URL Checking

The URL Validation function:

- Identifies URLs in the document (regardless of whether they are already Word hyperlinks)

- Converts them to clickable hyperlinks, if they are not already clickable, by turning them into Word Hyperlink fields

- Automatically corrects common problems with special characters or spaces found in URLs

- Checks on the Web to determine whether the link is valid

Valid URLs—URLs that may be followed without redirection to the indicated website—are not flagged. URLs that return any HTTP error code other than 200 (i.e., ones with no HTTP error) will have a Word comment inserted to bring attention to the invalid URL and provide an indication of the error encountered in attempting to follow the link.

For example, a link that points to a nonexistent page will have a Word comment inserted indicating that the server returned an HTTP 404 ("Not Found") status code:

> http://www.inera.com/~kdkdkd
>
> Comment: URL Validation failed because the page http://www.inera.com/~kdkdkd does not exist (HTTP error 404).

URL Validation will also try to fix some cases of URLs that contain invalid characters or spaces, as sometimes happens during the editing process. For example, spaces sometimes appear mid-URL, which would cause a URL to fail in a browser. When possible, URL Validation will remove the incorrect spaces and add a comment such as:

> The URL "http://dtd.nlm.nih.gov/ faq.html" has been corrected. One or more spaces were removed.

Examples of other cleanup cases include replacing en dashes with hyphens, or similar characters with a simple tilde.

URL Validation is best run after Bibliographic References processing because URL conversion to clickable hyperlinks when they're found in reference lists interferes with eXtyles recognition of electronic references during Bibliographic References processing.

Redirected URLs

Experience has shown that the most common warning from URL Validation is that a URL has been redirected to a different final URL. There are many reasons for this, including:

- A website has been reorganized and the (wise) webmaster has automatically redirected old URLs to the new pages. In cases such as this, you may want to update the URL in the document to reflect the most current URL.

- A "vanity" URL has been used. For example, some drug manufacturers will register a domain for a drug name, but when you follow the URL to the website of the drug, you actually visit the manufacturer's landing page. In such cases, you might want to keep the vanity URL for the drug (despite the redirection warning) because the drug manufacturer could be acquired, and the ultimate landing page would change. In other words, when a URL is redirected, we recommend keeping in the document the URL that will likely have the longest life.

- Some URLs are always redirected. For example, URLs to articles on the New York Times website have a new and unique final URL on every single visit, even though the URL you click is identical. Redirection is performed in this case for the purpose of tracking web visitors. In such cases, the redirect warning can be ignored.

Excluded URLs

URL Validation does not check URLs that link to PubMed and CrossRef in reference lists because these links have likely been added by eXtyles.

Limitations of URL Validation

URL Validation has the following limitations:

- If a hyperlink has no comment after URL Validation, there's no guarantee that it points to a working page. Some websites redirect links for now-dead pages to a standard page that indicates a page has moved. For critical cases, there is still no substitute for a manual check of each page.

- A Word hyperlink with an underlying "mailto" property, indicating that the hyperlink is an email address, will not be validated.

Connection Error Codes

When URL Validation encounters a connection error, it inserts a warning with a corresponding error code:

Error returned	Error code in comment
Connection Failures	
Authoritative: Host not found	ERROR_INTERNET_NAME_NOT_RESOLVED
Network subsystem failed	ENETDOWN
No file handles available	EMFILE
No buffer space available	ENOBUFS
Handle is not a socket	ENOTSOCK
Not connected	ENOTCONN
Address not available	EADDRNOTAVAIL
Connection aborted	ECONNABORTED
Connection reset	ERROR_INTERNET_CONNECTION_RESET
Connection timed out	ETIMEDOUT
Connection refused	ECONNREFUSED
Failed to read from connection	EREAD
Failed to write to connection	EWRITE
Host down	EHOSDOWN
Host unreachable	EHOSTUNREACH
Non-authoritive: host not found or server failure	TRY_AGAIN
Non-recoverable: refused or not implemented	NO_RECOVERY
Internal error	EINTERNAL
Valid name, no data record for type	NO_DATA
Connection Status Errors	
URL is incorrect	HTTP Error 400
Restricted source	HTTP error 403
Does not exist	HTTP error 404
Does not exist	HTTP error 410