RFC 8574 : cite-as: A Link Relation to Convey a Preferred URI for Referencing

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 18 avril 2019

Date de publication du RFC : Avril 2019

https://www.bortzmeyer.org/8574.html

Ce RFC décrit un nouveau type de liens hypertexte permettant d'indiquer l'URI sous lequel on préfère qu'une ressource soit accédée, à des fins de documentation ou de citation précise.

Imaginons que vous accédez à un article scientifique en ligne. Vous utilisez un URI qui identifie cet article. Vous voulez ensuite citer cet article dans un de vos textes. Allez-vous utiliser l'URI d'accès ? Ce n'est pas forcément le meilleur, par exemple ce n'est pas forcément le plus stable sur le long terme. Le lien « cite avec » permet au serveur d'indiquer l'URI le plus pertinent pour une citation.

Ce RFC s'appuie sur la formalisation du concept de lien faite dans le RFC 8288[1]. « Contexte » et « cible » sont donc utilisés comme dans cette norme, le contexte d'un lien étant le point de départ et la cible l'arrivée. En prime, notre RFC 8574 définit deux termes, l'**URI d'accès**, celui utilisé pour accéder à une ressource (par exemple une page Web) et l'**URI de référence**, celui qu'il faudrait utiliser pour la citation.

La section 3 du RFC décrit quelques scénarios d'usage. Le premier est celui des identificateurs stables. Normalement, lorsque le ou la webmestre est compétent(e) et sérieux(se), les URI sont stables, comme précisé dans l'article « "*Cool URIs don't change*" <https://www.w3.org/Provider/Style/URI.html>. Mais, en pratique, beaucoup de webmestres sont incompétents ou paresseux. Cela a mené à des « solutions » fondées sur la redirection, où il apparait une différence entre URI d'accès et URI de référence. C'est le cas avec des techniques comme les DOI (« "*use the Crossref DOI URL as the permanent [reference] link*" <https://blog.crossref.org/display-guidelines/> »), PURL ou ARK. Dans les trois cas, au lieu de gérer proprement les URI, on utilise un redirecteur supposé plus stable

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc8288.txt

(alors que rien ne le garantit) et on souhaite utiliser comme URI de référence l'URI du redirecteur (donnant ainsi des pouvoirs démesurés à des organisations privées comme l'IDF, qui matraque régulièrement `<https://arxiv.org/abs/1602.09102>` l'importance de n'utiliser que leurs identificateurs).

Un autre scénario d'usage est celui des ressources versionnées. C'est par exemple le cas de Wikipédia. La page Wikipédia sur l'incendie de Notre-Dame de Paris change souvent en ce moment. Comme toutes les pages Wikipédia, chaque version à un identificateur, et on peut se référer à une version particulier. Si renvoie à la dernière version, sans cesse en mouvement, renvoie uniquement à la toute première version, très sommaire et à une version intermédiaire déjà très fournie.

Souvent, quand on veut citer un article de Wikipédia, on veut se mettre à l'abri de changements ultérieurs, pas forcément positifs, et on souhaite donc citer exactement une version particulière. On clique donc sur « Lien permanent » (ou bien « Voir l'historique » puis sur la date la plus récente) pour avoir l'URI de la version qu'on vient de regarder. (Notez aussi le très utile lien « Citer cette page ».)

Troisième cas d'usage cité, celui des identifiants sur les réseaux sociaux. M. Toutlemonde a typiquement plusieurs pages le décrivant sur ces réseaux (dans mon cas, , , , et sans doute bien d'autres que j'ai oubliés, et ceux que j'ai eu la flemme de faire, comme FOAF). Or, on pourrait souhaiter décider qu'un de ces profils est meilleur que les autres, par exemple parce qu'il est plus directement contrôlé par l'utilisateur, ou mieux maintenu. Il serait alors intéressant d'indiquer lors de l'accès à chacun des autres profils quel est le profil de référence. (Le RFC est très irréaliste là-dessus : je vois mal un réseau « social » capitaliste permettre à ses utilisateurs de dire « allez plutôt voir là ».)

Enfin, un dernier cas d'usage est celui d'une publication composée de plusieurs ressources (par exemple un livre où chaque chapitre est accessible séparement, avec son propre URI). On souhaite alors que l'accès à chaque ressource indique, à des fins de citation, l'URI de référence (par exemple la page d'accueil).

La section 4 du RFC présente la solution : un nouveau type de lien, `cite-as`, qui permet de dire quel est l'URI de référence. (Le RFC recommande d'ailleurs que cet URI soit un URL : le but est d'accéder à la ressource !) Il est évidemment recommandé qu'il n'y ait qu'un seul lien de type `cite-as`. Ce lien n'interdit pas d'utiliser d'autres URI, il indique seulement quel est l'URI que le webmestre souhaite voir utilisé dans les références webographiques. `cite-as` est désormais dans le registre IANA des types de liens `<https://www.iana.org/assignments/link-relations/link-relations.xml>`.

La section 6 du RFC donne des exemples concrets, puisque les liens peuvent se représenter de plusieurs façons. Par exemple, l'article de PLOS One auquel vous accédez en pourrait contenir, en HTML, le lien avec l'attribut `rel="cite-as"` :

```
<link rel="cite-as"
          href="https://doi.org/10.1371/journal.pone.0171057" />
```

Cela indiquerait que les auteurs préfèrent être cités par le DOI (une mauvaise idée, mais c'est une autre histoire).

Autre exemple de syntaxe concrète pour les liens, imaginé pour arXiv, pour des articles avec versionnement, un lien dans un en-tête HTTP pour , qui pourrait indiquer qu'on est en train de regarder la première version, « v1 » (il existe une « v2 », essayez) :

```
HTTP/1.1 200 OK
Date: Sun, 24 Dec 2017 16:12:43 GMT
Content-Type: text/html; charset=utf-8
Link: <https://arxiv.org/abs/1711.03787v1> ; rel="cite-as"
```

Comme arXiv garde les différentes versions successives d'un article, cela permettrait de récupérer la version actuelle tout en sachant comment la référencer.

Revenons au HTML pour l'exemple des profils sur les réseaux sociaux, imaginons un utilisateur, John Doe, qui place dans le code HTML de sa page personnelle un lien vers son profil FOAF en Turtle :

```
<html>
    <head>
    ...
     <link rel="cite-as" href="http://johndoe.example.com/foaf"
           type="text/turtle"/>
    ...
    </head>
    <body>
...
```

Et un dernier exemple, celui d'une publication composée de plusieurs ressources. Ici, l'exemple est Dryad une base de données scientifiques qui permet l'accès à des fichiers individuels, mais où chaque jeu de données a un identificateur canonique. En HTTP, cela donnerait, lorsqu'on accès à (un fichier CSV qui fait partie de cette base de données) :

```
HTTP/1.1 200 OK
Date: Tue, 12 Jun 2018 19:19:22 GMT
Last-Modified: Wed, 17 Feb 2016 18:37:02 GMT
Content-Type: text/csv;charset=ISO-8859-1
Link: <https://doi.org/10.5061/dryad.5d23f> ; rel="cite-as"
```

Le fichier CSV est membre d'un jeu de données plus général, dont l'URI de référence est .

Ainsi, dans un monde idéal, un logiciel qui reçoit un lien `cite-as` pourrait :

— Lorsqu'il garde un signet, utiliser l'URI de référence,
— Identifier plusieurs URI d'accès comme ayant le même URI de référence, par exemple à des fins de comptage,
— Indexer les ressources par plusieurs URI.

D'autres solutions avaient été proposées pour résoudre ce problème de l'URI de référence. La section 5 de notre RFC les énumère. Il y avait notamment cinq autres types de liens qui auraient peut-être pu convenir, `alternate`, `duplicate`, `related`, `bookmark` et `canonical`.

Les trois premiers sont vite éliminés. `alternate` (RFC 4287) décrit une autre représentation de la même ressource (par exemple la même vidéo mais encodée différemment). `duplicate` (RFC 6249) désigne une reproduction identique (et cela ne traite donc pas, par exemple, le cas d'une publication composée de plusieurs ressources). Quant à `related` (RFC 4287), sa sémantique est bien trop vague. Un article des auteurs du RFC décrit en détail <http://ws-dl.blogspot.com/2016/11/2016-11-07-linking-to-persistent.html> les choix de conceptions et explique bien le problème. (Je trouve cet article un peu gâché par les affirmations sans preuves comme quoi les DOI seraient « permanents ». Si le registre disparait ou fait n'importe quoi, il y aura le même problème avec les DOI qu'avec n'importe quelle autre famille d'identificateurs.)

Le cas de `bookmark` (normalisé par le W3C <https://www.w3.org/TR/2017/REC-html52-20171214/>) est plus compliqué. Il est certainement proche en sémantique de `cite-as` mais ne peut pas être présent dans les en-têtes HTTP ou dans la tête d'une page HTML, ce qui en réduit beaucoup l'intérêt. Le cas compliqué de `bookmark` est décrit dans un autre article des auteurs du RFC <http://ws-dl.blogspot.com/2017/08/2017-08-26-relbookmark-also-does-not.html>.

Enfin, le cas de `canonical` (RFC 6596). Ce dernier semble trop restreint d'usage pour les utilisations prévues pour `cite-as`. Et il n'a pas vraiment la même sémantique. Par exemple, pour les ressources versionnées, `canonical` indique la plus récente, exactement le contraire de ce qu'on voudrait avec `cite-as`. Et c'est bien ainsi que l'utilise Wikipédia. Si je récupère :

```
<link rel="canonical" href="https://fr.wikipedia.org/wiki/Incendie_de_Notre-Dame_de_Paris"/>
```

On voit que `canonical` renvoie à la dernière version. Le cas de `canonical` fait lui aussi l'objet d'un article détaillé <http://ws-dl.blogspot.nl/2017/08/2017-08-07-relcanonical-does-not-mean.html>.

Je n'ai pas mis de tels liens sur ce blog, ne voyant pas de cas où ils seraient utiles.