WebCite Technical Background and Best Practices Guide

This document provides an overview of the WebCite system, including best practices and technical details. It describes preferred methods for both initiating caching of content by the WebCite system, and also for building links to content cached by WebCite such that those links are both permanent and unambiguous.

Overview

Archiving

Content is stored in WebCite based on one of the following types of requests;

1. An explicit request from a user visiting http://www.webcitation.org/archive to cache the contents of a single URL. Note that pages archived via the WebCite bookmarklet fall into this category, as it uses the above URL behind the scenes. This method is intended to be used by authors during their primary research, and is the ideal method for initiating WebCite caching, as it enables a snapshot of the page in question to be made in the condition it existed when it was initially viewed by the citing author.
2. An explicit request from a user can also be obtained by accessing http://www.webcitation.org/archive from within an application (such as bibliographic software). Given the correct parameters, this URL will return an XML document that specifies the response. Outlined in the Technical Specifications section below are the parameter specifications as well as the DTD of the XML response. A sample XML response can be found in the Examples section and the Best Practices for Bibliographic Software Vendors briefly explains how to utilize this approach.
3. Coming Soon: Requests can be made in batch by submitting an XML document (via FTP or through the web site). A document submitted in this form can contain any number of archive requests.
4. A request to comb the outbound links from a given web page, either specified by URL or uploaded to the WebCite server. The outbound links from such a page are presented as a checklist to the user, who can then choose to archive the content of any of the outgoing links on such a page. This method is intended to be used during the pre-publishing phase of journal article preparation, in order to capture the content of cited web pages which the author may have not archived with WebCite during their primary research. This method is deficient in that the captured version of the cited page may differ from the version the author intended to cite if it has changed between the original access date and the article's processing date. However in cases where the original author did not include WebCite backed links for their references, this is nonetheless a better approach than simply not caching references at all.

Querying

Once the page(s) in question have been cached by WebCite, they can be accessed by users and publishers any number of times. Depending on the information a user has at hand, items cached by WebCite can be queried based on one of three methods:

1. By explicit snapshot ID. Every item added to the WebCite database (including web pages, PDF files, and included images or stylesheets) is assigned a unique numeric ID. These IDs are unique, unambiguous and idempotent, and thus represent the ideal way of querying a given resource cached by WebCite. However, the use of this method requires knowing the ID for a given resource, and so can't be used without premeditation. Upon completion of a archiving request, WebCite sends an email to the user who requested the archiving operation (or, in the case of an FTP uploaded file, to the pre-arranged technical contact for the doi prefix of the article in question)
2. By URL and referring article. This method allows for publishers which use WebCite as part of their pre-publishing workflow to easily create WebCite queries for their cached references with minimal co-ordination with WebCite before publishing. These queries are unambiguous, but are not necessarily guaranteed to be idempotent (the content of the URL may be re-cached by multiple submissions of a given page for combing).
3. By URL and date. When queried in this manner, WebCite finds all cached versions of the given URL, and sorts them by proximity to the given date. Although this allows for a certain 'fudge factor' with timestamps, it also means that these types of queries are inherently ambiguous, and are not guaranteed to be idempotent across queries. As such, these queries are intended to be used when the user has no information in hand other than the URL to query, and possibly the approximate date of the snapshot they would like to see.

These queries can be performed by accessing the user interface (the web site) or by accessing the URL from within other software (such as bibliographic software). The response in the latter case can be either the contents of the cached web page, or an XML document outlining the result of the query. Both the input and output specifications can be found in the Technical Specifications section below.

Best Practices

Best practice workflows for various users are as follows:

Best practices for authors

1. Authors install the WebCite bookmarklet in their browsers.
2. For every web based reference they wish to cite, they do the following:
    a. Use the WebCite bookmarklet to cause a copy of the page in question to be cached.
    b. Wait for the WebCite system to send a response email containing the URL to use to query the WebCite database for this just cached article (this URL is based on the 'explicit snapshot' model described above).

c. Include the original URL in the references section of the paper accompanied by the WebCite URL.

Best practices for publishers (submitting content to WebCite)

When processing a manuscript for publishing, publishers submit the raw document to WebCite for processing. Ideally, this submission is done via FTP, and uses a well defined (preferably XML based) schema for article data. The exact dialect used for this purpose should be agreed on ahead of time by the publisher and WebCite. Currently, we support

- (X)HTML documents
- NLM Journal Publishing DTD documents
- 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.

Best practices for publishers (providing WebCite content to readers)

The preferred method for providing WebCite content to readers is to include the WebCite URL along with the original URL for each of the references. Alternatively, publishers can simply inform their readers that all web references have been archived with WebCite and leave the onus on the reader to search for the archived version. However, since there may be multiple caches of a given URL, the reader has no straightforward way of identifying which one was cached for a given manuscript. The former method provides an unambiguous link to the appropriate archive.

A sample reference can be found in the Examples section below.

Best Practices for Bibliographic Software Vendors

When a user of the bibliographic software chooses to add a Web reference, it is recommended that the software automatically submit a request to archive the page to WebCite. Requests to archive a page must be made by executing http://www.webcitation.org/archive?parameters. Where parameters must be in the standard http style. Details of the parameters and examples can be found in the Technical Specification section below.

The vast majority of programming languages and environments provide the capabilities for executing URLs. One highly recommended free and open source software package which provides the libraries for doing exactly this is libcurl (http://curl.haxx.se/)

Technical Specifications

This section outlines some of the technical details that are important in order to effectively use the WebCite system.

Input Variables

When performing a query through direct execution of a URL or through the user interface, parameter specification is as follows:

Query: http://www.webcitation.org/query

Parameters : [url [date] [refdoi] | id] [returnxml]

url
the URL of the desired archived page
must be a fully qualified URL
must begin with http://, https:// or ftp://

date
optional
the date that the URL was archived. Responses will be
date can be in any textual English datetime description
date is parsed by the php function strtotime (http://ca.php.net/strtotime)

refdoi
optional
the DOI of the page that caused the archive
the DOI must be an exact match to the one submitted in an XML request or the one harvested from an uploaded XML file

id
the numeric WebCite identifier for the archived web page
no other parameters can be specified since id uniquely identifies the archive

returnxml
when specified (by setting =true) provides an XML response
the detail of the response format is outlined below

Archive: http://www.webcitation.org/archive

Parameters : url email

url
the URL of the desired archived page
must be a fully qualified URL
must begin with http://, https:// or ftp://

email
a valid email address where notifications can be sent

returnxml
when specified (by setting =true) provides an XML response
the detail of the response format is outlined below

Responses

Query responses where returnxml has been set result in an XML document with the following DTD:

```
<!ELEMENT queryresult ( resultset, error ) >

<!ELEMENT error ( #PCDATA ) >
<!ELEMENT resultset ( result+ ) >

<!ELEMENT result ( webcite_id, timestamp, original_url, webcite_url?,
webcite_raw_url? ) >

<!ATTLIST result status (success, failure_404,
failure_unsupportedfiletype) #REQUIRED >

<!ELEMENT original_url ( #PCDATA ) >
<!ELEMENT timestamp ( #PCDATA ) >
<!ELEMENT webcite_id ( #PCDATA ) >
<!ELEMENT webcite_raw_url ( #PCDATA ) >
<!ELEMENT webcite_url ( #PCDATA ) >
```

Archive requests where returnxml has been set result in XML document with the following DTD:

```
<!ELEMENT archiverequest ( resultset ) >
<!ELEMENT resultset ( result , error+) >

<!ELMENT error ( #PCDATA ) >
<!ATTLIST error type (email, url)>

<!ELEMENT result ( webcite_id, original_url, webcite_url, email ) >
<!ATTLIST result status (success ,  onlySuccessForNow,
futureTypesSupportedHere) #REQUIRED >

<!ELEMENT email ( #PCDATA ) >
<!ELEMENT original_url ( #PCDATA ) >
<!ELEMENT webcite_id ( #PCDATA ) >
<!ELEMENT webcite_url ( #PCDATA ) >
```

Examples

Citing using WebCite:

1. Centre for Global eHealth Innovation. URL: http://www.ehealthinnovation.org [accessed 2006 Feb 1] [WebCite Cache]
2. Friedman M. How to Cure Health Care. The Public Interest. 2001. URL: http://www.thepublicinterest.com/archives/2001winter/article1.html [accessed 2005 Nov 21] [WebCite Cache]

Executing URL

Query

http://www.webcitation.org/query?id=1138911916587475
http://www.webcitation.org/query?url=http://www.ehealthinnovation.org
http://www.webcitation.org/query?url=http://www.ehealthinnovation.org&date=today
http://www.webcitation.org/query?url=http://www.ehealthinnovation.org&date=2006-02-02
http://www.webcitation.org/query?url=http://www.ehealthinnovation.org&date=2006-02-02&returnxml=true
http://www.webcitation.org/query?url=http://www.ehealthinnovation.org&date=February%202,%202006

Archive

http://www.webcitation.org/archive?url=http://www.ehealthinnovation.org&email=jalperin@ehealthinnovation.org
http://www.webcitation.org/archive?url=http://www.ehealthinnovation.org&email=jalperin@ehealthinnovation.org&returnxml=true

XML responses

Query

http://www.webcitation.org/query?returnxml=true&url=http://www.ehealthinnovation.org

```
<queryresult>
    <resultset>
        <result status="success">
            <webcite_id>1138911916587475</webcite_id>
            <timestamp>2006-02-02 15:25:18</timestamp>
            <original_url>http://www.ehealthinnovation.org/</original_url>
            <webcite_url>
                http://www.webcitation.org/query?id=1138911916587475
            </webcite_url>
```

```
            <webcite_raw_url>
        http://www.webcitation.org/cache/73e53dd1f16cf8c5da298418d2a6e452870cf50e
            </webcite_raw_url>
        </result>
    </resultset>
</queryresult>
```

same URL but when there exist multiple archives in WebCite

```
<queryresult>
    <resultset>
        <result status="success">
            <webcite_id>1138911916587475</webcite_id>
            <timestamp>2006-02-02 15:25:18</timestamp>
            <original_url>http://www.ehealthinnovation.org/</original_url>
            <webcite_url>
                http://www.webcitation.org/query?id=1138911916587475
            </webcite_url>
            <webcite_raw_url>
        http://www.webcitation.org/cache/73e53dd1f16cf8c5da298418d2a6e452870cf50e
            </webcite_raw_url>
        </result>
        <result status="success">
            <webcite_id>1138914101684001</webcite_id>
            <timestamp>2006-02-02 16:01:43</timestamp>
            <original_url>http://www.ehealthinnovation.org/</original_url>
            <webcite_url>
                http://www.webcitation.org/query?id=1138914101684001
            </webcite_url>
            <webcite_raw_url>
        http://www.webcitation.org/cache/73e53dd1f16cf8c5da298418d2a6e452870cf50e
            </webcite_raw_url>
        </result>
    </resultset>
</queryresult>
```

when an archive request did not succeed (in this case a 404 error occurred when attempting to archive)
http://www.webcitation.org/query?returnxml=true&id=1138914355712941

```
<queryresult>
    <resultset>
        <result status="failure_404">
            <webcite_id>1138914355712941</webcite_id>
            <timestamp>2006-02-02 16:05:55</timestamp>
            <original_url>http://www.mistypedurl.com</original_url>
        </result>
    </resultset>
```

```
</queryresult>
```

Archive

http://stage.webcitation.org/archive?returnxml=true&url=http://www.ehealthinnovation.org&email=jalperin@ehealthinnovation.org

```
<archiverequest>
    <resultset>
        <result status="success">
            <webcite_id>1138911916587475</webcite_id>
            <original_url>http://www.ehealthinnovation.org</original_url>
            <webcite_url>
                http://www.webcitation.org/query?id=1138911916587475
            </webcite_url>
            <email>jalperin@ehealthinnovation.org</email>
        </result>
    </resultset>
</archiverequest>
```

http://stage.webcitation.org/archive?returnxml=true&url=http://www.ehealthinnovation.org

```
<archiverequest>
    <resultset>
        <error type="email">No email address was provided</error>
    </resultset>
</archiverequest>
```

Notice that even an erroneous URL will give back a successful return if the archive operation completed successfully
http://www.webcitation.org/archive?returnxml=true&url=http://www.mistypedurl.com&email=jalperin@ehealthinnovation.org

```
<archiverequest>
    <resultset>
        <result status="success">
            <webcite_id>1138914355712941</webcite_id>
            <original_url>http://www.mistypedurl.com</original_url>-
            <webcite_url>
                http://www.webcitation.org/query?id=1138914355712941
            </webcite_url>
            <email>jalperin@ehealthinnovation.org</email>
        </result>
    </resultset>
</archiverequest>
```