Cataloging On-Line Health Information: A Content Analysis of the NC Health Info Portal

Catherine Blake[1&2], PhD, David West[1], Lili Luo[1], Gary Marchionini[1], PhD

[1]School of Information and Library Science, University of North Carolina, Chapel Hill
[2]Lineberger Cancer Center, University of North Carolina, Chapel Hill

Abstract

The unrelenting increase of health information on the World Wide Web has resulted in an urgent need for portals that provide consumers with trustworthy health information. In response to this need, the National Library of Medicine initiated the *Go Local* initiative, which extends MedlinePlus by providing consumers with links to local health services, programs and providers. NC Health Info (www.nchealthinfo.org) is the first NIH funded *Go Local* portal. Our goal is to gain insight into the nature of interactions that occur during the cataloging process of online health information resources. We conducted a content analysis of annotations made by catalogers on the NC Health Info portal between January 2000 and September 2004. Our analysis of 2369 online information resources revealed challenges with establishing the navigational, geographical and topical content of an on-line resource. Our analysis provides insights into the mechanisms that catalogers use to overcome those challenges and thus will be of value to future *Go Local* portal development.

Keywords:

Consumer Health Information, *Go Local*, NC Health Info, MedlinePlus, Annotation, Cataloging online material.

Introduction

Distributing information using the World Wide Web has never been easier. Although access to information can empower a consumer to make informed choices regarding their health care, the quantity of information often leaves consumers feeling inundated.

The National Library of Medicine (NLM) plays an active role in the provision of health information to consumers. MedlinePlus , which was launched in October 1998, typifies the NLM's commitment to providing trustworthy, well-organized health information [1]. The MedlinePlus critieria includes (1) the quality, authority and accuracy of content; (2) the primary purpose of the Web page (i.e. educational and not to sell a product or service), (3) the availability and maintenance of the Web page and (4) special features, for example providing content that is accessible to persons with disabilities[1].

The *Go Local* intuitive (www.nlm.nih.gov/medlinplus/golocal.html) augments information in MedlinePlus with local health services, programs, and health care providers. The first *Go Local* portal was funded in August 1999 at the University of North Carolina, Chapel Hill as a joint project between the Health Sciences Library and the School of Information and Library Science. The next portal was created at the University of Missouri, and since then *Go Local* portals have been initiated in Alabama, Arizona, California, Indiana, Maryland, Massachusetts, Michigan, Ohio, Texas, Tribal Four Corners (Arizona, Colorado, New Mexico and Utah), and Wyoming.

Several studies have explored the trustworthiness of health information on the WWW [2-4]. A cataloger on the NC Health Info portal fulfills two roles. The first is to ensure that the information resource (which we refer to as a web page) is trustworthy. Figure 1 captures the primary inclusion criterion used by catalogers for a *Go Local* portal. This inclusion criterion reflects both MedlinePlus and *Go Local* criteria.

In addition to trustworthiness, the catalogers on the NC Health Info portal assign terms from the *Go Local* controlled vocabulary. Although several projects have developed search engines specifically designed for health information, the controlled vocabularies in both MedlinePlus and the *Go Local* portals still play an important role in enabling a user to identify health information on the web.

In this paper, we characterize the communication patterns that occur between catalogers as they assign terms from the *Go Local* controlled vocabulary to each web page that satisfies the quality criterion shown in Figure 1. Our goal in this paper is to characterize the challenges faced by catalogers. Findings from this study will inform the design and development of automated tools that support catalogers as they provide the meta-data necessary for the *Go Local* portal initiative.

---

[1] Complete criteria are available from www.nlm.nih.gov/ medlineplus/criteria.html

Authority of source
- The sponsorship of the site is clear.
- There is a way to contact the site.
- Sites for an individual health care provider must include credential information.
- If the site is commercial, it acknowledges any commercial interest or personal point of view.

Content
- Pages contain a created, revised, or update date.
- Links on the site are reliable and relevant.
- Information on the site is unique and not readily available elsewhere in the database.
- The site does not contain inaccurate, erroneous, misleading or dangerous medical information, claims, or allegations.
- Most information is available at no charge.
- Registration, an account, or password is not required to access site information.

Audience
- The intended audience of the site is consumers

Local Relevance
- The site provides information about a local or regional organization, service or activity.

Figure 1 – Selection Guidelines from Appendix 1 of the *Go Local* Input System Training Manual, Version 1.2[1]

Materials and Methods

The NC Health Info project provided a snapshot of the trustworthy web pages that their team of catalogers had collected between 1 January, 2000 and 29 September, 2004. Of particular interest is the "note" field in the database that the catalogers began using in April 2002. Each web page has one or more notes, which we also refer to as an annotation[2]. Annotations are either substantive messages between catalogers, or non-substantive system messages; i.e., these are meant as part of the cataloging process rather than for the public.

Figure 2 provides an example of the cataloger's annotations for web page 46. On May 31, 2002, a cataloger working on the NC Health Info project established that web page 46 satisfied the authority, content, audience, and local relevance criterion shown in Figure 1. On the same day cataloger AA posed a question about the scope of the geographical location of this service (line 2), and on June 11, 2002 cataloger BB responded to AA's question (line 3). In September 2002, cataloger CC updated the geographical locations that were associated with the web page (line 4). The page was again updated in March 2003; however, the cataloger information for that update is not available (line 5). In April and October 2003 (line 6 and 7), the system proposed that the site should be reviewed, but the catalogers found that no changes to the database record were necessary. In April 2004, DD again reviewed the site with respect to the geographical location and posed a question (line 8), which cataloger FF answered (line 10) three days later.

1 5/31/2002   Original Record
2 5/31/2002   Women's Breast Health Center, Iredell Home Health (many served counties listed...should they be included on this page, or just on a child page for the home health itself (AA)
3 6/11/2002   Let's assume the hostpital serves the counties listed on the home health page, and use one record to reflect all aspects of the site (BB)
4 9/13/2002   I added a few more topics then approved the site (CC)
5 3/19/2003   Took out cataloging for Birth Center and made new record.
6 4/2/2003 (DD)
7 10/13/2003 (EE)
8 4/9/2004 DD:I do not feel that all of these counties should be listed just because the Home Health Service serves these counties because if you read the hospital mission they serve Iredell and Alexander counties. What do you think? (FF)
9 4/9/2004 (FF)
10 4/12/2004 FF: see this from their mission statement: Iredell Memorial Hospital was established to provide quality health care to citizens of Iredell County. In recent years, this mission has expanded to all contiguous counties. By carrying out this mission, the hospital has taken a leadership role in the provision of health care and health promotion programs for the citizens of Iredell County, Alexander County, and citizens of other counties who may come to the hospital for care or utilize its services at some other location. I think we should leave the contiguous counties in and delete any others. (DD)

Figure 2 – An example of annotations associated with web page 46 in the NC Health Info portal from May 2002.

The example database records shown in Figure 2 demonstrate the rich set of interactions that occurred between at least six catalogers over the two year period. Annotations 2-5, 8, and 10 capture interactions between catalogers as they work through the process of assigning the initial health services terms and topics, and as they continue to maintain the web pages. The maintenance schedule for the NCHealthInfo project is currently every six months. The system generated annotations 1, 6, 7 and 9 automatically.

In general, the kappa statistic is used to report inter-rater reliability [5]; however, studies of on-line web pages have shown that researchers rarely provide inter-rater reliability [3]. The example in Figure 2 demonstrates that catalogers seek consensus before assigning a controlled vocabulary term. Thus, the kappa statistic would not provide insight into the nature of the interactions shown in Figure 2 because disagreements are resolved as part of the cataloging process.

In contrast to the kappa statistic, a qualitative analysis can provide insight into the challenges faced by catalogers as they assign terms from the *Go Local* controlled vocabulary to each web page that satisfies the quality criterion. We used content analysis to characterize the nature of disagreements, such as the discussions shown in Figure 2.

Table 1 captures the content, format and function facets that we considered during the content analysis. These facets and categories were developed from our initial pilot study that comprised a random sample of 464 web pages (20%) from the web pages in the NC Health Info database. Removing non-substantive annotations yielded 371 substantive messages.

Two of the authors (LL and DW) characterized each of those 371 annotations, and Table 1 shows the categories that emerged. Once inter-rater reliability was established between the two authors, they labeled the remaining pages using these eleven categories shown in Table 1. The categories in each facet are not mutually exclusive, thus any given substantive annotation can have multiple categories assigned. Every annotation has at least one category for each facet.

Table 1 – Facets used during the Content Analysis

Facet	Category
Content	*Navigation:* issues involved in navigating and accessing the web page
	Geographic Scope: issues with defining the geographical scope of the web page
	Topical Scope: issues with defining the topical scope of the web page
	Miscellaneous: issues related to website cataloging that fall into none of the above categories
Format	*Question:* Annotation is formatted as a question, or can be reasonably inferred as a question.
	Answer: An annotation explicitly in response to a question and comments which probably answer unasked questions
	Statement: Declarative statements
Function	*Log of Action:* A statement of an action taken in the past.
	Reminder: A statement to remind catalogers of actions that should or should not be taken in the future and relevant information that they should notice in the future.
	Reach Consensus: A statement made in the process of reaching an agreement on a disputed point.
	Action Request: A comment that request a cataloger to take an action or provide information

The annotations for web page 46 (shown in Figure 2) reflect the content, format, and function facets. The catalogers discuss content, specifically the geographical scope in lines 2, 3, 8, and 10; and the topical scope in lines 4 and 5. Lines 2 and 8 are examples where the format of the annotation is a question, line 3 is an example of an answer format, and lines 4, 5 and 10 are examples of declarative statements. The function of lines 2, 3, 8 and 10 are to establish consensus, whereas lines 4 and 5 serve as a log of actions taken in the past.

Results

Catalogers added 2788 distinct web pages to the NC Health Info portal between Jan 2000 and Sept 2004. The following content analysis includes 2369 of the web pages. There were 10,462 annotations for these 2369 pages, of which 2301 (22%) captured interactions between catalogers. The number of substantive annotations per web page ranged between one and eight. For example, there are six substantive annotations associated with the web page shown in Figure 2. The average number of annotations made per web page is 3.02. The average number of substantive annotations per web page is 1.82; thus, there are an average of 1.20 system generated annotations per web page.

Content Analysis – Pilot Study

The content analysis from the pilot study indicated that most of the annotations related to establishing the topical scope of a website (n=192) and website navigation (n=147). A large number (n=266) of annotations took the form of a statement while 109 were posed as questions and 97 as answers. As for functions, 99 logged the cataloger actions, 29 were reminders for the cataloging team, 181 were requests for other catalogers to take an action or provide information, and 174 were messages exchanging ideas and reaching consensus on solving a particular problem arising from the cataloging process.

The pilot study indicated that 97 of the annotation fields comprised at least one round of discussion with regard to properly cataloging the website. Such consensus building is necessary to avoid low levels of inter-rater reliability with respect to the final catalog decision. This finding suggests that software tools that support collaboration between catalogers would enable catalogers to reach consensus in on- or off-line environments.

Content Analysis – Full Study

The content, format, and function facets capture the nature of the interactions between catalogers who worked on the NC Health Info portal during Jan 2002 through Sept 2004. We consider only entries after 2002, where the catalogers first started to use the optional annotation field.

Figure 3 captures discussions with respect to the content of the web page. In most cases, the catalogers discussed the appropriate *Go Local* topic for the page (n=1165, 47.1%). Catalogers also discussed navigation issues (n=712, 28.8%) associated with the web pages. For example, on 13 October 2003, a cataloger made the following annotation to web page 212: "This site contains links to many services on their homepage. Because we have separate records for "Birthing Center" and "Rehabilitation Services", I think separate records for these other services should be created - especially since some services encompass both Harris Regional AND Swain County Hospital. What do you think? Then, I'll change this record to "Hospital - Health facilities.""

The content facet also captured the cataloger's decisions regarding the geographical scope of an online resource (n=365, 14.8%). For example, on May 30, 2002 a cataloger stated that "they say they treat patients from western North Caro-

lina...how do I capture that for the county? …". The remaining 229 (9.3%) of the 2471 content annotations referred to miscellaneous content (the frequency differs between facets because the categories in each facet are not mutually exclusive).
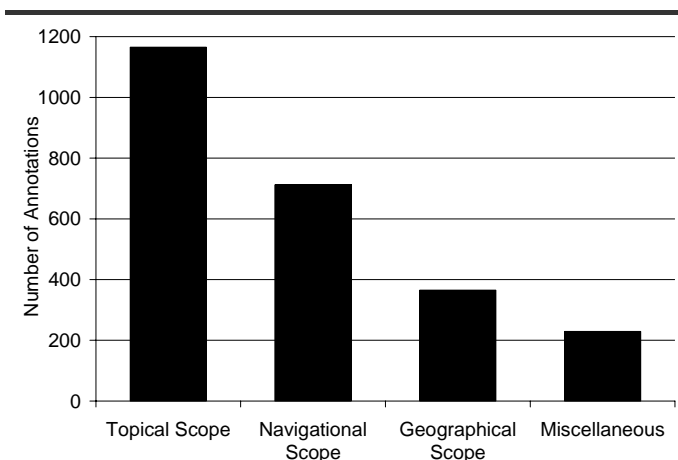
Figure 3 – Content Facet. Catalogers discuss the topical scope of the web page more often than they discuss navigational or geographical scope.

Figure 4 captures the format of the annotations made between catalogers. Annotations were most often in the form of a statement (n=1528, 64.2%), rather than a question (n=467, 19.6%) or an answer (n=384, 16.1%).
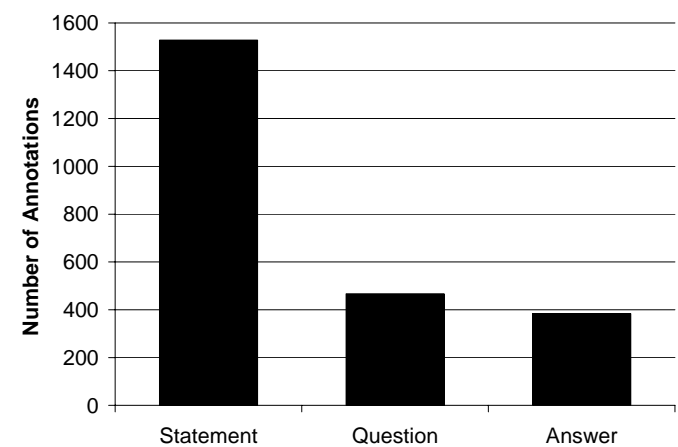
Figure 4 – Format Facet. Catalogers often make statements rather than asking or answering questions.

Figure 5 provides a breakdown of the annotation function. Catalogers most often used the annotations to leave reminders about their decision processes (n=1102, 47.7%). This is particularly useful for long term projects such as the *Go Local* portals because any given cataloger is unlikely to participate for the entire life of the project. For example, the NC Health Info project often selects catalogers from the pool of graduate students from degree programs in the School of Information and Library Science (SILS). Although the SILS program prepares students well for the cataloging task, graduation from the the two-year Masters program, result in relatively high staff turnover. If this high turnover is typical of other *Go Lo-*

cal sites (which we posit it is), the reminder annotations will play an important role in sharing information between catalogers, who work on the project during different time frames.
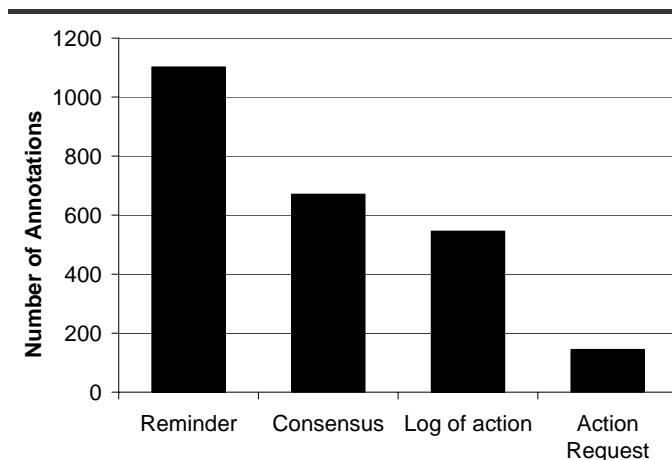
Figure 5 – Function Facet. Catalogers used annotations to leave reminders for themselves or other catalogers, to reach consensus or to log an action more often than they request an action from another cataloger.

Figure 5 also shows that catalogers often use the annotations to build consensus. Lines two and three, and lines eight through ten in Figure 2 are examples of consensus building. Of the substantive annotations, 671 (27.2%) involved consensus building. The remaining functions were to log an action (n=546, 22.2%) and to issue an action request (n=145, 5.9%).

Discussion

Although catalogers were not required to annotate their decision making process, more than half of the web pages in the NC Health Info portal had at least one annotation (1263 out of 2369). This suggests that catalogers found annotating web pages useful during the cataloging process, and leads us to recommend the inclusion of annotation in systems designed to support this important user population.

In order to understand if annotation behavior of catalogers changed over time, we compared the number of annotations for six different catalogers from April 2002 to September 2004 (data not shown). Our intuition behind this analysis was that the number of annotations made by an individual cataloger would decrease as their familiarity with the cataloging process increased. Such an analysis might indicate the time required to train a new cataloger. Contrary to our expectations experienced catalogers continued to provide annotations throughout the cataloging process.

Annotations enabled the catalogers to form consensus around the meaning of an existing information source, an activity that is not new in medicine. Scientists who conduct systematic reviews develop extraction worksheets that capture their consensus building activities [6]. Multiple reviewers independently extract information using the guidelines, then resolve differences. This process enables the group to establish group norms and verify the accurate extraction of information from each article.

Similarly, consensus building is an important consideration in recent efforts in bioinformatics to annotate scientific articles with terms from the gene ontology (www.geneontology.com). In both the systematic review and bioinformatics examples scientists have developed hierarchies of evidence that reflect the annotator's confidence in the final category assignment. In a systematic review, the stated study design reflects the level of evidence while in bioinformatics, scientists have invented a set of evidence codes[3] including "inferred from assay" and "inferred from genetic interaction" to measure their confidence. The annotations provided by the NC Health Info catalogers also reflect the cataloger's confidence in the final annotation. This serves as a surrogate for levels of evidence in this new area of web page annotation until accepted levels of evidence are developed.

The large number of annotations related to topical scope suggests that additional conversations are required to define the boundary of an online information resource. In this paper, we have hidden the complexity regarding annotating online information resources, by referring to each resource as a web page, which implies an individual page. However, catalogers do not catalog every individual web page on a site; rather they catalog an entire site, or a sub-set of pages within a site. The number of annotations indicates that defining these boundaries is very challenging.

The format facet provides insight into the nature of interactions that takes place between catalogers. A statement does not require an immediate response, but a question suggests that an interactive dialogue between catalogers is eminent. We are currently extending our analysis to explore interactions between catalogers.

Conclusion

This analysis is the first to explore the nature of interactions that occur between catalogers as they manually add terms from the *Go Local* vocabulary to online health information. The content analysis of 2301 annotations revealed that catalogers discuss the topical (n=1165), and navigational scopes, (n=712), of a web page more frequently than the geographical scope (n=365). Annotations were most often in the format of a statement (n=1528) rather than a question (n=467) or an answer (n=384). Catalogers made annotations as reminders to themselves or other catalogers (n=1102), to reach consensus (n=671) to log an action (n=546) and to issue a request (n=145).

Two of the challenges faced by catalogers are specific to an on-line environment. The first concerns the web page boundary, for example when should the cataloger assign a topic to an entire web site, and when should they assign topics to sub-domains? The second issue concerns the dynamic nature of on-line information compared to traditional information resources. Currently catalogers review web pages in the NC Health Info portal every six months. However, a cataloger need not review an unchanged web page, and should conduct a review if the page has changed within the six-month period.

---

[3] A complete list of evidence codes are available from the Gene Ontology URL http://www.geneontology.org/GO.evidence.shtml

Thus, an information system that detected change within a web page would aid in the allocation of resources for the review task.

None of the individuals who worked on the NC Health Info project was required to annotate their cataloging process; yet they provided annotations for more than half of the web pages. This suggests catalogers find annotations useful. Many of the catalogers on this project were students, so these annotations have long-term implications with respect to preserving organizational memory.

The NLM established the *GoLocal* initiative to provide consumers with information about health services, programs, and health care providers in their local community. As health care providers and health care consumers continue to use the online environment to disseminate and access information, the need for portals that provide high quality information will also increase. Studies such as ours, which characterize the challenges faced during the cataloging process, are the first step towards the development of information systems that support these important user communities.

References

1.Miller N, Lacroix E, Backus J. MedlinePlus:building and maintaining the National Library of Medicine's consumer health Web service. Bulletin of the Medical Library Association 2000; 88(1): 11-7.

2. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewer--Let the reader and viewer beware. JAMA 1997;277:1244-5.

3. Gagliardi A, Jadad AR. Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. BMJ 2002;324:569-73.

4. Eysenbach G, Powell J, Kuss O, Sa E-R. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web A Systematic Review. JAMA 2002;287(20):2691-2700.

5. Sagaram S, Walji M, Meric-Bernstam F, Johnson C, Bernstam E. Inter-observer Agreement for Quality Measures Applied to Online Health Information. In: MEDINFO 2004; 2004; 2004. p. 1308-12.

6. Alderson P, Green S, Higgins JPT, editors. Cochrane Reviewers' Handbook 4.2.2 [Updated March 2004]. Chichester, UK: John Wiley & Sons, Ltd; 2004.

Address for correspondence

For additional information, contact Catherine Blake at cablake@email.unc.edu