RFC 5987 : Character Set and Language Encoding for Hypertext Transfer Protocol (HTTP) Header Field Parameters

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 23 août 2010

Date de publication du RFC : Août 2010

https://www.bortzmeyer.org/5987.html

Les requêtes et réponses du protocole HTTP incluent des **en-têtes** (comme `User-Agent:` ou `Content-Disposition:` avec des valeurs, qui ne pouvaient se représenter directement qu'avec les caractères du jeu ISO 8859-1. Comme MIME, dans le RFC 2231 [1], prévoyait un mécanisme très riche pour encoder les en-têtes du courrier électronique, ce RFC 5987 réutilise ce mécanisme pour HTTP (plusieurs en-têtes l'utilisaient déjà, de manière pas vraiment officielle). Pour le corps du message (voir par exemple le RFC 7578), rien ne change. Ce RFC a depuis été remplacé par le RFC 8187.

Cette restriction à Latin-1 vient de la norme HTTP, le RFC 2616, dans sa section 2.2, qui imposait l'usage du RFC 2047 pour les caractères en dehors de ISO 8859-1. (Le RFC 7230 a changé cette règle depuis.)

Notre RFC peut être résumé en disant qu'il spécifie un **profil** du RFC 2231. Ce profil est décrit en section 3, qui liste les points précisés par rapport au RFC 2231. Tout ce RFC n'est pas utilisé, ainsi le mécanisme en section 3 du RFC 2231, qui permettait des en-têtes de plus grande taille, n'est pas importé (section 3.1 de notre RFC).

En revanche, la section 4 du RFC 2231, qui spécifiait comment indiquer la langue dans laquelle était écrite la valeur d'un en-tête est repris pour les paramètres dans les en-têtes. Ainsi, (section 3.2.2), voici un en-tête, avec un paramètre `title` traditionnel en pur ASCII :

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc2231.txt`

```
X-information: news; title=Economy
```

et en voici un avec les possibilités de notre RFC pour permettre les caractères [Caractère Unicode non montré [2]] et € (ce dernier n'existant pas dans Latin-1) :

```
X-information: news; title*=UTF-8''%c2%a3%20and%20%e2%82%ac%20rates
```

Par rapport au RFC 2231, deux jeux de caractères sont décrétés obligatoires (ISO 8859-1 et UTF-8) et la mention du jeu utilisée est également désormais obligatoire (section 3.2 de notre RFC). La langue elle-même est indiquée par une **étiquette**, selon la syntaxe du RFC 5646. Du fait de ces possibilités plus riches que celles prévues autrefois pour HTTP, les paramètres qui s'en servent doivent se distinguer, ce qui est fait avec un astérisque avant le signe égal (voir l'exemple ci-dessus). La valeur du paramètre inclus donc le jeu de caractères et l'encodage (obligatoire), la langue (facultative, elle n'est pas indiquée dans l'exemple ci-dessus) et la valeur proprement dite.

Voici un exemple incluant la langue, ici l'allemand (code `de`) :

```
X-quote: theater;
    sentence*=UTF-8'de'Mit%20der%20Dummheit%20k%C3%A4mpfen%20G%C3%B6tter%20selbst%20vergebens.
```

La section 4 couvre ensuite les détails pratiques pour les normes qui décrivent un en-tête qui veut utiliser cette possibilité. Par exemple, la section 4.3 traite des erreurs qu'on peut rencontrer en décodant et suggère que, si deux paramètres identiques sont présents, celui dans le nouveau format prenne le dessus. Par exemple, si on a :

```
X-information: something; title="EURO exchange rates";
               title*=utf-8''%e2%82%ac%20exchange%20rates
```

le titre est à la fois en ASCII pur et en UTF-8, et c'est cette dernière version qu'il faut utiliser, même si normalement il n'y a qu'un seul paramètre `title`.

Ceux qui s'intéressent à l'histoire du développement de cette nouvelle norme pourront regarder les minutes de la réunion IETF 72 <http://tools.ietf.org/wg/httpbis/minutes?item=minutes72.html>. Ces paramètres étendus sont déjà mis en œuvre dans Konqueror (à partir de la 4.4.1), Firefox et Opera. Des tests plus détaillés sont présentés en <http://greenbytes.de/tech/tc2231>.

2. Car trop difficile à faire afficher par LaTeX