Auditing Dynamic Links to Online Information Resources

Jianhua Li, MD; James J. Cimino, MD
Department of Biomedical Informatics, Columbia University, New York, New York, USA

Abstract

The Columbia University Infobutton Manager (IM) is a system that provides dynamically generated, context-specific links between clinical information systems and online information resources. The resources range from local documents, to commercially available document sets and search engines. The links provided by the IM can be reliably created, but there is no guarantee that they will function reliably, since the resources to which they point are subject to unannounced changes and failures. We have developed a set of tools to audit the links periodically to determine if the resources are available and if the IM has sufficient information to generate all the links needed by its users. These tools have been in use since February, 2006 and have provided timely warnings on many occasions. These warnings have allowed us to correct problems with resource access before they became apparent to our users and often before the resource maintainers were aware of the problems. The tools have thus helped us provide clinicians with a dependable level of service.

Introduction

One of the strengths of the World Wide Web is that it allows virtually infinite integration of disparate resources to create information services that are far greater than individual local providers could create on their own. The caveat is that these disparate resources may change or become unavailable without notice; indeed, even if they wanted to provide notice, they would have no way of knowing whom to notify.

The need to be able to rely on and monitor the functionality of links to outside resources is widely recognized.[1] There are many link checker tools – both free[2,3] and commercially available[2,4,5] – that can check static HTML pages for broken links. However, these tools cannot help when links are embedded in other information systems or are dynamically generated.

For example, Lobach and colleagues developed a Web link management tool called WILIAM that assisted clinical information systems developers with managing links to online information resources.[6] WILLIAM used a commercial product (now discontinued) called InfoLink (BiggByte Software, Fort Smith, AR). InfoLink provided a wealth of information that resulted from testing the static links defined in WILLIAM. However, neither WILLIAM nor InfoLink could generate and test the large set of potential links that can occur when a generic query string is instantiated with a user's topic of interest.

The Infobutton Manager (IM) is a program that provides dynamically generated, context-specific links between clinical information systems and on-line health information resources.[7] The IM includes a knowledge base that is used to match contextual information with information needs and information needs with resources to generate automated queries into the resources. We (JL and JJC) are responsible for the maintenance of this knowledge base, but we provide links to resources on an as-is basis, with the assumption that the curators of the resources are responsible for keeping the contents up to date. Nevertheless, we must assure that the resources have not changed in such a way that the links will fail, and that the resources have not changed in ways that cause the automated queries to fail. Despite an extensive review of published literature, we remain unaware of any previous work to assure the integrity of dynamically generated Web links and automated queries. This paper describes our automated auditing process for addressing this task.

Methods

Our auditing process requires two components. First, it must assure that the links in the IM knowledge base are functioning. In some cases, these links are simple hyperlinks that can be checked directly by a program that simulates a user's Web browser. However, many links in the knowledge base are "generic", requiring some completion in order to point to specific documents that correspond to particular concepts of interest. In other cases, the links require modification to customize them with context variables, such as the concept of interest, patient age and gender, etc., for use with resource search engines.

Link Checker: Simple links to static documents are stored in the IM's knowledge base, such as this one to Elsevier's iConsult product:

http://infonet.nyp.org/Pharmacy/Pharmacy-M/L---N/MagAdult042506.pdf

```
Date: Thu, 8 Mar 2007 02:04:12 -0500
From: Jianhua Li
<jil7001@dbmi.columbia.edu>
To: ciminoj@dbmi.columbia.edu
Subject: Infobutton_links

----LINK CHECK REPORT: Infobutton----
This report was generated on 03/08/07
at 02:03:02
It took 1 minutes and 10 seconds to
generate this report

NUMBER OF BROKEN LINKS: 1
NUMBER OF Warning LINKS: 0
NUMBER OF LINKS PARSED: 103

BROKEN LINK(S):
3342: http://www.geri-ed.com/modules
/assess/assess/performance_oriented_
assessment_of_mobility.htm

This report was generated by
/localmisc/webcisdev13/cgi/webcis_
linkcheck_cron.cgi

------END OF REPORT------
```

Figure 1: Sample e-mail from the Link Checker

Other links require customization and contain placeholders for the required additional information, such as this one to Elsevier's iConsult product:

http://prod.iconsult.elsevier.com/iConsult/perform lookup?source=CU001&content=FC&output=red irect&primterm=<234>&gender=<g>

This link requires filling in values for the iConsult variables "primterm" and "gender". The "<234>" refers to an attribute from our terminology server, the Medical Entities Dictionary (MED).[8] Normally, the IM would fill in this attribute by querying the MED for attribute 234 for the concept of interest, and the MED would return a preferred name for a clinical finding. For the purpose of monitoring, the Link Checker uses preselected values ("Hyperkalemia" and "F", respectively) to produce a static link to the iConsult search engine. Executing the link will determine whether the search engine is available and functioning, but not whether it will actually return anything that relates to hyperkalemia in females.

The Link Checker is scheduled as a nightly job that collects all of the links in the IM's knowledge base, customizes them as needed, and then tests each link. The links are tested using a standard Perl library function (LWP::UserAgent) that carries out two World Wide Web functions (HTTP::Request and HTTP::Response). When a link is passed to the user agent, the following codes may be returned:

2xx - Success
3xx - Redirection
4xx - Client error
5xx - Server error

For example, many users are familiar with the return code "404 – Not Found".[9] We interpret 2xx as "Good", 3xx as "Warning", and 4xx and 5xx as "Broken".

The return codes are assembled into a report and the report is e-mailed to the system administrators each night. Figure 1 shows a sample report. The Link Checker can also be executed as needed from our IM online management site to produce a similar report, shown in Figure 2.

MED Checker: In cases where the resource is a set of hyperdocuments, with no search engine, the IM uses the MED to keep track of pages that relate to specific concepts. For example, for each test carried out by the clinical laboratory, the MED maintains a document name for the local laboratory manual. The IM contains this link in its knowledge base:

http://cpmclabinfo.cpmc.columbia.edu/<251>

and uses the MED to fill the "<251>" placeholder by

LINK CHECK FULL REPORT
GENERATED ON 02/26/06 AT 22:28:48

Infobutton_ID	URL	STATUS
15	http://search.atomz.com/search/?sp-q=tacrolimus&sp-a=sp1001878c	GOOD
21	http://teamportal.nyp.org/sites/NYP/COLE/IT/Eclipsys/default.aspx	GOOD
52	http://cancerweb.ncl.ac.uk/cgi-bin/omd?query=dilantin	GOOD
53	http://www.medicinenet.com/script/main/srchCont.asp?li=MNI&SRC=dilantin	GOOD
54	http://www2.merriam-webster.com/cgi-bin/mwmednlm?book=Medical&va=dilantin	WARNING
55	http://www.onelook.com/?w=dilantin&ls=a	GOOD
71	http://cpmclabinfo.cpmc.columbia.edu/chapter/mono/cl004500.htm#Container	BROKEN
81	http://www.crlonline.com/crlsql/leaflets-english/hf060200.htm	GOOD
82	http://www.crlonline.com/crlsql/pleaflets-english/hf000300.htm	GOOD

Figure 2: Example of manually generated report from the Link Checker

querying for the value of attribute 251 of the concept of interest (in this case, a laboratory test). If, for example, the concept of interest was a serum glucose test, the MED would return:

chapter/mono/cl001400.htm

which would then be used to create the complete link to the lab manual page for glucose tests:

http://cpmclabinfo.cpmc.columbia.edu/chapter/mono/cl001400.htm

One reason for using the MED to maintain this information is that it provides for inheritance in its class hierarchy. Thus, "chapter/mono/cl001400.htm" can be included in the MED as the value for attribute 251 for the class "Intravascular Glucose Measurement", and all of the descendants of this class (currently, there are 59) will return this value when the MED is queried for *their* attribute 251 values. The MED currently maintains the names of 520 documents for 9,010 laboratory test terms.

The CPMC laboratory manual undergoes periodic maintenance, with addition, removal, and renaming of pages. Therefore, the 520 links in the MED need periodic auditing to determine if they are still valid. Auditing also requires periodic reviews of the laboratory test terms in the MED that do not have corresponding lab manual pages (currently 2,188) to check to see if relevant pages have been added.

The MED maintains other document names for other resources. Notably, it contains 5,289 references to 786 adult drug information documents (attribute 252), 5,237 references to 472 pediatric drug information documents (attribute 253), and 7.080 references to 881 drug images, all from LexiComp.

The MED maintains other attributes besides document names that are useful for the automated information retrieval. For example, most disease terms in the MED have an official (attribute 48) or preferred (attribute 49) ICD-9-CM code associated with them and many finding terms have the preferred name (attribute 234) used by DXplain, the diagnostic decision support system.[10] The values for these attributes can be used as parameters for retrieving information from resources such as Micromedex, iConsult, and DXplain.

The MED checker provides an automatic audit of the attributes in the MED that need to be maintained for use in the Infobutton Manager. The first step is to determine which attributes are relevant. This step simply requires retrieving all the current links in the IM knowledge base and determining which ones contain MED attributes (such as "<234>" and "<251>", in the above examples). The results are used to customize and check links in a manner similar to that used by the Link Checker.

The second step is to determine which MED terms should have values for these attributes, so that the MED can provide them when requested by the IM. This is determined by checking the IM knowledge base to see which classes of concepts are considered "concepts of interest" for selecting links.[11] For example, the link to iConsult (described in the description of the Link Checker, above) is evoked whenever the concept of interest is in the class "Finding". The MED is queried to determine all of the descendants of the class Finding and the values for attribute 234 are obtained for each term. Those terms without values are reported (see Figure 3, for example).

MED Attribute Checker

Attribute	Broken Value
254	6867.htm
254	6738.htm
254	5923.htm
254	6925.htm

Attrib	Missing Value
251	1311: IgG/Total Protein Ratio Calculation
251	1319: Stool Starch Measurement
251	1320: Stool Fat Measurement
251	1324: Urine Xylose Measurement
251	1339: Serum LD1/LD2 Ratio Calculation

Figure 3: Sample report from the MED Checker

Results

Link Checker: The Link checker was placed into service February, 2006 and has been checking an average of 101.5 links daily. A 7-month summary of the reports is presented in Figure 4. Initially, one site that was referenced in two IM links was producing warning messages, but links to this site were removed on August 25th, 2006 (see Figure 4). Two single sites produced the majority (224 out of 317) of "BROKEN" messages – one was a local resource and the other was a guideline site; each of these were discontinued by their stewards but remained in the IM knowledge base for archival purposes. Most of the other broken links were due to sporadic unavailability of individual sites. In most cases, the sites were available again later the same day.

On November 2, 2006, access to one commercial site, available through our health sciences library, became unavailable because the library moved the resource. Four links in the IM knowledge base that referenced

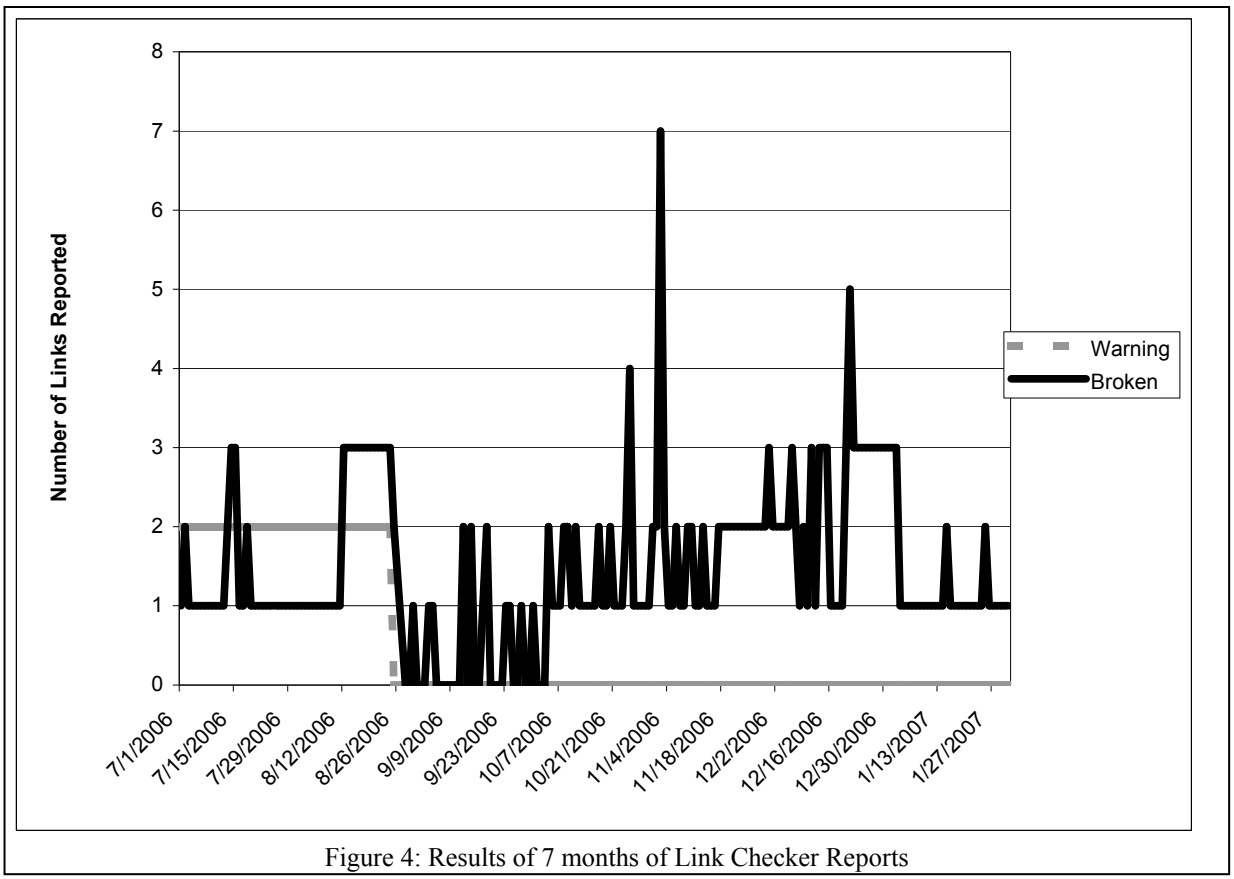

Figure 4: Results of 7 months of Link Checker Reports

that site were reported broken. We identified the problem and corrected it the same day. The same event occurred on December 21, and it was again corrected the same day.

MED Checker: The MED Checker is run on a periodic basis. In one run in February 2007, it discovered that 787 document names were no longer valid – this was entirely due to changes to the aforementioned commercial site, which recently began renaming its documents. (Note that the Link Checker showed no problem, since it was looking for the presence of a single, preselected document.) The document names were updated in the MED run in July 2007 showed no errors.

The MED checker also reported on 36,735 instances of missing attribute values. The majority of these (26,888) were missing "finding names" from attribute 234, mentioned above. These were not serious omissions, since the IM can use the preferred name for a term if it does not have a name in a specific terminology (in this case, the DXplain finding names). The remaining 11,047 missing values relate to a pair of attributes for ICD-9-CM codes. One attribute provided the code when the term was an official ICD9 term and the other provided the code when the term was not an official ICD9 term. No term should have a value for both attributes, but all disease terms in the MED should have a value for at least one of these. In fact, only 731 terms had neither. Codes were assigned to all these terms and, in the July analysis, only 104 new terms lacked a value for both attributes.

Discussion

The ability to link on-line information resources to clinical information systems is an important step in the translation of biomedical knowledge into practice. The Infobutton Manager is a system designed to carry out this task for real clinicians caring for real patients. As such, the performance requirements of the IM are identical to the rest of the clinical information systems in which it resides: it must be available seven days a week, 24 hours a day. Unfortunately, the information resources are held to a lower performance standard. In addition, they are designed for interaction with humans, who may be relatively oblivious to minor changes in the user interface, whereas the IM depends on complete stability. When information resources become unavailable, there is often no notification process for the maintainers, let alone the users. In many cases, the Link Checker has discovered problems before they were discovered by the resource maintainers.

The IM knowledge base currently contains over 100 links and the MED contains over 100,000 terms. Making sure these two resources work together to successfully interact with online resources is crucial to the function of the IM and, just as clearly, is beyond reach of manual maintenance. The Link Checker and MED Checker have helped us many times to catch problems before they became problems for the users. As we expand the IM for use in other institutions, this management task will grow geometrically.

The Web is a heterogeneous set of resources, with each having its own information model and interaction method. Nevertheless, we have found that these models and methods generalize into a relatively small set of paradigms. We have also found that relatively simple methods can be used to interact successfully with each of these paradigms.[12] We have exploited that approach with our automated auditing tools.

Anyone who is harnessing the richness of the Web by incorporating links to outside resources into their own applications, such as those who are creating infobutton managers,[13,14] will face challenges similar to ours. A simple catalogue of "canned" links, with a program to perform status checking using standard programming libraries, may be sufficient for many applications. However, when the links that an application uses are dynamically customized, and when the set of links itself is subject to frequent change, a similarly-dynamic approach, such as ours, will be needed for monitoring.

The ability to resolve clinician information needs by linking clinical systems to online knowledge depends on a chain of resources that must be maintained. The set of questions a clinical user might ask is relatively stable over time, and it is the task of resource maintainers to keep their knowledge up to date. Maintaining the links between them requires creative informatics solutions.

Conclusions

The Columbia University Infobutton Manager depends on reliable availability of outside information resources that are beyond its control. Our automated auditing tools provide a valuable mechanism for ensuring that the resources are available and accessible for the dynamic interactions required of them by the Infobutton Manager and its clinician users.

Acknowledgments

This work is supported in part by NLM grant R01LM07593. The authors thank Dr. Soumitra Sengupta ("Sen"), for inspiration.

References

1. Kasal P, Janda A, Feberova J, Adla T, Hladikova M, Naidr JP, Potuckova R. Evaluation of health care related web resources based on web citation analysis and other quality criteria. *Conf Proc IEEE Eng Med Biol Soc*. 2005;3:2391-4.
2. W3C Link Checker. http://validator.w3.org/checklink
3. Website Quality Management. http://www.web-ceo-site-auditor.com/
4. Web Link Validator. http://www.relsoftware.com/web-link-validator/
5. Link Checker Pro. http://www.link-checker-pro.com/
6. Lobach DF, Spell RU, Hales JW, Rabold JS. A Web link management tool for optimizing utilization of distributed knowledge in health care applications. Proc AMIA Symp. 1999;839-43.
7. Cimino JJ, del Fiol G. Infobuttons and point of care access to knowledge. In: Greenes RA, ed. *Clinical Decision Support: The Road Ahead*. Amsterdam: Elsevier, 2007:345-371.
8. Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. *Am Med Inform Assoc*; 2000;7(3):288-297.
9. http://www.hwg.org/lists/hwg-servers/responsecodes.html
10. Barnett GO, Famiglietti KT, Kim RJ, Hoffer EP, Feldman MJ. DXplain on the Internet. *Proc AMIA Symp*. 1998;:607-11.
11. Cimino JJ, Li J, Bakken S, Patel VL. Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users. *Proc AMIA Symp*. 2002;:170-4.
12. Cimino JJ, Li J, Allen M, Currie LM, Graham M, Janetzki V, Lee NJ, Bakken S, Patel VL. Practical considerations for exploiting the World Wide Web to create infobuttons. *Medinfo*. 2004;11(Pt 1):277-81.
13. Del Fiol G, Rocha RA, Clayton PD. Infobuttons at intermountain healthcare: utilization and infrastructure. *AMIA Annu Symp Proc*. 2006;:180-4.
14. Maviglia SM, Yoon CS, Bates DW, Kuperman G. KnowledgeLink: impact of context-sensitive information retrieval on clinicians' information needs. J *Am Med Inform Assoc*. 2006 Jan-Feb;13(1):67-73.