Internet Engineering Task Force (IETF)
Request for Comments: 6596
Category: Informational
ISSN: 2070-1721

M. Ohye
J. Kupke

April 2012

The Canonical Link Relation

Abstract

RFC 5988 specifies a way to define relationships between links on the web. This document describes a new type of such a relationship, "canonical", to designate an Internationalized Resource Identifier (IRI) as preferred over resources with duplicative content.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6596.

1. Introduction

The canonical link relation specifies the preferred IRI from resources with duplicative content. Common implementations of the canonical link relation are to specify the preferred version of an IRI from duplicate pages created with the addition of IRI parameters (e.g., session IDs) or to specify the single-page version as preferred over the same content separated on multiple component pages.

In regard to the link relation type, "canonical" can be described informally as the author's preferred version of a resource. More formally, the canonical link relation specifies the preferred IRI from a set of resources that return the context IRI's content in duplicated form. Once specified, applications such as search engines can focus processing on the canonical, and references to the context (referring) IRI can be updated to reference the target (canonical) IRI.

2. Notational Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. The Canonical Link Relation

The target (canonical) IRI MUST identify content that is either duplicative or a superset of the content at the context (referring) IRI. Authors who declare the canonical link relation ought to anticipate that applications such as search engines can:

- Index content only from the target IRI (i.e., content from the context IRIs will be likely disregarded as duplicative).
- Consolidate IRI properties, such as link popularity, to the target IRI.
- Display the target IRI as the representative IRI.

The target (canonical) IRI MAY:

- Specify a relative IRI (see [RFC3986], Section 4.2).
- Be self-referential (context IRI identical to target IRI).
- Exist on a different hostname or domain.

- Have different scheme names, such as "http" to "https" or "gopher" to "ftp".

- Be a superset of the content at the context IRI.

  * As an example, each component page (e.g., page-1.html, page-2.html) of a multi-page article MAY specify the "view-all" version (e.g., page-all.html), the superset of their content, as the target IRI. This is because the content from each component page is contained within the view-all version. Given this implementation, applications can mark page-1.html and page-2.html as duplicates of page-all.html, process content only from page-all.html, and disregard the component pages. All references can then be made to the view-all version (page-all.html, the target IRI), and no content will have been lost in this process.

  * Using the same example above, page-2.html SHOULD NOT designate page-1.html as the target (canonical) IRI because this may cause a loss of data. When page-2.html designates page-1.html as the canonical, only content from the target IRI, page-1.html, will be processed. page-2.html may be marked as a duplicate of page-1.html and its content disregarded.

- Be the source IRI of a temporary redirect. For HTTP, this refers to status codes 302, 303, or 307 (Sections 10.3.3, 10.3.4, and 10.3.8, respectively, of [RFC2616]).

To better ensure that applications properly handle the canonical link relation, administrators ought to consider the following guidelines:

- Specify only one canonical link relation for a resource. (It would be confusing to consider/label/designate more than one IRI as authoritative.)

- Avoid designating the target (canonical) as:

  * The source IRI of a permanent redirect (for HTTP, this refers to 300 and 301 response codes, defined in Sections 10.3.1 and 10.3.2 of [RFC2616]).

  * An IRI that also specifies a canonical link relation to an IRI other than itself.

  * An IRI that returns an error code, such as a 4xx response in HTTP (Section 10.4 of [RFC2616]).

* The first page of a multi-page article or multi-page listing of items (since the first page is not duplicative or a superset of the context IRI). For example, page-2.html and page-3.html of an article SHOULD NOT specify page-1.html as the canonical. This may cause a loss of data from page-2.html and page-3.html as they will be marked duplicative of page-1.html with only content from page-1.html being processed.

When the canonical link relation is declared improperly, such as creating chained canonicals (i.e., target IRI specifies the source IRI of a permanent redirect) or designating a target IRI that returns a 4xx response, applications can use their own heuristics when processing the resource. For instance, an application can choose to ignore any improper canonical designation and continue to process the remaining content on a page.

4. Examples

The following example illustrates:

- Three IRIs that serve duplicate content.
- One IRI that is the canonical or "preferred version".
- Two IRIs with additional query parameters, making them the non-preferred version of the content (duplicates). The canonical link relation is therefore specified on these duplicates.

If the preferred version of a IRI and its content exists at:

http://www.example.com/page.php?item=purse

Then duplicate content IRIs such as:

http://www.example.com/page.php?item=purse&category=bags
http://www.example.com/page.php?item=purse&category=bags&sid=1234

may designate the canonical link relation in HTML as specified in [REC-html401-19991224]:

```
<link rel="canonical"
        href="http://www.example.com/page.php?item=purse">
```

or as a relative IRI:

```
<link rel="canonical" href="page.php?item=purse">
```

or alternatively, in the HTTP header field as specified in Section 5 of [RFC5988]:

Link: <http://www.example.com/page.php?item=purse>; rel="canonical"

This signals to applications, such as search engines, that these are duplicates of the target (canonical) IRI:

http://www.example.com/page.php?item=purse.

Applications may then select the canonical value as the display IRI (such as in search results), and additional IRI properties such as indexing and ranking signals can be transferred as well.

5. Recommendations

Before adding the canonical link relation, verification of the following is RECOMMENDED:

1. The content of the context IRI is duplicated within the content of the target (canonical) IRI.

2. For HTTP, permanent HTTP redirects (Section 10.3.2 of [RFC2616]), the traditional strong indicator that a IRI’s content has been permanently moved, could not be implemented in place of the canonical link relation.

3. In the case where the target (canonical) IRI is a superset of content from the context IRI (i.e., the case where page-1.html and page-2.html designate page-all.html as the canonical), that the user experience is strongly taken into consideration, both in regard to possible increased load time and potential complexity in navigation.

6. IANA Considerations

IANA has registered the Canonical Link Relation below as per [RFC5988].

Relation Name:

canonical

Description:

Designates the preferred version of a resource (the IRI and its contents).

Reference:

This specification.

Notes:

None.

Application Data:

None.

7. Security Considerations

When a site is compromised, the canonical link relation can be implemented with malicious intent to designate the attacker's IRI as the preferred version of the content. While this technique is largely unnoticeable to humans, automated programs may cluster the compromised resource as duplicative of the attacker's target IRI, transferring properties such as link popularity away from the compromised resource to the attacker's designated canonical. (Naturally, even a site that is not compromised could provide inaccurate or misleading information about which URI is canonical.)

8. Internationalization Considerations

Internationalization considerations for link relations are provided in Section 8 of [RFC5988].

9. Normative References

[REC-html401-19991224] Raggett, D., Le Hors, A., and I. Jacobs, "HTML 4.01 Specification", W3C Recommendation REC-html401-19991224, December 1999, <http://www.w3.org/TR/1999/REC-html401-19991224>.

Latest version available at <http://www.w3.org/TR/html401>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2616] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, January 2005.

[RFC5988] Nottingham, M., "Web Linking", RFC 5988, October 2010.

Appendix A. Implementations

Automated programs that implement functionality with regard for the canonical link relation include:

- Google, canonical link relation HTML and HTTP header support, within the same domain and across domains:
  - <http://googlewebmastercentral.blogspot.com/2009/02/specify-your-canonical.html>
  - <http://googlewebmastercentral.blogspot.com/2011/06/supporting-relcanonical-http-headers.html>
  - <http://googlewebmastercentral.blogspot.com/2009/12/handling-legitimate-cross-domain.html>
- Yahoo, canonical link relation HTML support within the same domain:
  - <http://www.ysearchblog.com/2009/02/12/fighting-duplication-adding-more-arrows-to-your-quiver/>
- Bing, canonical link relation HTML support within the same domain:
  - <http://www.bing.com/community/site_blogs/b/webmaster/archive/2009/02/12/partnering-to-help-solve-duplicate-content-issues.aspx>

Authors' Addresses

Maile Ohye

EMail: maileohye@gmail.com
URI: http://maileohye.com/

Joachim Kupke

EMail: joachim@kupke.za.net