# Preserving The Scholarly Record With WebCite (www.webcitation.org): An Archiving System For Long-Term Digital Preservation Of Cited Webpages

*Gunther Eysenbach*[1,2,3]

[1] Centre for Global eHealth Innovation, University Health Network,
190 Elizabeth St, Toronto M5G2C4, Canada
e-mail: geysenba at uhnres.utoronto.ca
[2]Department of Health Policy, Management, and Evaluation, University of Toronto
[3]Knowledge Media Design Institute, University of Toronto

**Abstract**

Scholars are increasingly citing electronic "web references" which are not preserved in libraries or full text archives. WebCite is a new standard for citing web references. To "webcite" a document involves archiving the cited Web page through www.webcitation.org and citing the WebCite permalink instead of (or in addition to) the unstable live Web page. Almost 200 journals are already using the system. We discuss the rationale for WebCite, its technology, and how scholars, editors, and publishers can benefit from the service. Citing scholars initiate an archiving process of all cited Web references, ideally before they submit a manuscript. Authors of online documents and websites which are expected to be cited by others can ensure that their work is permanently available by creating an archived copy using WebCite and providing the citation information including the WebCite link on their Web document(s). Editors should ask their authors to cache all cited Web addresses (Uniform Resource Locators, or URLs) "prospectively" before submitting their manuscripts to their journal. Editors and publishers should also instruct their copyeditors to cache cited Web material if the author has not done so already. Finally, WebCite can process publisher submitted "citing articles" (submitted for example as eXtensible Markup Language [XML] documents) to automatically archive all cited Web pages shortly before or on publication. Finally, WebCite can act as a focussed crawler, caching retrospectively references of already published articles. Copyright issues are addressed by honouring respective Internet standards (robot exclusion files, no-cache and no-archive tags). Long-term preservation is ensured by agreements with libraries and digital preservation organizations. The resulting WebCite Index may also have applications for research assessment exercises, being able to measure the impact of Web services and published Web documents through access and Web citation metrics.

**Keywords:** Internet archiving, digital preservation, citing web material

## 1. Introduction

Scholars (but also legal professions1 and authors of lay publications) are increasingly citing electronic npn-journal "web references" (such as Wikis, Blogs, homepages, PDF reports) 2 which are generally not permanently preserved in libraries or repositories such as Pubmedcentral and are prone to become inaccessible over time (Error 404 – Not found).

The unstable nature of web references is increasingly recognized as a problem within the scientific community and has been the subject of recent articles and discussions [1-11].

In a seminal article, Dellavalle et al have shown that 13% of Internet references in scholarly articles were inactive after only 27 months. Even if URLs are still accessible, another problem is that cited webpages

may have changed, so that readers see something different than what the citing author saw, sometimes without realizing this. Dellavalle et al. have concluded that "publishers, librarians, and readers need to reassess policies, archiving systems, and other resources for addressing Internet reference attrition to prevent further information loss" and called this an issue "calling for an immediate response" by publishers and authors[1].

Until recently (before WebCite was available), the only option readers of articles which cite inaccessible URLs had to retrieve the "lost" webmaterial was to consult services such as the Internet Archive (Wayback Machine) or the Google archive, hoping that they may - by pure chance – have a version of the cited web document in their archive (hopefully a version which is close to the access date). However, the Internet Archive, Google, and other Internet archiving initiatives commonly use unspecific crawlers to harvest the Web in a shotgun-approach, not focussing on academic references, and the archiving process cannot be initiated by authors, editors, or publishers wanting to archive a specific web reference at a specific time and date, as they saw it when they quoted it. Moreover, certain webmaterial may be part of the "hidden web" and not be accessible to archiving crawlers. Therfore, these traditional approaches are inadequate.

The objective of this paper is to present and discuss a solution called WebCite (http://www.webcitation.org), an on-demand archiving system for authors, journal/book editors, and publishers for long-term preservation of cited webreferences.

WebCite is a tool specifically designed to be used by authors, readers, editors and publishers of scholarly material, allowing them to permanently archive cited "non-journal" Web material, such as cited homepages, wiki pages, blogs, draft papers accessible on the web, "grey" PDF reports, news reports etc.

To prevent "link rot", authors simply have to cite the WebCite snapshot ID and/or a link to the permanent WebCite URL, in addition to citing the original URL.

WebCite is now used by an increasing number of authors and journals3, ensuring permanent availability of cited webreferences for future readers and scholars. WebCite has built a XML-based webservice architecture which enables for example publishers, webmasters, editors, institutions, and vendors of bibliographic software packages to exchange data (e.g. metadata) and to trigger an archiving request. As such, WebCite can be seen as an intermediary between the scholarly community (authors/editors/publishers) and the digital preservation community.

As member of the International Internet Preservation Consortium (IIPC), WebCite works together with IIPC members such as the Internet Archive to create 1) a distributed storage infrastructure (so that WebCite snapshots are automatically fed into other digital archives, such as the Internet Archive or National Libraries and Archives), and 2) an interoperability infrastructure which would allow a federated search across different archives (e.g., if a reader clicks on a WebCite link with the format www.webcitation.org/query?date=..&url=..., the system looks across different archives to locate snapshots of that URL cited on a specific date).

WebCite is also working on advanced Web 2.0 functionalities, allowing users to share and recommend documents with other users.

Finally, WebCite creates alternative statistics on the usage and citation of websites and non-journal webmaterial, which can be used (in analogy to the impact factor) to measure the scholarly "impact" of a given webservice or web document, based on citations in scholarly publications.

## 2. Methods: WebCite Architecture, Workflow and Functionality

### 2.1. History

The WebCite idea was first conceived in 1997 and mentioned in a 1998 article on quality control on the Internet, alluding to the fact that such a service would also be useful to measure the citation impact of webpages [12]. In the same year, a non-functional mockup was set up at the address webcite.net (see archived screenshots of that service at the Internet Archive [1]). However, shortly after, Google and the Internet Archive entered the market, both apparently making a service like **WebCite** redundant. The idea was revived in 2003, when a study published in Science [1] concluded that there is still no appropriate and agreed on solution in the publishing world available. Both the Internet Archive and Google do not allow for "on-demand" archiving by authors, and do not have interfaces to scholarly journals and publishers to automate the archiving of cited links. In 2005, the first journal [*Journal of Medical Internet Research*] announced using WebCite routinely [13], and dozens of other journals followed suit. Biomed Central, publisher of hundreds of open access journals, has been using WebCite routinely since 2005 (all URLs cited in Biomed Central articles are automatically archived by WebCite)[2].

### 2.2 Functionality Overview

Authors and journal editors ensure long-term accessibility of cited URLs by using WebCite-enhanced references. A WebCite-enhanced reference is a reference which contains - in addition to the original live URL (which can and probably will disappear in the future, or its content may change) - a link to an archived copy of the material, exactly as the citing author saw it when he accessed the cited material.

There are two basic formats of a WebCite URL: The opaque and the transparent format - the former can be used to be added to a cited URL, the latter can be used to replace a cited URL. Both formats will be returned in response to an archiving request, usually initiated by the citing author.

The *opaque URL* is very short and handy, containing a short ID like 5Kt3PxfFl (http://www.webcitation.org/5Kt3PxfFl). This format should only be used in a reference where the original URL is still visible:

> Example References "enhanced" by WebCite:
> [1] Lawrence, Lessig: "this is a fantastically cool idea" (Blog). Sept 8, 2006. http://lessig.org/blog/2006/09/ Archived by WebCite at http://www.webcitation.org/5UzgHmsS7 on 20-01-2008)

Alternatively, the cited URL and the cited date can be part of a single WebCite URL (the *transparent format*), making it redundant to spell out the original URL. The drawback is that the WebCite URL can become pretty long:

> Alternative format:
> [2] Lawrence, Lessig: "this is a fantastically cool idea" (Blog). Sept 8, 2006.
> Archived by WebCite at http://www.webcitation.org/query?url=lessig.org/blog/2006/09/&date=2008-01-20

These are just examples, the actual citation formats preferred by different editors may differ. Most style guides currently give little or no guidance on how to cite URLs and their archived version, but most editors will accept something along the lines of citing the original URL together with the archived URL in a submitted manuscript. It is possible to omit the archiving or "accessed on" date (which is recommended in most style guides when citing URLs), because WebCite always tells the reader when the snapshot was taken and in the transparent format it becomes part of the URL.

Another form of a WebCite link contains the cited URL and the DOI (Digital Object Identifier) of the *citing*

document (refdoi):

[3] IMEX.
http://www.webcitation.org/query.php?url=http://imex.sourceforge.net&refdoi=10.1186/jbiol36

This format is used by publishers who use the DOI system to identify their articles and who have implemented WebCite (e.g. Biomed Central) by sending us their citing articles shortly before or at publication (which our software combs for URLs which have not been archived by the citing author). A WebCite URL containing a refdoi implies that a snapshot of the cited URL was taken when the citing paper (identified by its DOI) was published. Thus, the archiving date of the snapshot retrieved by WebCite if the reader clicks on this link will be close to the publishing date of the article. The example above is from the citing paper with the DOI 10.1186/jbiol36, which cites the URL http://imex.sourceforge.net (see reference 45).

A fourth way to retrieve a WebCite snapshot is using a hash sum (a sort of a digital fingerprint of a document), however, citing a document by URL and date is currently much more common than using a hash.

A fifth way is that WebCite may have (on request of the cited author/publisher) assigned a DOI to an archived snapshot, so that the link has the format http://dx.doi.org/10.2196/webcite.xx" (where xxx is the hash key of the material in WebCite). The DOI resolver at dx.doi.org (which is commonly also used to resolve cited journal and book references) would then resolve to a WebCite snapshot page (or an intermediary page pointing to the same work in other archives, to other manifestations such as print or pdf, or - in the case of online preprints - to "final" publications).

Before the WebCite URL can be cited, the archiving process has to be initiated, as described in detail below. The archiving process can be initiated by citing authors, editors, publishers, or cited authors. Various degrees of "automation" exist, from manual initiation of the archiving of a specific page, to automatic harvesting of cited URLs and subsequent archiving of these URLs. Authors usually use the archiving form, the WebCite bookmarklet, or upload an entire citing manuscript to the WebCite server via the comb page, which initiates the WebCite tool to comb through the manuscript and to archive all cited non-journal URLs. Participating journal or book editors, publishers, or copyeditors, participate through inserting a note in their "Instructions for authors" asking authors to use webcitation.org to permanently archive all cited webpages and websites before manuscript submission, and to cite the archived copy in addition to the original link. Participating Publishers (such as BioMed Central) submit manuscript XML files to WebCite at the time of publication, so that WebCite can comb through the manuscript and archive cited webpages automatically. Citable authors, i.e. academic bloggers and authors of non-journal scholarly webpages who foresee the possibility to be cited in the scholarly literature ("citable Web-author" or subsequently called "cited author"), but are concerned about the persistence and citability of their work can add a "WebCite this!" link to their work, which links dynamically to the archiving form.

### 2.3 Detailed Architecture and Workflow (see Figure 1)

#### 2.3.1 "Cited" Author-initiated archiving

These pathways are initiated by the author of webmaterial who is concerned about the citability and long-term preservation of his unpublished online work (we call these cited author or citable author). The work can be a blog, a discussion paper published on a website, an online preprint/draft of a paper, a posting in

a newsgroup etc. It is presumed that the cited author is also the copyright holder.

1. Self-archiving through the form at http://www.webcitation.org/archive (or a bookmarklet or browser-plugin). Optionally, authors may post a "cite-as" statement on their document, containing the WebCite link, indicating that they prefer that people (citing authors) prefer to cite a given version.
2. For dynamically changing content, authors can publish a button (Cite this page!) which dynamically creates a new archived version[1] whenever somebody (a "citing author") clicks on that link . This (this is a link to the archiving form which is populated with metadata and the URL to be archived handed over in the URL, for example http://www.webcitation.org/archive?url=http%3A%2F%2Foalibrarian.blogspot.com%2F2006%2F07%2Fopen-source-open-access-and-open.html&title=Open+Source%2C+Open+Access+-+And+Open+Search%3F&author=Giustini+Dean&date=2006-07-13&source=OA+Librarian&subject=electronic+publishing%3B+open+access.

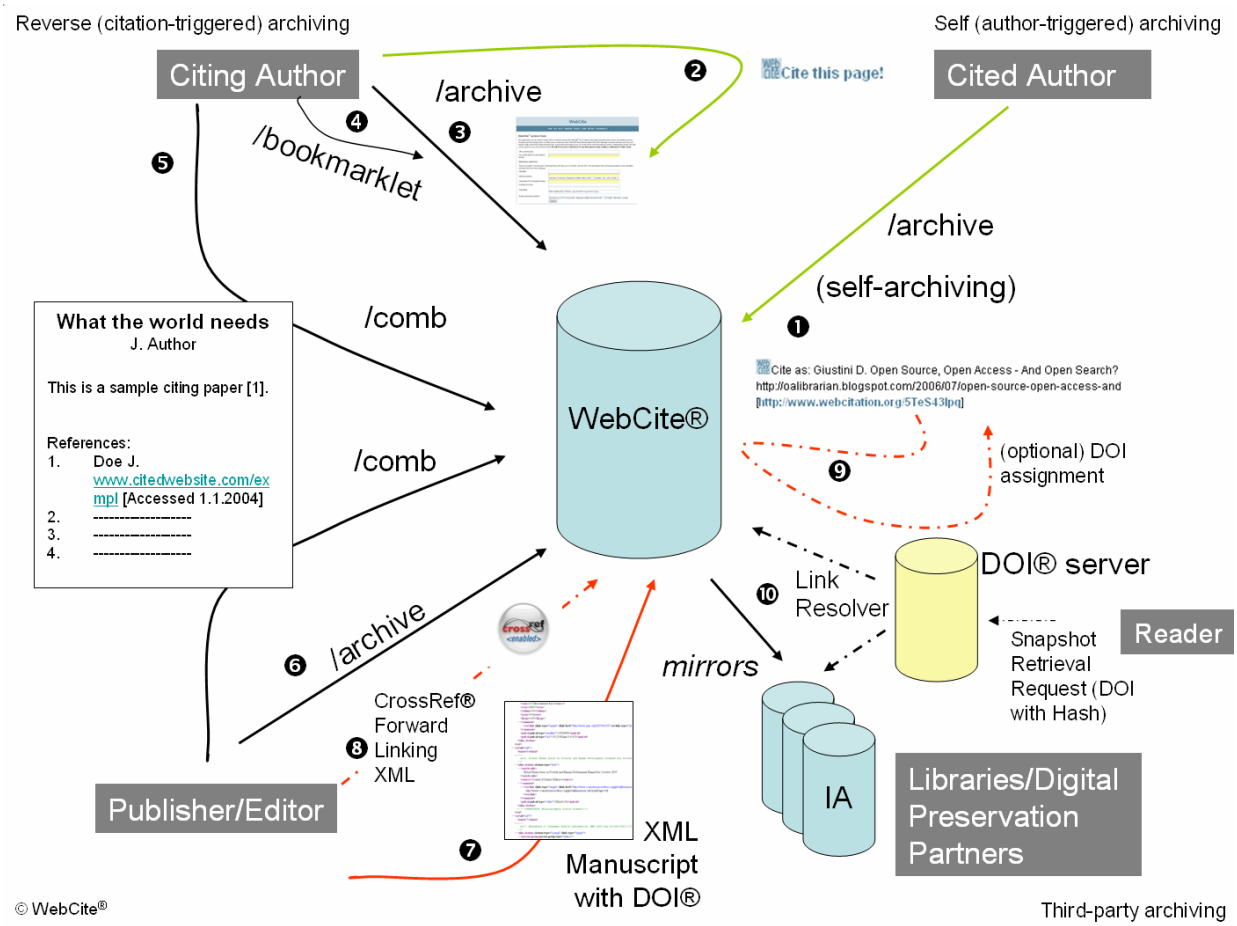



**Figure 1: WebCite architecture**

### 2.3.2 "Citing" Author-initiated archiving

2. See above. While the "cited author" facilitates archiving by publishing an archiving link, the actual archiving process is initiated by the citing author
3. Most citing authors will initiate archiving of a webpage through the form at http://

www.webcitation.org/archive ....

4. or through a bookmarklet (or through other yet-to-be-developed browser plugins which facilitate metadata entry)
5. Citing authors can also publish a draft manuscript on the web, and then let WebCite "comb" through this html-manuscript to archive all (or selected) cited URLs. Further functionality to comb through Word/ODT/RTF files is planned.

### 2.3.3 Publisher/Editor-initiated archiving

6. Publishers, editors, copyeditors who publish "citing" documents (academic papers, books etc) can use thesame tools as citing authors to archive cited URLs (/archive, /comb)..,
7. When processing a manuscript for publishing, publishers can submit a tagged document to WebCite for processing. WebCite will automatically archive all cited URLs (unless they have already been archived by the citing author). Ideally, this submission is done via FTP, and uses a well defined (preferably XML based) schema for article data. Currently WebCite support (X)HTML documents, NLM Journal Publishing DTD documents, and BioMed Central Article DTD documents. Adding new document types to this list is a straightforward process, and can be undertaken on a publisher by publisher basis by providing WebCite with a document DTD and sample document for testing. Note that if publishers use this avenue and if they are using DOIs for their citing articles, they can link to the archived URL simply by using a format like this: http://www.webcitation.org/query?url=http%3A%2F%2Fwww.iom.edu%2F%3Fid%3D19750&refdoi=10.2196/jmir.8.4.e27, where url is the archived URL and refdoi is the DOI of the citing document. That way, publishers know the link to the WebCite-cached copy ahead of time, and do not necessarily have to analyze the XML document which WebCite returns and which contains the WebCite IDs and success/failure messages. Providing a refdoi in the query implies that WebCite retrieves an archived copy of the URL which is close to the submission date.
8. *(possibly to be implemented in collaboration with CrossRef)* Publishers who are members of CrossRef currently supply their article metadata in an XML file formatted according to the CrossRef XSD schema version 3.0.1 for forward linking. This schema provides for the optional inclusion of citation lists attached to the existing journal article metadata, however, the current version does not support of non-journal, non-book, non-conference citations. Should the schema be revised to include cited URLs of webcitations, and if CrossRef makes these cited URLs (together with the referring DOIs of citing articles available, then WebCite could automatically archive cited URLs without publishers having to upload their content separately to WebCite as described under (7).

### 2.3.4 Mirrors, Creating a redundant infrastructure

9. "Cited authors" can request assignment of a DOI to their archived work. The DOI will be something like 10.2196/webcite.863eb3546af7384e44fb0e422cca1fa97704abeb, with the part after webcite being the digital fingerprint (hash sum) of the archived document. This enables citing authors to use the DOI resolver at dx.doi.org instead of using a "direct" link to WebCite, adding a layer of w
10. *(to be implemented)* WebCite deposits archived copies in mirrors and secondary archives

(e.g. Internet Archive and other IIPC members). While not implemented yet, in the future citing authors and readers could use the link-resolver at dx.doi.org to retrieve archived documents which have a DOI. Alternatively, cross-archive searches for snapshots with a certain URL/Date could be conducted.

## 2.4 Use cases

### 2.4.1 Using WebCite as a (citing) author to archive webpages

WebCite is an entirely free service for authors who want to cite webmaterial, regardless of what publication they are writing for (even if they are not listed as members).

The author of a citing manuscript can:

- Either manually initiate the archiving of a single cited webpage (by using either the WebCite bookmarklet or the archive page) and manually insert a citation to the permanently archived webdocument on webcitation.org in his manuscript, or;
- Upload an entire citing manuscript to the WebCite server via the comb page, which initiates the WebCite tool to comb through the manuscript and to archive all cited non-journal URLs. The WebCite software also replaces all URLs in the manuscript with a link to the permanently archived webdocument on webcitation.org.

### 2.4.2 Using WebCite as a reader

Readers simply click on the WebCite link provided by publishers or citing authors in their WebCite-enhanced references to retrieve the archived document in case the original URL stopped working, or to see what the citing author saw when he cited the URL.

Readers can also search the WebCite database to see how a given URL looked like on a given date - provided somebody has cited that URL on or near that date. The date search is "fuzzy", i.e. the date does not have to match exactly the archiving date - we always retrieve the closest copy and give a warning if the dates do not match. A drop-down list on top of the frame with different dates tells readers that snapshots were taken on these dates. Select any of these dates to retrieve the respective snapshot.

### 2.4.3 Using WebCite as an editor

Participating journal or book editors, publishers, or copyeditors should insert a note in their "Instructions for authors" asking their authors to use webcitation.org to permanently archive all cited webpages and websites, and to cite the archived copy in addition to the original link7.

Secondly, editors/copyeditors should initiate the archiving of cited webpages (either manually, or through the automated comb mechanism, which involves uploading the entire manuscript so that the WebCite engine can crawl and archive all cited URLs) and replace all webcitations in a manuscript with links to the archived copy, before the manuscript is published.

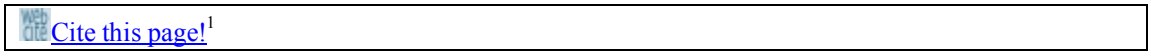
Thirdly, editors are encouraged to become a member of WebCite (which is free) so that their journal is listed on webcitation.org as participating journal.

### 2.4.4 Using WebCite as a library, internet archive, or digital preservation organization

WebCite works with digital preservation partners who are running dark mirrors, and is working on a federated cross-archive search and common API infrastructure. This ensures that archived content remains accessible for future generations, without being dependant on a single server. This is currently work in progress (interested potential partners are encouraged to contact the author).

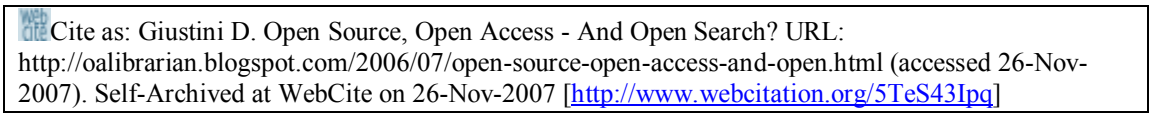
### 2.4.5 Using WebCite as a citable Web-author (blogger, copyright holder of Webpages, etc), or for self-archiving?

Academic bloggers and authors of non-journal scholarly webpages who foresee the possibility to be cited in the scholarly literature ("citable Web-author" or subsequently called "cited author"), but are concerned about the persistence and citability of their work can add a WebCite link to their work. If their content is dynamically changing, then they are encouraged to publish a button which creates a new archived version whenever somebody cites the work:




This links to the WebCite archiving request form – the link can contain prepopulated variables containing the Dublin Core metadata of the cited page, which will ensure that the reader ("citing author") knows exactly how to cite the work, and makes sure that a snapshot of the cited work is preserved in WebCite and its digital preservation partners.

If your online content is static, and bloggers (and other web authors) may want to encourage readers to cite a specific version. In this case, they first self-archive their work using the WebCite archiving request form, which simply requests the URL of the page to be archived. After successful archiving (which is done in an instant), web authors may then publish a button and link to the snapshot in the WebCite archive within their work, or when referring to it, for example:




Thus, WebCite can be used by authors as a one-click self-archiving tool, to ensure that for example preprints, discussion papers, and other formally unpublished material remains citable and available. All they have to do is to publish a preprint online, and then to self-archive it (one-click self-archiving).

Note that all these buttons should not be used for journal articles, which are presumably already archived through other mechanisms (LOCKSS etc.)

WebCite premium members will also be able to search the archive, assign a DOI to their work archived in WebCite, specify whether ads can be displayed (they receive a proportion of the ads revenues) etc. If a DOI is assigned, citing authors/publishers can use a link to dx.doi.org, which enables doi.org to resolve the link to either WebCite or another archive, e.g. the Internet Archive, if a archived copy with an identical hash is found.

### 2.4.6 Using WebCite as a publisher

Participating publishers include publishers of scholarly journals like BioMed Central who use WebCite to preserve cited webmaterial routinely and largely automatically.

They do this by encouraging their editors to instruct their authors and copyeditors to cache all cited URLs "prospectively" before submission (level 1) or during the copyediting process (level 2), respectively, and/or by submitting manuscript XML files to WebCite at the time of publication (level 3), so that WebCite can comb through the manuscript and archive cited webpages automatically. WebCite can also analyze back-issues of your journal(s) and archive the cited documents "retrospectively" (level 4).

**Implementation example at publishers**

Biomed Central uses a WebCite webcite logo to link to the archived copy in every webreference.

Another example is the *Journal of Medical Internet Research* - almost all articles in this Journal cite URLs, and since 2005 all are archived. See http://www.jmir.org/2005/5/e60#ref9 for an example.

## 3. Results

Since 2005, WebCite has been used by over 200 scholarly journals, and has archived over 3 Million scholarly important files and webpages.

## 4. Discussion

### 4.1 Copyright Issues

Caching and archiving webpages is widely done (e.g. by Google, Internet Archive etc.), and is not considered a copyright infringement, as long as the copyright owner has the ability to remove the archived material and to opt out. WebCite honors robot exclusion standards, as well as no-cache and no-archive tags. WebCite also honors requests from individual copyright owners to have archived material removed from public view.

A U.S. court has recently (Jan 19th, 2006) ruled that caching does not constitute a copyright violation, because of *fair use* and an *implied license* (Field vs Google, US District Court, District of Nevada, CV-S-04-0413-RCJ-LRL). *Implied license* refers to the industry standards mentioned above: If the copyright holder does not use any no-archive tags and robot exclusion standards to prevent caching, WebCite can (as Google does) assume that a license to archive has been granted. Fair use is even more obvious in the case of WebCite than for Google, as Google uses a "shotgun" approach, whereas WebCite archives selectively only material that is relevant for scholarly work. Fair use is therefore justifiable based on the fair-use principles of *purpose* (caching constitutes transformative and socially valuable use for the purposes of archiving, in the case of WebCite also specifically for academic research), the *nature* of the cached material (previously made available for free on the Internet, in the case of WebCite also mainly scholarly material), *amount* and substantiality (in the case of WebCite only cited webpages, rarely entire websites), and *effect* of the use on the potential market for or value of the copyrighted work (in the case of Google it was ruled that there is no economic effect, the same is true for WebCite).

In the future, WebCite will further reduce its liability by feeding content into third party archives such as National Libraries and Archives, which (often) have a legal deposit mandate. Secondly, WebCite is working on a more sophisticated infrastructure which would allow copyright holders to not only withdraw their content, but to specify a fair royalty fee.

### 4.2 Business Model

In order to cover the costs for ongoing and sustainable operations, WebCite will have to generate a

revenue stream, for example through the following mechanisms:

- Premium membership accounts for individuals (e.g. "cited authors") and institutions (publishers, universities) with an annual fee, which enables the assignment of a DOI to unpublished online work, the automatic processing of "citing" XML documents, and offer the opportunity to display advertisements (own ads, or Google Adsense) together with the archived content, if the member owns the copyright of the content. Premium accounts could also offer services such as access to certified citation statistics (for promotion and tenure, individuals and universities may want to have this information), citation recommendations (people who cited this webpage also cited …) etc.
- Advertising: "Cited authors" may decide to enable Google Adsense ads with their content (with them receiving royalties)
- Royalty collection: Cited authors/copyright holders will get the option to specify a per-pay-view royalty, which WebCite could collect for the copyright holder (receiving a commission)
- Developing tools and consulting for companies seeking to integrate WebCite into their products (e.g. vendors of bibliographic software packages such as RefMan or Endnote).

## 5. Conclusions

The current state of scholarly communication on the web can be characterized by the following paradox:

- blogs (and other Internet venues such as wikis) are - at least in theory - important venues for scholarship to publish hyptheses, analyses etc. outside of the traditional journal publishing system
- yet, they are not considered "citable" or "publications" - which in turn affect their use, usefulness, and acceptance among researchers as tools for scholalry communication.

WebCite aims to make Internet material (any sort of digital objects) more "citable", long-term accessible, and hence more acceptable for scholarly purposes. Without WebCite, Internet citations are deemed ephemeral and therefore are often frowned upon by authors and editors. However, it does not make much sense to ignore opinions, ideas, draft papers, or data published on the Internet (including wikis and blogs), not acknowledging them only because they are not "formally" published, and because they are difficult to cite. The reality is that in the age of the Internet, "publication" is a continuum, and it makes little sense to not cite (therefore acknowledge) for example the idea of a scholarly blogger, the collective wisdom of a wiki, ideas from an online discussion paper, or data from an online accessible dataset only because online material is not deemed "citable". By making Internet material more "citable" (and also by creating incentives such as mechanisms and metrics for measuring the "impact" of online material by calculating and publishing WebCite impact factor), we hope that this will encourage scholars to publish ideas and data online in a wide range of formats, which in turn should accelerate and facilitate the exchange of scientific ideas. While we do see the value of scholarly peer-reviewed journals for publishing research results, we also acknowledge that much of the scientific discourse takes place before it is "formally" published, and that peer-review can also take on other forms (e.g. post-publication peer-review, which is something WebCite plans to implement).

Another broader societal aspect of the WebCite initiative is advocacy and research in the area of copyright. We aim to develop a system which balances the legitimate rights of the copyright-holders (e.g. cited authors and publishers) against the "fair use" rights of society to archive and access important material.

We also advocate and lobby for a non-restrictive interpretation of copyright which does not impede digital preservation of our cultural heritage, or free and open flow of ideas. This should not be seen as a threat by copyright-holders - we aim to keep material which is currently openly accessible online accessible for future generations without creating economic harm to the copyright holder. This is a challenging, but feasible goal, and future iterations of this service may include some sort of revenue sharing mechanism for copyright holders.

Yet another angle is that WebCite enables "one-click self-archiving", making it very easy for scholarly authors to create a permanent, openly accessible record of their own work and their ideas. While the primary pathway in the WebCite system is third-party initiated archiving (triggered by a citing author), WebCite also provides a very simple mechanism for authors to self-archive their own work.

Another perspective is that WebCite is an innovative Internet archiving process that could be referred to as a "reverse archiving" or "Archiving 2.0" approach. Rather than to let librarians or archiving crawlers (such as the Wayback Machine) archive material (and to let curators assign metadata), WebCite puts the initiation of the archiving process into the hands of the scientific community, who – by virtue of citing it – decides what is considered worthy archiving. The assignment of metadata is also a highly decentralized, bottom-up process (which involves the community of "citing" authors, but also the cited author).

## 5. Notes

1 A New York Times article, published Jan 29, 2007, "Courts Turn to Wikipedia, but Selectively" by Noam Cohan mentions WebCite: "*(...) 'citation of an inherently unstable source such as Wikipedia can undermine the foundation not only of the judicial opinion in which Wikipedia is cited, but of the future briefs and judicial opinions which in turn use that judicial opinion as authority.'*
*Recognizing that concern, Lawrence Lessig, a professor at Stanford Law School who frequently writes about technology, said that he favored a system that captures in time online sources like Wikipedia, so that a reader sees the same material that the writer saw.*
*He said he used* ***www.webcitation.org*** *for the online citations in his amicus brief to the Supreme Court in Metro-Goldwyn-Mayer Studios v. Grokster Ltd., which "makes the particular reference a stable reference, and something someone can evaluate. (...)".*

2 It is important to understand that WebCite focuses on documents exclusively available on the web, not documents such as journal articles which can be assumed to be archived in libraries.

3 For a full list of journals using WebCite see http://www.webcitation.org/members. Accessed: 2008-06-04. (Archived by WebCite® at http://www.webcitation.org/5YJvduH5t)

4 http://web.archive.org/web/19990203173551/webcite.net/home.htm

5 Cockerill M. Webcite links provide access to archived copy of linked web pages. BioMed Central Blog. URL: http://blogs.openaccesscentral.com/blogs/bmcblog/entry/webcite_links_provide_access_to [Archived in WebCite at http://www.webcitation.org/5Tb2FDt4e on 2007-11-14]

6 WebCite® automatically determines whether there is a need for storing another physical copy, or whether the content has already been archived, in which case a new WebCite ID is generated which points .

7 For an example see: http://www.jmir.org/cms/viewInstructions_for_Authors:Instructions_for_Authors_of_JMIR#webcite

[8] This links to the WebCite archiving request form – the link can contain prepopulated variables containing the Dublin Core metadata of the cited page, e.g. http://www.webcitation.org/archive?url=http%3A%2F%2Foalibrarian.blogspot.com%2F2006%2F07%2Fopen-source-open-access-and-open.html&title=Open+Source%2C+Open+Access+-+And+Open+Search%3F&author=Giustini+Dean&date=2006-07-13&source=OA+Librarian&subject=electronic+publishing%3B+open+access

## 6. References


[1] Dellavalle RP, Hester EJ, Heilig LF, Drake AL, Kuntzman JW, Graber M *et al*. Going, Going, Gone: Lost Internet References. *Science* 2003;**302**:787-8.

[2] Crichlow R, Davies S, Winbush N. Accessibility and Accuracy of Web Page References in 5 Major Medical Journals. *JAMA* 2004;**292**:2723-b.

[3] Hester EJ, Heilig LF, Drake AL, Johnson KR, Vu CT, Schilling LM *et al*. Internet citations in oncology journals: a vanishing resource? *J Natl.Cancer Inst.* 2004;**96**:969-71.

[4] Johnson KR, Hester EJ, Schilling LM, Dellavalle RP. Addressing internet reference loss. *Lancet* 2004;**363**:660-1.

[5] Kelly DP, Hester EJ, Johnson KR, Heilig LF, Drake AL, Schilling LM *et al*. Avoiding URL reference degradation in scientific publications. *PLoS.Biol.* 2004;**2**:E99.

[6] Schilling LM, Kelly DP, Drake AL, Heilig LF, Hester EJ, Dellavalle RP. Digital Information Archiving Policies in High-Impact Medical and Scientific Periodicals. *JAMA* 2004;**292**:2724-6.

[7] Schilling LM, Wren JD, Dellavalle RP. Bioinformatics leads charge by publishing more Internet addresses in abstracts than any other journal. *Bioinformatics* 2004;**20**:2903.

[8] Badgett RG, Berkwits M, Mulrow C. Scholarship Erosion. *Ann.Intern.Med.* 2006;**145**:77-a.

[9] Wren JD, Johnson KR, Crockett DM, Heilig LF, Schilling LM, Dellavalle RP. Uniform Resource Locator Decay in Dermatology Journals: Author Attitudes and Preservation Practices. *Arch Dermatol* 2006;**142**:1147-52.

[10] Evangelou E, Trikalinos TA, Ioannidis JPA. Unavailability of online supplementary scientific information from articles published in major journals. *FASEB J.* 2005;**19**:1943-4.

[11] Aronsky D, Madani S, Carnevale RJ, Duda S, Feyder MT. The Prevalence and Inaccessibility of Internet References in the Biomedical Literature at the Time of Publication. *J Am Med Inform Assoc* 2007;**14**:232-4.

[12] Eysenbach G,.Diepgen TL. Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. *BMJ* 1998;**317**:1496-500.

[13] Eysenbach G,.Trudel M. Going, Going, Still There: Using the WebCite Service to Permanently Archive Cited Web Pages. *J Med Internet Res* 2005;**7**:e60.