RFC 6596 : The Canonical Link Relation

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 9 avril 2012

Date de publication du RFC : Avril 2012

https://www.bortzmeyer.org/6596.html

Depuis le RFC 5988 [1], il existe un mécanisme standard pour exprimer les types des liens entre deux ressources sur le Web. Ce très court RFC spécifie un nouveau type de lien, `canonical`, qui permet d'indiquer quel est l'URI canonique d'une ressource Web.

Le but de ces liens est de permettre d'exprimer l'idée « Quel que soit l'URI que vous avez utilisé pour arriver sur cette ressource, sachez que l'URI **canonique**, la référence, est celui indiqué par ce lien. » Cela permet notamment à un moteur de recherche de n'indexer les ressources que sous l'URI canonique (au cas où du contenu soit dupliqué sous plusieurs URI). Cela permet également à un navigateur de ne mémoriser que l'URI canonique, sans d'éventuels paramètres (options d'affichage, identificateurs de session et autres trucs qui viennent souvent polluer les URI).

L'auteur qui place un lien `canonical` doit donc veiller à ce que l'URI canonique désigne bien une ressource qui mérite ce titre (identique à la ressource de départ, ou un sur-ensemble de celle-ci, ce dernier cas est explicitement autorisé par le RFC). Voir la section 5 qui donne de bons conseils aux auteurs.

Par contre, l'URI canonique ne doit notamment pas :

— Être la source d'une redirection HTTP (sinon, par définition, il n'est pas canonique).
— Ne doit pas spécifier un URI canonique autre que lui-même (en d'autres termes, pas le doit de faire des chaînes).
— Ne doit évidemment pas renvoyer 404 lorsqu'on essaie d'y accéder.

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc5988.txt`

La section 4 donne des exemples concrets. Si la version canonique d'une ressource est désignée par l'URI `http://www.example.com/page.php?item=purse`, alors les URI `http://www.example.com/page.ph` ou `http://www.example.com/page.php?item=purse&category=bags&sid=1234` qui sont des URI possibles de la même ressource peuvent indiquer `http://www.example.com/page.php?item=purse` comme canonique.

Pour cela, deux techniques, le classique lien HTML avec l'attribut `rel` :

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

Il est également utilisable en version relative :

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

Et la deuxième technique (pratique notamment pour les ressources qui ne sont pas en HTML, une image, par exemple), l'en-tête HTTP (section 3 du RFC 8288) :

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

Pour prendre un exemple réel, si on demande à Wikipédia (qui fait face à des homonymies nombreuses) l'URL `<http://fr.wikipedia.org/wiki/M%C3%A9lenchon>`, on est redirigé vers la page sur Jean-Luc Mélenchon qui contient :

```
<link rel="canonical" href="/wiki/Jean-Luc_M%C3%A9lenchon" />
```

qui indique que la page canonique est celle avec le nom complet.

Le nouveau type `canonical` est désormais enregistré à l'IANA `<https://www.iana.org/assignments/link-relations/link-relations.xml>`.

Petit avertissement de sécurité (section 7). Si une ressource est modifiée par un attaquant, il peut mettre un lien vers un URI canonique de son choix. Bien sûr, il pourrait aussi massacrer complètement la ressource. Mais le changement d'URI canonique est discret et risquerait de ne pas être noté par un humain, alors même que certaines implémentations en tiendraient compte, par exemple pour apport du trafic à l'URI de l'attaquant.

Qui gère aujourd'hui ce type de liens ? Chez les moteurs de recherche, Google le fait (voir leurs articles « *"Specify your canonical"* `<http://googlewebmastercentral.blogspot.fr/2009/02/specify-your-canonical.html>` », « *"Supporting rel="canonical" HTTP Headers"* `<http://googlewebmaster-blogspot.fr/2011/06/supporting-relcanonical-http-headers.html>` », l'article de conseils pratiques « *"About rel="canonical""* `<http://support.google.com/webmasters/bin/answer.py?hl=en&answer=139394>` » et enfin « *"Handling legitimate cross-domain content duplication"* `<http://googlewebmastercentral.blogspot.fr/2009/12/handling-legitimate-cross-domain.html>` »). Pareil chez Yahoo (« *"Fighting Duplication : Adding more arrows to your quiver"* `<http://www.ysearchblog.com/2009/02/12/fighting-duplication-adding-more-arrows-to-your-quiver/>` ») et Bing (« *"Partnering to help solve duplicate content issues"* `<http://www.bing.com/community/site_blogs/b/webmaster/archive/2009/02/12/partnering-to-help-solve-duplicate-content-i-aspx>` »). Par contre, je ne sais pas si les sites de *"bookmarking"* comme del.icio.us ou SeenThis `<https://www.bortzmeyer.org/seenthis.html>` font ce travail de canonicalisation.