RFC 7111 : URI Fragment Identifiers for the text/csv Media Type

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 17 janvier 2014

Date de publication du RFC : Janvier 2014

https://www.bortzmeyer.org/7111.html

Le classique format CSV est normalisé dans le RFC 4180[1]. Lorsqu'un URI désigne un fichier CSV, comment indiquer une partie spécifique du fichier ? Jusqu'à présent, il n'y avait pas de mécanisme pour cela. Ce nouveau RFC comble ce manque et définit une syntaxe ("*fragment identifier*") pour désigner une ligne particulière, une colonne spécifique, voir une seule cellule d'un fichier CSV. `http://www.example.org/data.csv` fonctionnera désormais, pour indiquer qu'on veut sauter à la douzième ligne du fichier `data.csv`.

Les ressources au format CSV sont identifiées par le type `text/csv`. Ce type a deux paramètres, `charset` qui indique l'encodage et `header` qui indique si l'en-tête facultatif de CSV est présent ou pas (cet en-tête donne le nom des colonnes). Rappelons que la syntaxe et la sémantique des identificateurs de fragment dans un URI dépendent du type de la ressource. Par exemple, « la cellule en 2ème ligne et 5ème colonne » a un sens en CSV mais pas en texte brut, alors que « le 292ème caractère » a un sens en texte brut (RFC 5147 pour les identificateurs de fragments du texte brut).

Attention, les identificateurs de fragments sont interprétés uniquement par le client Web. Leur bon fonctionnement dépend donc du déploiement du logiciel nécessaire chez les clients, ce qui va prendre du temps (je me demande combien de navigateurs Web gèrent le RFC 5147...). Le serveur n'a même pas un moyen de savoir si le client gère ces identificateurs de fragment. Mais ce n'est pas forcément un problème grave : si le client ne comprend pas la syntaxe des identificateurs de fragments pour CSV, il charge quand même le fichier CSV, il ne peut juste pas aller directement à la partie intéressante. Le repli se fait donc en douceur. Pour l'instant, à ma connaissance, il n'existe pas de logiciel qui gère ces fragments CSV.

Bref, voyons maintenant la sémantique des identificateurs de fragments pour CSV (section 2 du RFC). Le RFC contient un exemple avec des données de température mais je vais plutôt prendre les données BGP de la récente panne de l'Internet à Saint-Pierre-et-Miquelon <https://www.bortzmeyer.org/panne-saint-pierre-miquelon.html>. Les données originales étaient au format MRT (RFC 6396) mais bgpdump <https://bitbucket.org/ripencc/bgpdump/> peut les traduire en quasi-CSV, il suffit d'un petit coup de sed ensuite :

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc4180.txt`

```
% bgpdump -m updates.20140104.1930.bz2 | sed 's/|/,/g' > updates.20140104.1930.csv
```

Le fichier utilisé pour les exemples est (en ligne sur `https://www.bortzmeyer.org/files/panne-spm.csv`) (oui, il contient une ligne optionnelle d'en-tête au début, mais ne l'indique pas dans le type MIME, violant ainsi la section 3 du RFC 4180). Chaque ligne comprend notamment l'heure (deuxième colonne, vous pouvez traduire 1388864007 avec `date -u --date=@1388864007`), le type de message (troisième colonne, avec A = "*Announce*" et W = "*Withdraw*"), le pair qui a envoyé le message (quatrième colonne), le préfixe d'adresse IP concerné (sixième colonne), etc.

D'abord, on peut sélectionner une **ligne** particulière, la première étant numérotée 1 (pas zéro, attention, les informaticiens). La ligne d'en-tête, facultative, compte pour une ligne normale si elle est présente. Ainsi, la ligne 5 est `BGP4MP,1388864007,W,195.66.224.32,3257,70.36.8.0/22`. On peut aussi indiquer un intervalle des lignes. Ainsi, 5-7 est :

```
BGP4MP,1388864007,W,195.66.224.32,3257,70.36.8.0/22
BGP4MP,1388864007,A,195.66.225.76,251,70.36.8.0/22,251 3257 11260 3695,IGP,195.66.225.76,0,0,3257:4000 3257:
BGP4MP,1388864007,W,195.66.236.32,3257,70.36.8.0/22
```

Le signe * indique « la dernière ligne » donc * désigne :

```
BGP4MP,1388864142,W,195.66.224.215,31500,70.36.12.0/22
```

On peut aussi désigner une **colonne** donnée. Là aussi, on part de 1. La sixième colonne sera :

```
70.36.8.0/22
70.36.8.0/22
70.36.8.0/22
70.36.8.0/22
70.36.8.0/22
70.36.8.0/22
70.36.8.0/22
70.36.8.0/22
70.36.8.0/22
70.36.8.0/22
...
```

(Les Unixiens noteront que c'est l'équivalent de `cut` : `cut -d, -f6 panne-spm.csv`.) Et les colonnes 2-4 :

```
1388864007,A,195.66.224.32
1388864007,A,195.66.225.111
1388864007,A,195.66.236.32
1388864007,W,195.66.224.32
1388864007,A,195.66.225.76
1388864007,W,195.66.236.32
1388864007,A,195.66.224.32
...
```

Le signe * est également utilisable.

Enfin, on peut indiquer une **cellule** particulière, identifiée par une ligne et une colonne. 4,2 désignera donc `195.66.224.32`. La syntaxe des intervalles marche aussi, donc 2-3,4-5 désignera quatre cellules :

```
1388864007,A
1388864007,W
```

Les identificateurs de fragments de ressources CSV peuvent aussi être composés de plusieurs sélections disjointes, mais, bon, cela devient un peu compliqué.

La syntaxe précise figure en section 3. Pour indiquer si le numéro désigne une ligne, une colonne, ou une cellule, on le préfixe avec `row=` (« rangée »), `col=` ou `cell=`. Comme montré dans l'exemple ci-dessus, les valeurs d'un intervalle sont séparées par un -, les lignes et colonnes d'une cellule par une ,. Si on a plusieurs sélections disjointes, elles sont séparées par un ;. Et, comme toujours avec les URI, l'identificateur de fragment est séparé du reste de l'URI par un #. Ainsi, l'URL pour récupérer la cellule 3,5 de mon fichier d'exemple sera `<http://www.bortzmeyer.org/files/panne-spm.csv#cell=3,5>`. Et, si vous avez un navigateur Web qui gère ce RFC, `<http://www.bortzmeyer.org/files/panne-spm.csv#row=10>` devrait vous pointer directement sur la dixième ligne.

La section 4 donne des détails sur la manière exacte dont un client devrait gérer ces fragments, notamment en cas d'erreurs. S'il y a une erreur de syntaxe (par exemple `<http://www.bortzmeyer.org/files/panne-spm.csv#line=10>`, avec `line` au lieu de `row`), l'identificateur doit être ignoré (tout se passera donc comme si on était simplement allé en `<http://www.bortzmeyer.org/files/panne-spm.csv>`). Même chose si les nombres pointent en dehors du fichier, parce que trop grands.

L'enregistrement du type `text/csv` a été mis à jour `<https://www.iana.org/assignments/media-types/text/csv>` pour inclure cette possibilité supplémentaire.