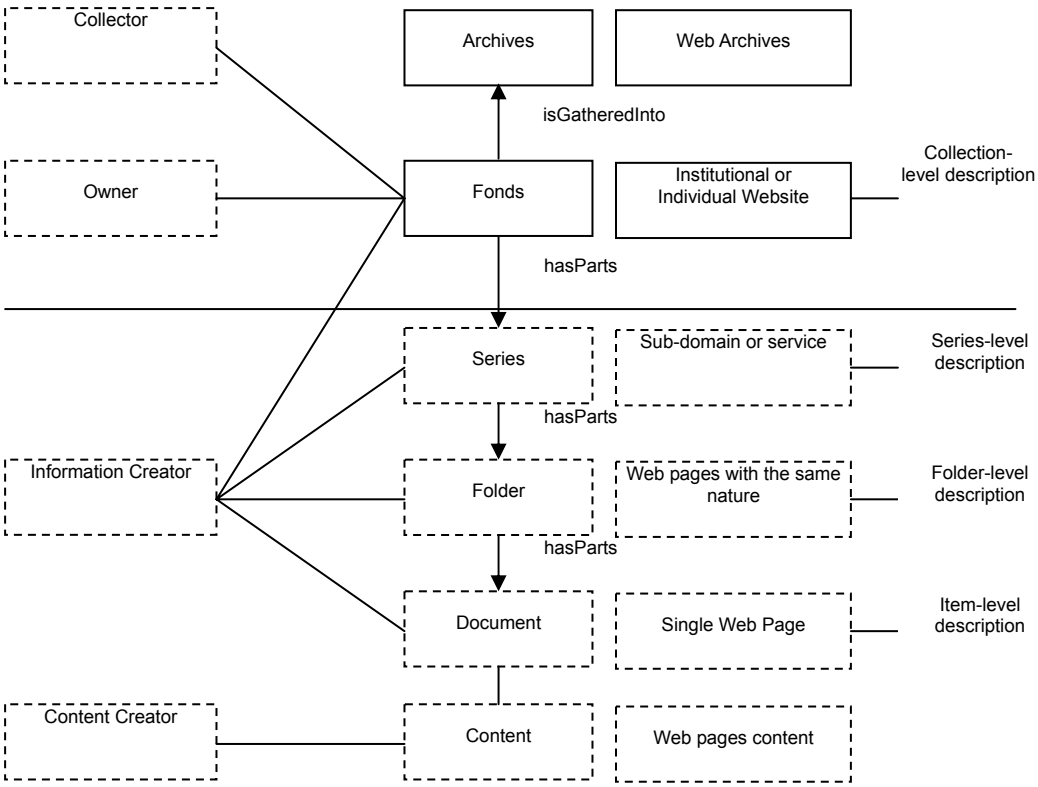

Figure 1. Multilevel Description Model of Web resources (Adopted from RSLP, RSLP collection description model)

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, 1994). 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 proposed a revised clause for identifying information objects that share a common topic (ISO JTC1/SC18/WG8, 1997). 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. 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. (Yu and Chen, 2001) So TopicMaps.Org established in 2000 adopted XML syntax to develop XML Topic Maps (XTM) 1.0 Specification in 2001 (Pepper and Moore, 2001). In 2002, ISO/IEC 13250: Topic Maps containing 2 syntax structures which are HyTime and XTM are approved.

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 Map s, from the initials of the constructs for representing find aids: topics, associations, occurrences (Pepper, 2000),subject descriptor, and scope (Daconta, Obrst, and Smith, 2003):

I. *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. Subject could be regarded as one "what", and one topic is the information representation of "what". So a topic represents the subject that is referred to and it is an essential component of constructing Topic Maps.

II. *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.

III. *Association*

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

IV. *Subject Indicator*

Originally named as subject descriptor in ISO/IEC 13250, a subject indicator is a resource that is intended by the topic map author to provide a positive, unambiguous indication of the identity of a subject. When two topics use the same resource to indicate their subject, they are by definition "about" the same thing, and must therefore be merged during processing.

V. *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. If two topics use the same base name in same scope, and must therefore be merged.

To sum up, the nature of Topic Maps is very simple. The basic components, *Topics*, are clustered via the description of topic type. Then event-related topics are gathering together via the element *Occurrence*, and the semantics between related events are linked by the element *Association*. Besides, users could make advantage of the element *Scope* to limit the range of name, resources guide and relations to form the original entirety web resource.

Namely, Topic Maps can organize abstract knowledge content into a structure with coordinates and form a structured semantic web. A web page could be regarded as a unit gathering related information together, and link with others through hyperlinks. Being used for describing the web page contents, Topic Maps could be the navigator of the information contained in the web resources, and they could also reflect the structure of web resources. As Fig. 2 shows, users can search for web resources needed through description of metadata, besides, they could find out links between web resources and recompose them via the structure displayed by Topic Maps. When searching for a specific web resource, users don't have to search in a large database. With the links provided by Topic Maps, users could find out related topics and events.

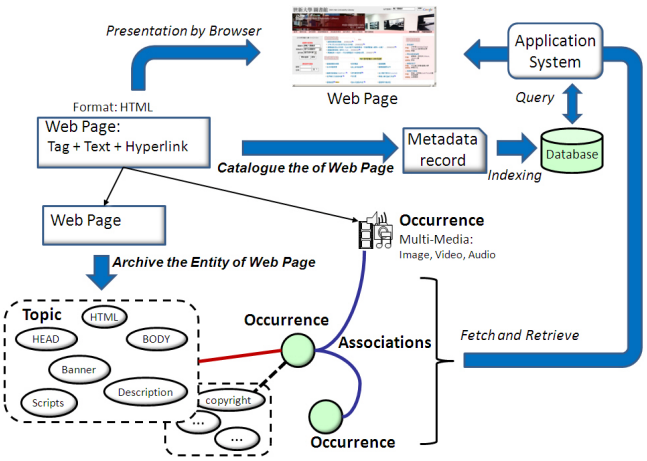

Figure 2. The Major process flow of this study

Because HTML markup is used for information presentation and not providing the structure of content, web pages written in HTML can't be searched accurately and the contents of them couldn't be use further. Considering that fact, the study tries to deconstruct HTML elements of web pages, describe and markup them as various Topics, and apply XTM elements *association* and *occurrence* to keep original link and related reference of web resources. There are several advantages of this strategy:

1. Providing the stability of long-term preservation of documents;
2. Offering the convenience of data reproduction and recomposing;
3. Realizing the feasibility of content mining;
4. Supporting the interoperation of systems;
5. Solving the problem of unstructured HTML document.

3.3 Deconstruction web page

Single web page shown on browser usually is a HTML document. As Fig. 3 shows, its content is within the range which <HTML> Tag declares. The HTML document could be divided into two parts: the <HEAD> Tag providing identification information for application software such as browser, searching engine, and the <BODY> Tag providing content represented by the web page (Smiraglia, 2005).

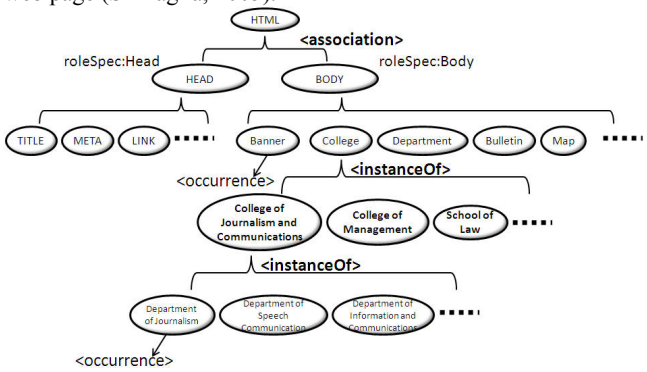

Figure 3. Deconstruction web page using Topic Maps example

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 Map, each Topic Map could be combined via <mergeMap> Tag of XTM. If a webpage is a combination of several web pages which associate with each other through <FRAME> Tag, the Topic Maps of single webpage could be combined via <mergeMap> Tag, too.

<BODY> Tag declares the actual content of a webpage resource. Distinguishing webpage 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 information presentation. Tags could be converted to independent Topics generally, and Tags with attribute could be recorded one by one using the <parameters> Tag under 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 webpage 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.

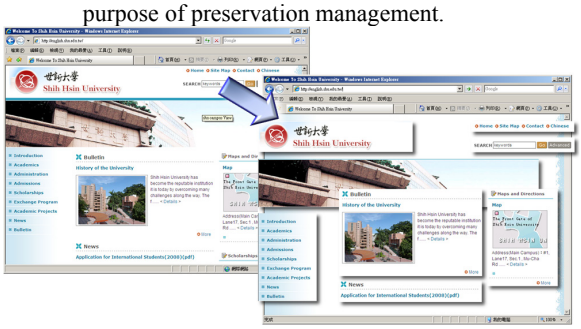
Figure 4. Deconstruction web page example

After analyzing the structure and the syntax of HTML document, the result of deconstruction of webpage is illustrated in Fig. 4. That offers archive application program to map the HTML text of each component to XTM syntax and convert the content of web pages to Topic Maps document. Additionally, this study use open-source tools developed by the Topic Maps for Java program to parse the webpage shown in Fig.5 and get the converted Topic Maps document. By presenting 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.

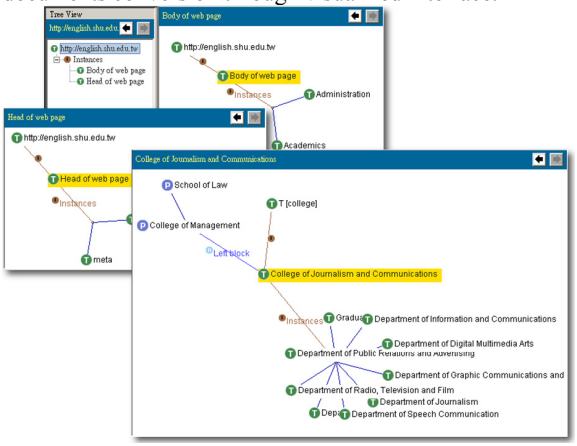

Figure 5. Apply TM4J tools to review the converted Topic Maps document

Basically, 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.

4. Conclusion

Be used for organizing the information of web resources in the long run, multilevel description model is an objective tool and with reasonable purpose. It works according to the archival principle of provenance, with the respect to creator's definition as well as the original structure of web resources, controls the levels of description, and represents the context in which resources are created. Besides, the contents of web resources are including text, image, sound, video, etc. For the purpose of future use of web resources, this study apply the system analysis method to deconstruct web pages, use RDF standards to describe the contents by their attributes and link the relationship through Topic Maps. Combing Topic Maps with the metadata records created through multilevel description model, the web resources would be preserved in an effective way and provided for sustained access.

References:

[1]Ahmed, Kal et al. (2001). *Professional XML Meta Data.* Birmingham, UK: Wrox. p.251-253.

[2]Coombs, James H. & Renear, Allen H. & DeRose, Steven J. (1987). "Markup systems and the future of scholarly text processing", *Communications of the ACM*, 30(11):933-947.

[3]Daconta, Michael C. & Obrst, Leo J. & Smith, Kevin T., (2003). *The Semantic Web*. Indiana: Wiley, pp.170-176.

[4]ISO JTC1/SC18/WG8 (1997). "Information Processing -- SGML Applications -- Topic Navigation Maps", available at http://www1.y12.doe.gov/capabilities/sgml/wg8/document/1860.htm

[5]Kal Ahmed (2002). "TM4J Developer's Guide", http://tm4j.org/tm4j/docs/devguide/

[6]Pepper, Steve(2000). "The TAO of Topic Maps: finding the way in the age of infoglut", available at http://www.gca.org/papers/xmleurope2000/pdf/s11-01.pdf

[7]Pepper,Steve & Moore, Graham (2001). "XML Topic Maps (XTM) 1.0", available at http://www.topicmaps.org/xtm/1.0/

[8]Pitti, D. V. (1999). "Encoded Archival Description: An Introduction and Overview", *D-Lib Magazine*, 5(11), available at http://dlib.ejournal.ascc.net/dlib/novemeber99/11pitti.html

[9]RSLP, "RSLP collection description model", available at http://www.ukoln.ac.uk/metadata/rslp/schema/

[10]SGML SIGhyper (1994). "A Brief History of the Development of SMDL and HyTime", available at http://www.sgmlsource.com/history/hthist.htm

[11]Smiraglia, Richard P. (2005). *Metadata: a cataloger's primer.* Binghamton, NY: Haworth, p.8-9.

[12]W3C (1998), "Extensible Markup Language (XML) 1.0: W3C Recommendation", available at http://www.w3.org/TR/1998/REC-xml-19980210

[13]Wang, Li-Chiao & Liao, Tsai-Hui(2007). "Mutual Proof between Max Weber's Theory of Bureaucracy and the Principle of Archival Provenance", *Bulletin of Library and Information Science*, 60:77-89.

[14]Wang, Li-Chiao (2007). "Web Archives: The Concept and Application of Multi-level Description Model", *Journal of Educational Media & Library Sciences,* 44(4):455-471.

[15]Yu, Shien-chiang & Chen, Ruey-shun(2001). "An XML framework for an electronic document delivery system", *The Electronic Library*, 19(2):102-110.