

URL Checking

The eXtyles URL Checking module:

- Identifies URLs in the Word document (whether or not 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 if the link is valid

Valid URLs, meaning those URLs which may be followed without redirection to the indicated webpage, are not flagged. URLs that return any HTTP error code other than 200 (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 non-existent 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 Checking 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 in mid-URL, which would cause a URL to fail. 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 ndashes with hyphens, or similar characters with a simple tilde.

Note: URL Checking is best run after reference processing because the process of converting URLs in reference lists into clickable Hyperlinks interferes with eXtyles reference processing recognition of electronic references.

Redirected URLs

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

- A web site 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.
- 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 web site of the drug, you actually visit the manufacturer's landing page for the drug as located on the manufacturer's web site. In such cases, you may 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 web site have a new and unique final URL on every single visit, even though the URL you click is identical. This is done for the purpose of tracking web visitors. In such cases, the redirect warning can be ignored.

Excluded URLs

URL Checking does not check URLs to PubMed and CrossRef links in reference lists since these links have likely been added by eXtyle and should not be validated by this process.

Limitations of URL Checking

URL Checking has the following limitations:

- If a Hyperlink does not have any comment, it does not *guarantee* that it points to a working page. Some web sites will 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.