

RFC 5147 : URI Fragment Identifiers for the text/plain Media Type

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 10 Avril 2008

Date de publication du RFC : Avril 2008

<http://www.bortzmeyer.org/5147.html>

Les URI disposent depuis longtemps d'un moyen de désigner une **partie** d'un document, par les biais des « identificateurs de fragments », précédés d'un dièse, comme par exemple `http://www.example.org/foobar.html#section3`. Entièrement interprétés par le client Web, ces identificateurs étaient définis seulement pour XML et HTML et notre RFC ajoute le texte brut.

La façon classique d'utiliser ces identificateurs est de modifier le document **cible** pour y ajouter un élément à viser, par exemple :

```
<p id="avertissement-sante">Fumer est dangereux pour la santé.</p>
```

fera qu'un navigateur pourra, lorsqu'on lui demandera `http://www.example.org/document.html#avertissement-sante`, aller directement à ce paragraphe. On peut aussi utiliser, au lieu d'un simple nom comme `avertissement-sante`, utiliser une expression XPath, si le navigateur accepte XPointer. En outre, avec XLink, on peut même pointer vers un document qu'on ne peut ou ne veut modifier, ce qui évite de devoir ajouter des attributs `id` ou `name`.

Le texte brut (normalisé dans MIME par le RFC 2046¹) n'étant pas, lui, structuré, on ne pouvait pas compter sur des « marques » particulières. D'où le choix de notre RFC de compter les caractères ou bien les lignes.

¹Pour voir le RFC de numéro NNN, <http://www.ietf.org/rfc/rfcNNN.txt>, par exemple <http://www.ietf.org/rfc/rfc2046.txt>

C'est ainsi que les URI de ce RFC ressembleront à `http://library.example/2007/11/5311.txt#char=6234` (le curseur du navigateur ira sur le 6234ème caractère) ou bien `http://library.example/2004/03/228.txt#line=534,580` (le navigateur affichera le texte compris entre la 534ème ligne et la 580ème).

Mais les documents accessibles sur le Web tendent à changer. De tels identificateurs de fragments sont clairement moins robustes que ceux utilisant un attribut du document XML (comme `avertissement-sante` plus haut), ou bien que ceux utilisant une expression XPath. Comment détecter le cas où un paragraphe a été ajouté, faussant ainsi les liens ? Cette norme permet d'ajouter à l'URI un mécanisme de contrôle d'intégrité, en indiquant la **longueur** ou bien la somme de contrôle MD5 (RFC 1321) du document. Si elles ne correspondent pas, le navigateur sait qu'il ne doit pas utiliser l'identificateur de fragment. Ainsi, `http://www.myblog.example/2006-12/mespensees.php#line=50;length=19576,UTF-8` amènera le navigateur à compter le nombre de caractères du document et à ignorer l'identificateur de fragment s'il ne trouve pas 19576 caractères.

Notre RFC doit aussi traiter des problèmes plus subtils comme la façon de compter les fins de ligne du document (il existe plusieurs façons de les représenter, et la norme décide qu'une fin de ligne compte toujours pour **un** caractère).

Un autre problème subtil est la nécessité de tenir compte de l'encodage des caractères. En effet, `char=` compte bien le nombre de caractères, pas le nombre d'octets.

Deux implémentations existent, une dans Amaya et une en Python, incomplète mais quand même utile : (en ligne sur `http://www.bortzmeyer.org/files/TextFrag.py`).

À noter aussi une présentation par un des auteurs (`http://dret.typepad.com/dretblog/2007/11/fragment-identi.html`), un amusant et intéressant article de Tim Bray, "*Matthew-6:9.txt#line=1*" (`http://www.tbray.org/ongoing/When/200x/2007/12/03/Text-Plain-Hash`) et un intéressant texte de Karl Dubost (`http://www.la-grange.net/2010/01/05/compter-texte`) élargissant le problème.