Jie Jin. NC Health Info and Go Local: An Analysis of Web Change Impacts on Metadata Quality and A Proposed Framework for Semi-Automatic Metadata Maintenance. A Master's Paper for the M.S. in I.S degree. April, 2008. 35 pages. Advisor: Jane Greenberg

This paper reports on a two-part study focusing on automatic metadata maintenance for NC Health Info, a health information web portal for North Carolina residents. The first examination is an analysis of web changes and their potential impacts on metadata quality. The second examination proposes a framework for semi-automatic metadata maintenance, which builds off of a hierarchy of evaluation proposed by Greenberg (2005). This work is supported by a baseline counting of the average number of Dublin Core Metadata elements on Dublin Core website and a conceptual framework enhanced with mathematical Set Theory. The results lead to a model that enables catalogers to review only a portion of the web resources with detected changes. The proposed approach aims to benefit the metadata generation and maintenance by reducing human effort and time consumption while ensuring high metadata quality.

Headings:

Web Change Detection

Consumer Health Website

Subject Metadata

Metadata Maintenance

Metadata Quality Evaluation

NC HEALTH INFO AND GO LOCAL: AN ANALYSIS OF WEB CHANGE IMPACTS ON METADATA QUALITY AND A PROPOSED FRAMEWORK FOR SEMI-AUTOMATIC METADATA MAINTENANCE

by
Jie Jin

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2008

Approved by

_____

Jane Greenberg

Table of Contents

1. Introduction

This study is based on the cataloging process of NC Health Info and Go Local (NCHI) which is a web portal to provide access to information of heath conditions, diseases, wellness and local health services, especially for the North Carolina residents. It is available at http://www.nchealthinfo.org/. This site is developed and maintained by Health Science Library (HSL) of University of North Carolina at Chapel Hill. Also, joint efforts between HSL and School of Information and Library Science, University of North Carolina at Chapel Hill have been made to improve the work of NCHI catalogers. The NCHI database stores various information including URL, Name, Phone Number, City, County etc. which are treated as metadata records for each web resource. Nevertheless, the knowledge of health issues is constantly changing due to new discoveries and important breakthroughs, which are always quickly disseminated on the web. In addition, health services are also considered highly dynamic in terms of rapid and frequent changes of service hours, locations and personnel who provide varying services etc. Therefore, coping with changes by ensuring the metadata quality and timeliness of its maintenance become the great challenge to the catalogers.

Based on the understanding of that manual metadata generation and maintenance is high labor-cost and time-consuming, which has been proved by the NCHI cataloging

operation, this study has conducted an analysis of web changes occurring to resources archived by NCHI so as to find out the metadata field that constitutes the greatest changes within the study period. Moreover, based on the result of this analysis, it proposes a conceptual application model for metadata maintenance using the concept of metadata quality evaluation. As a semi-automatic approach, this model attempts to increase the degree of automation and reduce the amount of time required by human while ensuring high quality metadata records. Specifically, there are two main tasks discussed in this study:

1) Among all the data tables where web resources' metadata records are stored, determine the field that constitutes the greatest change in the study period
2) Develop an application model for metadata quality evaluation at both empirical level and semantic levels so that a combinational evaluation score can be produced, on which the decision of the need of further human reviews will be based.

2. Literature Reviews

2.1 Web Change Detection and Study

World Wide Web is a highly dynamic space, which means the resources residing on the Internet involve great changes over time. In order to study such constant changes and to better manage them, several applications, such as WATZNEW (http://www.watznew.com), URL-Minder (www.urlminder.com) have been developed for tracking and viewing the changes. In addition to the application development, other

researches have been conducted under the same topic. As is discussed by Francisco-Revilla et al. (2001), many of the existing applications have the disadvantages of treating presentational changes as substantial relevant changes. In their study, they classified the nature of changes into categories such as content changes, presentational changes, structural changes, behavioral changes. Based on this classification, they attempted to develop an approach to automatically evaluate the relevance of changes in the web pages. However, their system, Path Manager, also had limitations in lacking support of indirection, which means frames embedded in the web pages cannot easily be handled; because, the application does not check URLs contained in the frames, unless these URLs appear as part of the path being checked. Also, monitoring JavaScript or other page behaviors was regarded as another challenge due to the complexity of web specifications adopted by various browsers to handle dynamic page objects. Therefore, human efforts still seem to be indispensable in such process. Flesca & Masciari (2002) also attempted to propose a technique to provide a change monitoring service on the web. The unique aspect of this technique was that it represents a web document as a tree and allows the users to concentrate on a particular part of the tree, which they call “sub-trees”. Then a query can be built up to check if a change to specific portion of the document has occurred. This approach was derived from the similarity measure between two documents using graph theory. The similarity of two trees (documents) is defined by considering the edge or association between two nodes on the two sub-trees of user’s interest. Such edge is expected to give the “maximum degree of similarity”. However, the

maturity of it was still limited to be at personal web updating level. Douglis (1997) and other fellow researchers conducted their study to evaluate the rate and nature of change of Web resources. Their results showed that the content type and rate of access have a great impact on the metrics used to measure the changes such as ratio of changes, access frequency etc., while size of the web resource had little effect. Koehler (2002) also confirmed the point that " Web pages change when either the content of the page is modified or when the hypertext links form the page are changed, added to, or deleted form the page. Using software, it is possible to ascertain whether change occurs, it is less easy to determine whether change is substantive or not. The determination of substantive change is inherently subjective." He concluded that the longevity of a web page is depending on its domain type and purpose; content pages have less survival rate than navigations pages.

2.1 Consumer Health Web Portal

More and more people are using the Internet to search for health-related information. According to Bendict (2000), about 52 million Americans look for health information online at least once a month, and 40% among whom show that this type of information affect their health care decision. More recently, according to health report written by Fox (2006), 80% of American internet users, or some 113 million adults, have now searched for health information online now. Apparently, the growth of online health information seeking is significant. Studies have been conducted to understand the characteristics of sites providing health information and their users' concerns and information seeking

behavior. Chin (2002) mentioned that the quality and reliability of health information on the web have become a great concern to the consumers. In their article, Luo and Najdawi (2004) conducted a review on how many of the health portals are making effort to use different measures to build the trust between the sites and consumers. They considered that the health information catalog, as one of the common features, is a mechanism for organizing health or medical treatment and services information and providing links to other health-related sites where consumers may further seek other information or obtain answers to their questions. This feature was available in most of health portals studied in their research. Lastly, they concluded that though maybe difficult and varied, trust-building measures need to be used by health portals to increase their trustworthiness to the consumers. MedlinePlus is a web-based consumer health information resource built by National Library of Medicine (NLM). It is widely considered as one the most authoritative health information portals. According to Miller, Lacroix and Backus (2000), with the rapid growth of using Internet for health information seeking, health portals become especially value and useful to the consumers. Facing high usage intensity, building and maintaining such service is challenging. The authors talked about various concerns in their study, including: Keeping information up to date with high quality and reliability as well as efficiently organizing and selecting the web resources are the goals of MedlinePlus's team. In particular, they described how the health topics are developed based on the analysis of users' search terms along with other statistics. As part of the content management process, such task is of great importance, yet combined with

difficulties due to the continuous growth of the availability of health information on the web. Cline and Haynes (2001) confirmed the fact of vastly increasing use of the Internet for health information seeking. Although providing "widespread access to information and advantages of interactivity, information tailing and anonymity…", inequity in accessing information, design features, information quality and evaluation skills have been perceived as problems. The authors advocated that the Internet needs to be viewed as "part of the larger health communication system" and "taken advantage of incorporating extant communication concepts". They have raised the significance of understanding consumers' Internet health information seeking behavior, considerations on potential benefits and quality concerns and identification of criteria for service evaluation. Besides discussing about the navigational difficulties caused by information overload, disorganization of web content, etc., authors also pointed out that, since there is hardly a proper filtering process, health information with low quality can be very dangerous to its users. Therefore, the best case will be ensuring the quality at its source where it is generated by authors or catalogers; and for evaluating such type of information, a consensus has to be reached that "Health-related websites should be judged by the quality of health information found on them and by design features that may facilitate or impede the use".

2.3 Web Resources Classification

Researchers have been attempting to explore methods for web resources classification to automatically label them with proper topics or subjects and specific

assignment of subject descriptor terms from a predefined controlled vocabulary. Text Mining is a typical technique implemented for this purpose. Given the machine power, the human subjectivity can be maximally avoided and degree of automation is expected to be increased. Golub (2003) gave a brief of different automated classification approaches, including text categorization, document clustering and document classification. The purpose of his study was to determine the utilization of controlled vocabulary in automated classification of textual web pages, in the context browsing. He attempted to classify web pages based on controlled vocabulary by using the term frequency and weight assignment. Yang and Lee (2005) proposed an automatic approach to generate semantically related labels/themes for web resources by using text mining techniques, a process including clustering, labeling and generation. However, besides necessary technical skills, the implementation of text mining approaches do require sufficient training data at its initial stage before the assignments can be done, this somehow presents a difficulty to its adopters.

2.4 Metadata Generation and Quality Evaluation

In this study, “Metadata Generation” is concentrated on subject metadata and to correctly classify web resources under different health concepts and assigning them with most relevant health topics. The quality of the metadata records is directly reflected by the how accurately these web resources are given the correct subjects. Unfortunately, in practice, it is proved to be hard to avoid the metadata records being subjectively or inconsistently generated by catalogers depending on their experiences and

domain-specific knowledge. With more understanding of automatic classification techniques, researches have been trying to use them for metadata generation. Liddy et al. (2002) applied Natural Language Processing and Machine Learning techniques to automatically generate metadata for educational resources and drew controlled vocabulary terms from he ERIC Thesaurus. Within this particular context, they have found the minimal difference between automatically generated metadata and manually assigned metadata during evaluation may be reached. Meanwhile, applications for the same purpose are being built. Dublin Core Metadata Initiative has a dedicated web page (http://dublincore.org/tools/) for available tools and software built based on Dublin Core standard. AMeGA project, Greenberg, Spurgin and Crystal (2006), was to explore and produce a list of recommended functionalities for applications supporting automatic metadata generation in the library / bibliographic control community through a survey conducting. In Greenberg (2004), her research results indicated the positive contribution of extraction and processing algorithms and harvesting metadata from META tags created by authors to automatic metadata generation. In Paynter (2005), besides talking about implementing text classification to assign metadata values from a given controlled vocabulary to the web resources, Paynter also described the development of an automatic metadata assignment tool, IVIA, which is used to give descriptive metadata to resources kept in virtual libraries, metadata repositories and digital libraries with reference to Library-Standard Metadata. This study covered both metadata assignment and evaluation using human and automatic approaches. As far as the metadata quality evaluation is

concerned, Bruce and Hillmann (2004) provided a comprehensive top-level list of metadata quality measures and metrics such as completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility. It has been mentioned that "automated metadata validation or evaluation is usually cheaper than human validation. Automated techniques potentially enable humans to use their time to make more sophisticated assessments. Cost-effective machine-based techniques represent 'the least we can do' to ensure metadata quality, possibly with more expensive human techniques following on..." In addition, Jane Greenberg has developed a layer cake of criteria for metadata quality evaluation in Greenberg (2005). Both Bruce and Hillmann (2004) and Greenberg (2005) are the fundamental principles where the proposed application model are based upon.

3. Research Goal

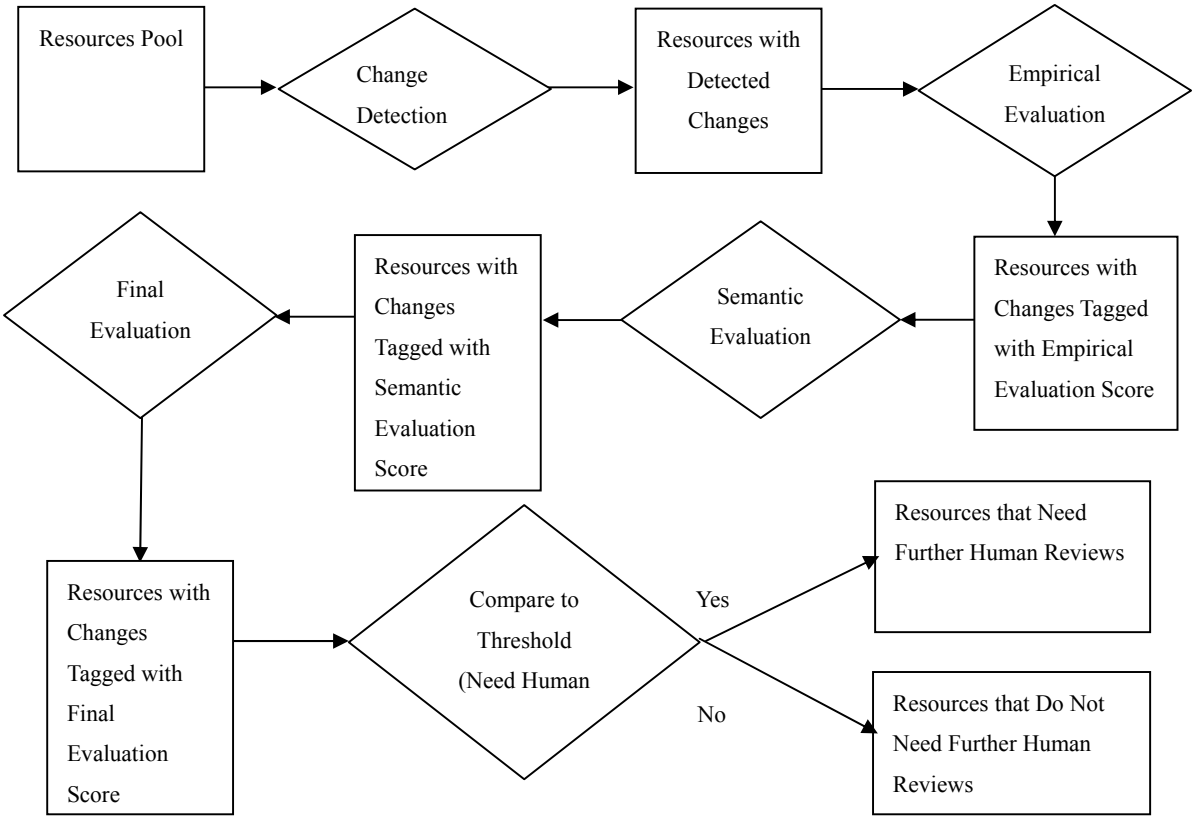
The goal of this study is to identify the aspect of web resources' metadata records structured in NCHI database that constitutes the greatest changes over the study period from December 5, 2007 to March 5, 2007, based on result of which, a conceptual framework is to be developed by using a set of metadata quality evaluation criteria as a semi-automatic approach for metadata maintenance. It is aimed that this research can benefit the metadata generation and updating operation as part of the whole cataloging process by ensuring the metadata quality and reducing the cost of human labor and amount of time consumption.

4. Methodology and Framework

The data collection period was from December 5, 2007 to March 5, 2007. There are 5 copies of databases have been archived on each of the following dates: December 5, 2007, December 27, 2007, Jan 15, 2008, February 5, 2008 and March 5, 2008, from which December 5, 2007, Jan 15, 2008, February 5, 2008 and March 5, 2008 are used to for this study. Although it is not the main focus of this study, a special feature, "Local Service Term - Health Topic Paring", used by NC Health Info site's content organization needs to be described for the benefit of understanding how metadata records are assigned to the web resources. NCHI staff have developed a list of service terms that are specifically used for describing health services in the State of North Carolina, while the list of Health Topics are used by MedlinePlus. Local Service Terms are qualified by pairings with appropriate Health Topics. Each web resources archived will be labeled with one or more parings. Based on this concept, there are two types of display pages in NC Health Info site: Local Service Term Display and Health Topic Display. Local Service Term display shows only resources cataloged to that term for a certain location, for instance, "Nursing Homes", then "Orange County"; on the other hand, the Health Topic display works the same way; users may choose "Breast Cancer", then "Orange County", the resultant display also includes resources necessary to manage Breast Cancer labeled by local service term such as Clinics, Oncologists and Support Groups. Hence, users are allowed to choose either of these two displays and then a specific location to reach the information they are looking for. In practice, although there is no direct search

function provided by NC Health Info site, the paring mechanism is sufficiently serving the purpose of facilitating the health information seeking through the site. The idea of the methodology can be summarized by Figure 1. This method for the case study of NCHI is at its conceptual and experimental level.

Figure 1: Proposed Framework



4.1 Capture and Study Web Changes

The Web is an information space with high changing rate and frequency. Moreover, in the consumer health field, the knowledge of health issues is constantly changing due to new discoveries and important breakthroughs, which are always quickly disseminated on the web. In addition, health services are also considered highly dynamic

in terms of rapid and frequent changes of service hours, services provided, locations and personnel. Therefore, in order for an efficient and effective semi-automatic metadata maintenance approach to be developed, the web changes occurring to the web resources collected by NC Health Info site have to be captured and studied. These “changes” are expected to be the changes that are substantively affecting the quality of metadata records generated by catalogers, which, without timely and accurate updating, may decrease the usefulness of the site. Thus, catalogers are monitoring web resources on a regular basis. As soon as there are content changes found, they will manually review them one by one. Based on their experience, if one change is deemed to be substantial, then a metadata record updating is executed to the database where the resources’ descriptive data is stored. This is considered as a part of metadata maintenance operation.

All of the archived databases are sharing the same schema in which a set of dedicated tables are created to hold metadata records used for web resources description. Among all the data fields, the URL, City, Phone_prefix, Phone_last_4_digits, Local Service Term – Health Topic Paring are empirically believed to be the fields that may constitute the majority of the changes. Therefore, in this study, the change capturing is focused on them. It needs to be kept in mind that each resource may have more than one city_id assigned to it to reflect its service locations and more than one parings gned to it to reflect the health topics its content covers. For convenience, the field-ideassintifprocess is completed through a simple approach: table to table comparison crossing all of the archived databases.

Table 1: Total Number of Unique Web Resources Collected by Each Database Archived

Date	Dec 5,2007	Jan 15, 2008	Feb 5, 2008	Mar 5, 2008
# of Records	5080	6206	6208	6208

Table 2: Fields Change Tracking for Dec 5,2007 to Jan 15,2007, Jan 15, 2008 to Feb 5, 2008 and Feb 5, 2008 to Mar 5, 2008

	Dec 5,2007 to Jan 15,2008		Jan 15, 2008 to Feb 5, 2008		Feb 5, 2008 to Mar 5, 2008	
	Number of Changes	Changes in percentage (%)	Number of Changes	Changes in percentage (%)	Number of Changes	Changes in percentage (%)
City Added	1774	7.92	2	1.14	12	7.27
City Removed	542	2.42	8	4.55	21	12.73
Paring Added	16354	73.04	42	23.86	74	44.85
Paring Removed	2807	12.54	93	52.84	18	10.91
URL	505	2.26	24	13.64	26	15.76
Phone Last Four Digits	240	1.07	4	2.27	8	4.85
Phone Prefix	170	0.76	3	1.70	6	3.64

Figure 2: Column Chart of Tracked Changes for 3 Time Intervals

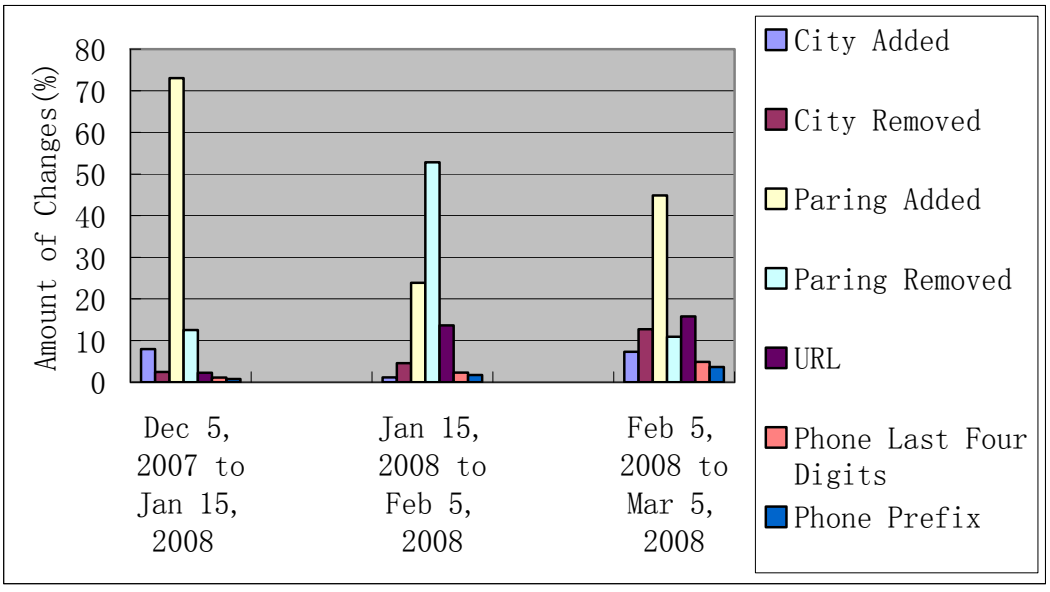
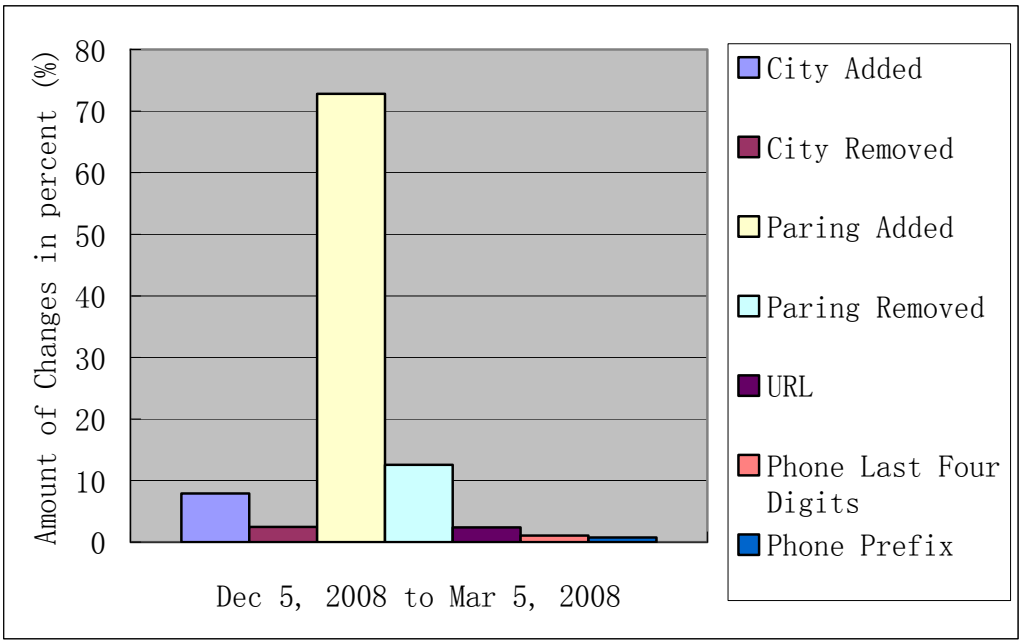


Table 3: Fields Change tracking from Dec 5, 2007 to Mar 5, 2008

	Number of Changes	Changes in percentage (%)
City Added	1791	7.90
City Removed	563	2.48
Paring Added	16505	72.79
Paring Removed	2849	12.57
URL	545	2.40
Phone Last Four Digits	247	1.10
Phone Prefix	174	0.77

Figure 3: Column Chart of Tracked Changes for Dec 5, 2008 to Mar 5, 2008



From the results above, it can be apparently seen that the paring is the data field that constitutes the most changes among all. The nature of this field is practically keywords for describing the main idea of the content. Hence, it is reasonable to regard these parings as descriptive metadata or subject metadata of each web resource. The function of this type of metadata is defined as "describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords…" Understanding Metadata (2004).

4.2 Metadata Quality Evaluations

After the portion of web resources with changes are captured, they will be fed to an application whose output is able to determine whether human evaluation is needed for a particular web resource based on a set of evaluation criteria. The conceptual model of this application is proposed in this study. The entire process of metadata quality evaluations consists of three sequential stages namely Empirical Evaluation, Semantic Evaluation and Final Evaluation Score Calculation, which is based on Metadata Evaluation Hierarchy discussed in Greenberg (2005). Each of Empirical Evaluation, Semantic Evaluation stages produces a score for the web resource and based these scores, a final evaluation score is calculated at the third stage and to be compared with a threshold so that the decision on whether catalogers will need to review the resources can be made.

4.2.1 Empirical Evaluation

The purpose of Empirical Evaluation is to measure the degree of a web resource's compliance to a particular metadata standard and furthermore, depending on the standard used, to determine its completeness and redundancy.

In this study, The Dublin Core Metadata Standard (www.dublincore.org) is selected due to fact that it provides a metadata element set intended to facilitate discovery of electronic resources. Dublin Core Metadata Standard was originally built for author-generated description of web resources, now it has gained the significant amount of attention of formal resource description communities such as museums, libraries,

government agencies and commercial organizations. 15 elements are defined in Dublin Core Metadata Element Set including: Title, Creator, Subject, Description, Publisher, Contributor, Date, Resource Type, Format, Identifier, Source, Language, Relation, Coverage, Rights. There are two important steps for this level of evaluation: 1) Dublin Core Metadata Element(s) Generation and 2) Empirical Evaluation Score Calculation.

1) Dublin Core Metadata Element(s) Generation

In order to be able to automatically extract the elements that are used by a web resource, a free online application named **DC-dot** (http://www.ukoln.ac.uk/metadata/dcdot/) is employed. It is listed as one of the utilities for metadata generation and extraction on Dublin Core Official website (http://dublincore.org/tools/). The function of DC-Dot is to "retrieve a Web page and automatically generate Dublin Core metadata, either as HTML <meta> tags or as RDF/XML, suitable for embedding in the <head>...</head> section of the page…" (http://www.ukoln.ac.uk/metadata/dcdot/). By using DC-Dot's extraction and generation results, it is easy to determine the number of Dublin Core elements that a certain web site can be generated from.

2) Empirical Evaluation Score Calculation

What was surprisingly found in this study is that it seemed to be impractical to assume if a web resource can be extracted or generated with 15 elements by DC-Dot, it will get a score of 100%; because, after running DC-Dot against the commonly used pages on www.dublincore.org which can be reasonably believed to have full

implementation of Dublin Core Metadata Standard, it was found that there has been none of these pages have entirely 15 elements extracted or generated. The result is shown in the Table 4

Table 4: Number of Dublin Core Elements Extracted or Generated from Web Pages on Dublin Core Official Site

URL	Number of Dublin Core Elements Generated
http://www.dublincore.org/	**8**
http://www.dublincore.org/documents/dcmi-terms/	**7**
http://www.dublincore.org/documents/dces/	**9**
http://www.dublincore.org/documents/abstract-model/	**8**
http://www.dublincore.org/resources/expressions/	**9**
http://www.dublincore.org/schemas/	**9**
http://www.dublincore.org/documents/usageguide/	**7**
http://www.dublincore.org/tools/	**9**
http://dublincore.org/news/2008/	**9**
http://dublincore.org/groups/	**9**
http://dublincore.org/about/	**9**
http://dublincore.org/projects/	**9**
http://dublincore.org/translations/	**8**

We can see that the average number of Dublin Core terms extracted or generated for this set of commonly accessed pages on www.dublincore.org is approximately 8. Therefore, it is proposed that a web resource will be given a score 100% at Empirical Evaluation level if the number of unique Dublin Core Metadata Elements it can be extracted or generated with is equal to or more than 8. In order for this concept to be illustrated in a more detailed way, one example is presented to show how it works in practice. A health-related sample web site, http://acra-org.com, is randomly picked from NCHI database. This

particular website is created for AIDS Community Residence Association. The DC-Dot result is shown below where Dublin Core elements are bolded.

Figure 4: The DC-Dot Extraction and Generation Result for http://acra-org.com

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />

<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />

<meta name="**DC.title**" content="ACRA - AIDS Community Residence Association" />

<meta name="**DC.subject**" content="Home; Case Management; Links; Housing; Information For Clients; Newsletter; Contact Us; Frequently Asked Questions; How Can I Help" />

<meta name="**DC.date**" scheme="DCTERMS.W3CDTF" content="2007-09-15" />

<meta name="**DC.type**" scheme="DCTERMS.DCMIType" content="Text" />

<meta name="**DC.format**" content="text/html" />

<meta name="**DC.format**" content="5225 bytes" />

<meta name="**DC.identifier**" scheme="DCTERMS.URI" content="http://acra-org.com" />

There are 6 unique Dublin Core elements for this health information site. Hence, for the purpose of Empirical Evaluation, the web resource will get a score of 6/8 that is 75%.

4.2.2 Semantic Evaluation

As is perceived, both knowledge of health issues and health service information is constantly changing and can be quickly distributed over the Internet. In the proposed application model, once a content change occurring to the web resource has been detected, its metadata records existing in NCHI database is no longer accurate. Therefore, besides being evaluated empirically, the resource will be also assessed semantically. Extending

from the result in 4.1, the proposed application model at semantic level in this study is to mainly deal with subject metadata changes occurring to the web resources. The reason of applying Semantic Evaluation is to determine how accurately terms picked from existing controlled vocabulary are assigned to the web resource as its metadata records. In the case of NC Health Info website (NCHI), when there is a change occurred to a web resource in terms of addition, removal or alteration of its content. A certain web resources will be first assigned one or more pairings of local service terms and health topics, this set will then be compared with the existing set of parings given to this web resource. Lastly, a score of Semantic Evaluation is to be calculated based on the result of the comparison. Figure 5 and Figure 6 demonstrate how content changes affect the subject metadata assignment in the case of content addition. A sample site, http://www.myocare.net, is randomly chosen from NCHI database and artificial changes are created solely for demonstration purpose in this paper. They do not reflect what happens to the site in reality.

Figure 5: Sample Site's Main Page at Time Point 1

Based on "Our Mission" on the left navigation bar as well as other information provided by the website, the catalogers assigned following local service – health topics parings as subject metadata to this piece of resource:

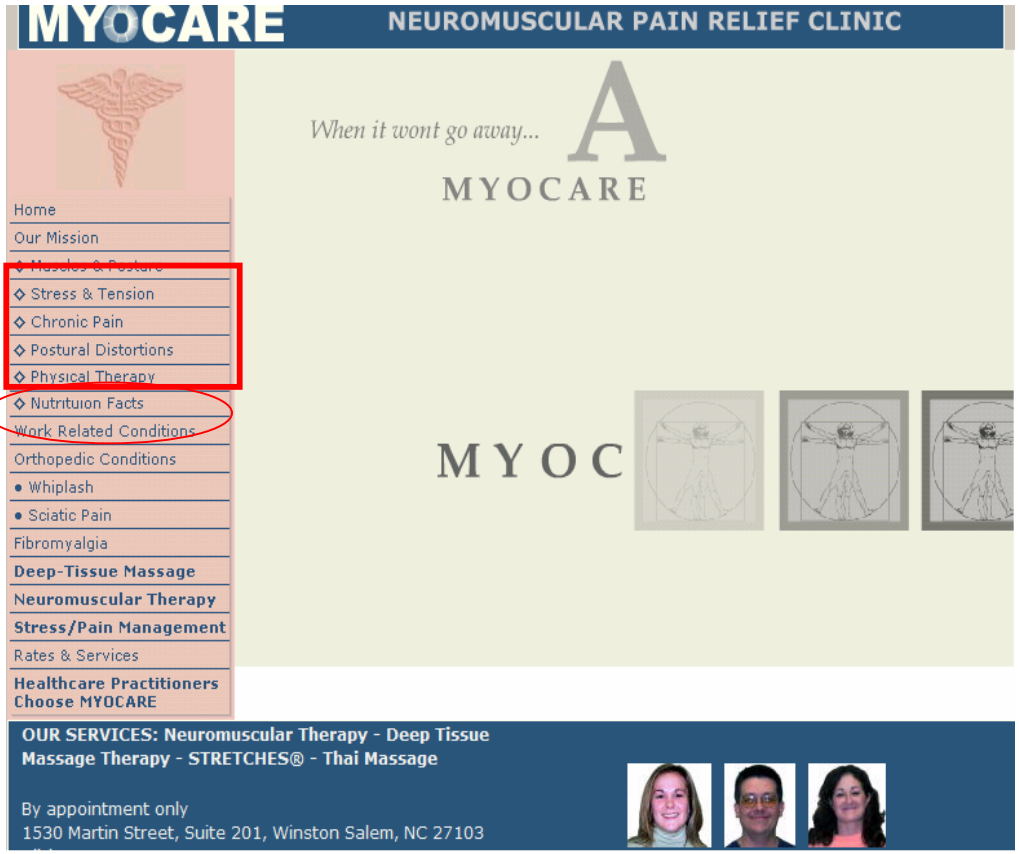
- Massage/Bodywork Therapists -- Stress
- Massage/Bodywork Therapists -- Pain
- Massage/Bodywork Therapists –Fibromyalgia
- Massage/Bodywork Therapists –Complementary and Alternative Medicine
- Clinics – Health Facilities
- Clinics – Fibromyalgia

- Clinics – Stress
- Clinics – Pain

However, after certain period of time, there may be two new terms are added to "Our Mission" shown in Figure 6 to reflect the service expansion of the site and the rest of the content remains the same. With the addition to the content, it is certain that new subject metadata will need to be added to the metadata records of this web resource, which could be the following pairings:

- Health Education – Nutrition
- Physical Therapist – Pain

Figure 6: Sample Site's Main Page at Time Point 2 (Change is Detected)



There are two important steps in this process: 1) Assignment of Subject Metadata and 2) The Semantic Evaluation Score Calculation based on the comparison result.

1) Subject Metadata Assignment

Certainly, Text Mining techniques can find their role here to automatically assign subject metadata to the web resources, when changes occur. However, how accurate this process can be is still under researchers' speculation. In this study, it is believed that a straightforward and simple word for word / phrase for phrase mapping can be used to automatically select the terms or health topics from the existing list for topic annotation purpose. There are three potential matching rules are proposed in this study so that a match between terms in content and terms in the predefined list of health topics

- *Exact Matching*: a word or phrase will match to a member in the predefined list of health topics. For instance, "Drug Abuse" in the content matches to the "Drug Abuse" in the lists.
- *Partial Matching*: a word or phrase will not completely match to a member in the predefined list of health topics, but partially. For instance, "Massage Therapy" in the content partially matches to "Massage / Body Therapist" in the lists.
- *Stemming Matching*: with the availability of many existing stemming algorithms such as Porter Stemming Algorithm (http://tartarus.org/~martin/PorterStemmer/), it is feasible to stem all the words in both content and lists which does require certain amount of computational power so that a match may be found if two terms are sharing the same root. For instance: "Clinical" in content matches to "Clinics" in the predefined

list of health topics.

Moreover, it is highly likely that web resources are using terms that do not appear exactly the same as those in the predefined lists while sharing an identical meaning. Hence, the use of synonym list has been brought to the scope of this study. The necessity of including such a list is to increase the ability of the application for capturing the extensiveness of the resource content in practice. For instances, “Bone Specialist” is considered as a synonym of “Orthopedists”, “Bodywork Therapists” is considered as a synonym of “Massage Therapists”. Thus, the use of synonyms is a special case of Exact Matching Rule. Therefore, as along as a certain term in the web resource content follows any one of the matching rules(Exact Matching, Partial Matching, Stemming Matching) or falls into the synonym list, the content where this term appears will be labeled with corresponding member of predefined list of health topics.

2) Semantic Evaluation Score Calculation

At this step, the method to calculate Semantic Evaluation Score Calculation can be explained by borrowing the concept of Set Theory. For a particular web resource R with detected changes, we can assume Set A represents the set of existing subject metadata generated previously for describing R and Set B represents the set of subject metadata generated through the assignment process in Step 1. Then we can count the number of elements in both Set A and Set B, which are denoted as N_A and N_B, the number of elements in their intersection set, which is denoted by $N_{A \cap B}$ and their union set which is denoted by $N_{A \cup B}$. This score is essentially representing, for a particular web

resource R, the ratio of the number of set elements shared by the existing set of subject metadata(Set A) and the newly generated set of subject metadata(Set B) to the total number of set elements of the union of Set A and Set B. This is can be represented as Figure 7.

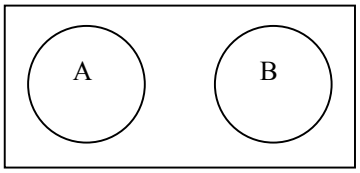
Figure 7: Semantic Evaluation Score for R

$$\frac{N_{A \cap B}}{N_{A \cup B}}$$

There are possibly four cases that may take place in this step:

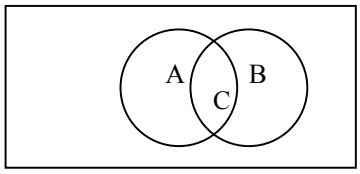
Case 1: The Intersection of Set A and Set B will be an empty set, denoted by $A \cap B = \emptyset$, which is shown in Figure 8. For this case, it means there are no overlapping subject metadata terms shared by both sets. Thus, the Semantic Evaluation Score of this web resource is 0.

Figure 8: Case 1



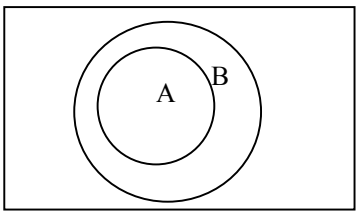
Case 2: There is a overlapping part between Set A and Set B shown in Figure 9. For this case, it means there is certain number of subject metadata terms shared by both sets.

Figure 9: Case 2



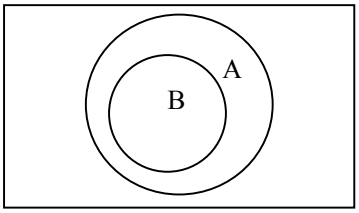
Case 3: The Intersection of Set A and Set B is Set A, denoted by $A \cap B = A$, which is shown in Figure 10. For this case, it means the set of newly assigned subject metadata includes all the members of the subject metadata set previously assigned to the resource. Mostly, it is due to content addition.

Figure 10: Case 3



Case 4: The Intersection of Set A and Set B is Set B, denoted by $A \cap B = B$. which is shown in Figure 11. For this case, it means the set of previously assigned subject metadata includes all the members of the subject metadata set newly assigned to the resource. Mostly, it is due to content removal.

Figure 11: Case 4



The following example is to illustrate how Semantic Evaluation Score Calculation works in practice.

Case 1: Set A = {Topic A, Topic B},

Set B = {Topic C, Topic D}

$N_{A \cap B} = 0$ and $N_{A \cup B} = 4$

Score = 0 = 0%

Case 2: Set A = {Topic A, Topic B, Topic C},

Set B = {Topic B, Topic C, Topic D, Topic E}

$N_{A \cap B} = 2$ and $N_{A \cup B} = 5$

Score = 2 / 5 = 40%

Case 3: Set A = {Topic B}

Set B = {Topic A, Topic B}

$N_{A \cap B} = 1$ and $N_{A \cup B} = 2$

Score = 1 / 2 = 50%

Case 4: Set A = {Topic A, Topic B, Topic C, Topic D}

Set B = {Topic A, Topic B}

$N_{A \cap B} = 2$ and $N_{A \cup B} = 4$

Score = 1 / 2 = 50%

4.2.3 Final Evaluation Score

As the last step where the need of further human evaluation will be decided on, the Empirical Evaluation and Semantic Evaluation Scores are combined to produce the Final Evaluation Score. The formula shown in Figure 12 will be used for this calculation.

Figure 12: Final Evaluation Score Calculation

$$\frac{\text{Empirical Evaluation Score in percentage} + \text{Semantic Evaluation Score in Percentage}}{100\% + 100\%}$$

For example, if a web resource gets an Empirical Evaluation Score of 60% and a Semantic Evaluation Score in of 80%, its final score will be (60% + 80%) / 200% = 0.7(70%). Finally, a threshold value will be determined by users who are adopting this framework, one potential method is to use average final score of all the web resources with detected changes. This value will be taken as a cut-off point and catalogers will need to manually review those web resources have the scores lower than this cut-off point.

5. Future Research

Some aspects of the study can be furthered in the future. The proposed application model for semi-automatic metadata maintenance is still at its conceptual level. It will be necessary and helpful to build a working version so that more testing and data analysis of its efficiency and accuracy can be conducted. The study of web changes still presents a great space for researchers; for instance, how to distinguish presentational changes from substantial changes is still an important issue. Specifically for websites providing health information that are constantly subject to the content changes, researches of how to make correct judgment on the substantiality of changes will be extremely valuable. Moreover, in this study, there are three matching rules are proposed are they are treated equally, which may be considered as a limitation at some point; because, the nature of different rules may actually affect the accuracy level of subject metadata assignment. Thus, weights may be assigned to each type of matching rules to differentiate them. Also, the

score calculation methods discussed in this paper is still subject to adjustments. Users who would like to adopt this approach may add some other measures or adjust how these score are produced based on their own needs and acceptance level so that they can more accurately and realistically determine the need of human effort for metadata quality evaluation. Furthermore, for the same purpose of developing semi-automatic approaches, instead of working at data or element level, usability studies from an interface design angle can also be potentially implemented to the current cataloging system so that visual enhancements with the functionalities such as keyword highlighting may be added to the existing system to ease the human work.

6. Conclusion

There is no doubt that more and more people have been using Internet to seek for health-related information. And this is one of the important reasons of the birth and development of NC Health Info (NCHI) site. Ensuring high content quality and building the firm trust are deemed to be the key to the success of consumer health website, such as NCHI. However, the rapid growth of the Internet and high frequency of web content changes have presented a great challenge to the catalogers at NCHI in generating and maintaining the metadata records of web resources archived in the database. They are required to properly update and assign relevant health subjects to the resources in a timely manner so that users can always easily find the latest health information they are looking for. Unfortunately, the existing manual approach for metadata generation adopted

by NCHI catalogers is considered to be high labor-cost and time consuming. This study is to inject the computational power to metadata quality evaluation as part of the metadata generation and maintenance process.

Based on the case study of NCHI web site, it has been confirmed that among all the fields used to describe a particular health-related web resource, the one representing the health topics which are considered as subject metadata records constitutes the most changes over the study period, and therefore, requires the most maintenance effort. Therefore, a semi-automatic approach is proposed from a metadata quality evaluation angle, which is essentially a conceptual model of an application performing metadata quality examination at empirical and semantic levels. The output of this model is an evaluation score assigned to each web resource that is to be compared to a threshold. The catalogers will only need to review the web resources that are scored lower than the threshold. Although at conceptual level, this application is expected to be able to reduce the amount of subjectivity, inconsistency and time consuming involved in the human metadata generation and maintenance process while ensuring high quality of the metadata records.

7. Bibliography

Bendict, C. (2000, December 11). E-health: Act 2. *Los Angeles Times.*

Bruce, T. R. & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann & E. L. Westbrooks (Eds.) Metadata in Practice (pp.238 - 256). Chicago, IL: American Library Association.

Bang, D.L., Farrar S., Sellers, J. W., & Buchanan, D. H. (1997). Consumer Health Information Services: Preliminary Findings About Who Is Using Them. *Journal of Medical Systems*, 22(2), pp. 103-115.

Chin, T. (2002, June 17). Patients Put Trust in Internet Health Information. *American Medical News.*

Cline, R. J. W., & Haynes, K. M. (2001). Consumer Health Information Seeking on t he Internet: The State of the Art. Health Education Research, 2001 December, 16(6), pp. 671-92.

Douglis, F., Ball, T., Chen, Y., & Koutsofio, E. (1998). The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web. *World Wide Web*, 1(1), pp. 27 -44.

Douglis, F., Feldmann, A., Krishnamurthy, B. & Mogul, J. (1997). Rate of change and other metrics: A live study of the World Wide Web. *USENIX Symposium on Internet Technologies and Systems*, pp.147-158.

Flesca, S., & Masciari, E. (2002). Efficient and effective Web Change Detection. Data & Knowledge Engineering. August, 2003, 46(2), pp. 203 – 224.

Fox, S. (2006). Report on Online Health Search 2006. PEW Internet and American Life Project. October 29, 2006. Website: http://www.pewinternet.org/PPF/r190/report_display.asp

Francisco-Revilla, L., Shipman F., Furuta R., Karadkar U., &Arora A. (2001). Managing Change on the Web. *In Proceedings of the 1st ACM/IEEE-CS Joint Conference on*

Digital Libraries. June 24-26, 2001, Roanoke, Virginia.

Golub, K. (2003). Using Controlled Vocabularies in Automated Subject Classification of Textual Web Pages, in the Context of Browsing. *Bulletin of the IEEE Technical Committee on Digital Libraries, 2(2)*. Retrieved October 29, 2005, from http://www.ieee-tcdl.org/Bulletin/v2n2/golub.golub.html

Greenberg, J., Spurgin, K. & Crystal, A. (2006) Functionalities for automatic Metadata Generation Applications: A Survey of Metadata Experts' Opinions. *International Journal of Metadata, Semantic and Ontologies*, 1(1): pp. 3-20

Greenberg, J. (2004) Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. *Journal of Internet Cataloging*, 6(4): 59 -82

Greenberg, J. (2005). Metadata Quality: A Hierarchy of Criteria. *In the Proceedings of the American Society for Information Science and Technology, Charlotte, North Carolina, October 28-November 2, 2005, 42(1).*

Koehler, W. (2002).Web Page Change and Persistence – A Four-Year Longitudinal Study. Journal of the American Society for Information Science and Technology. January, 2002, 53 (2), pp. 162 - 171 .

Liddy, E. D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Diekema, A., et al. (2002). Automatic Metadata Generation & Evaluation. *In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Luo, W. H., &Najdawi, M.(2004). Trust-building Measures: A Review of Consumer Health Portals. Communications of the ACM. January, 2004, 47(1), pp.: 108 – 113.

Miller N., Lacroix, E., & Backus, J. E. B. (2000) MEDLINE*plus*: building and maintaining the National Library of Medicine's consumer health Web service. Bull Med Libr Assoc. January, 2000, 88(1), pp.11–17.

Paynter, G. W. (2005). Developing practical automatic metadata assignment and evaluation tools for internet resources. *InProceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries.*

Understanding Metadata. (2004). Bethesda, MD: NISO Press: http://www.niso.org/standards/ resources/UnderstandingMetadata.pdf.

Yang, H. C., &Lee, C. H. (2005) Automatic Metadata Generation for Web Pages Using a Text Mining Approach. *In Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05).*

Yeh, P.J., Li, J.T., &Yuan, S.M.. (2001). Tracking the Changes of Dynamic Web Pages in the Existence of URL Rewriting. *In Proceedings of the Fifth Australasian Conference on Data Mining and Analystics* Vol. 61**,** pp. 169 - 176 , Sydney, Australia.