Prevalence and Inaccessibility of URLs in the Biomedical Literature

Sina Madani, MD, Randy J. Carnevale, BA, Stephany Duda, MS, Michael Feyder, BS, Dominik Aronsky, MD, PhD
Dept. of Biomedical Informatics, Emergency Medicine, Vanderbilt University, Nashville, TN

The World Wide Web is a dynamic environment that does not guarantee permanent access or content stability. We determined the prevalence of URLs in forthcoming, biomedical papers when they are first released in MEDLINE® and prospectively evaluated the rate of inaccessible URLs during a 19-day period. Among 96,153 references from 2,614 forthcoming papers (739 journals) the prevalence of URLs was 0.59%. The rate of inaccessible URLs was 12.4% when first available to the public community.

INTRODUCTION

The World Wide Web has changed the information access patterns of the scientific community. The electronic availability of scholarly work through the Internet has improved information access and accelerated the speed of knowledge dissemination. Many journals have adopted the practice of publishing electronic preprints of accepted papers leading to earlier availability in searchable databases such as MEDLINE®. Authors are increasingly citing information from the Internet by providing uniform resource locator (URL) references. Without a permanent digital library, citing URLs has the disadvantage that accessibility and content stability are not guaranteed. Researchers have demonstrated that the URL mortality is up to 55% after 5 years (1,2), however, the actual prevalence of URL citations in the biomedical literature is unknown. The rate of inaccessible URLs that are cited in articles upon their first release to the general public is also unknown.

The goals of the study were to:
1. Examine the prevalence of URLs in the biomedical literature; and
2. Prospectively evaluate the rate of inaccessible URLs in preprint (“forthcoming”) papers when they are first released in MEDLINE.

METHODS

During a 19-day study period (2/21/2006–3/11/2006) we prospectively collected all forthcoming papers published in PubMed® MEDLINE on a daily basis using the search term {pubstatusaheadofprint AND yyyyy/mm/dd[edat]}. This marks the first time when information about an article becomes accessible to the general public. The full text articles of a 20% random sample were retrieved. All references were copied and pasted into a spreadsheet. A URL extraction script searched the references for protocol terms, commonly used URL substrings, top domains, and other commonly used terms, and noted return codes (2). Inaccessible URLs were manually checked from 2 independent networks. All URL access attempts occurred within 1 day after MEDLINE appearance. MEDLINE covers ~4,800 journals. Our biomedical library subscribes to ~3,300 and our general library to ~12,000 electronic journals.

RESULTS

Table 1 displays URL occurrence and inaccessibility rates for preprint articles within 1 day of publication. HTTP error return codes for the 70 inaccessible URLs included 40 “not found,” 24 timed out, 3 “forbidden,” 2 “server error,” and 1 “unauthorized access.”

<table>
<thead>
<tr>
<th>URLs in Preprint Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
</tr>
<tr>
<td>Articles</td>
</tr>
<tr>
<td>Journals</td>
</tr>
<tr>
<td>References</td>
</tr>
<tr>
<td>URLs</td>
</tr>
<tr>
<td>Inaccessible URLs</td>
</tr>
</tbody>
</table>

CONCLUSION

We found a 0.59% prevalence rate of URL citations among forthcoming biomedical articles. The 12.4% inaccessible rate upon publication is considered high and calls for careful scrutiny of URL references from authors, journal editors, and publishers prior to releasing a publication.

Acknowledgment: The 2nd and 3rd authors were supported by Training Grant T15 LM 007450-03

References