RFC 6657 : Update to MIME regarding Charset Parameter Handling in Textual Media Types

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 10 juillet 2012

Date de publication du RFC : Juillet 2012

https://www.bortzmeyer.org/6657.html

Lorsqu'un programme envoie des données en les étiquetant avec le type MIME `text/quelquechose`, quels sont le jeu de caractères et l'encodage utilisés par défaut ? ASCII ? UTF-8 ? Un autre ? Si vous l'ignorez, vous n'êtes pas le seul. Très peu de gens connaissaient les règles appliquées par défaut et ce RFC choisit donc de les simplifier en les supprimant. Pour le résumer : il n'y aura plus de "*charset*" par défaut pour les nouveaux types MIME `text/*`.

Le RFC 2046 [1], qui normalisait ces étiquettes MIME, disait pourtant (section 4.1.2) que le "*charset*" par défaut était US-ASCII (un point de vocabulaire au passage : le terme de "*charset*" est fréquent dans les normes IETF mais étymologiquement incorrect, puisqu'il désigne en fait à la fois un jeu de caractères et un encodage). Mais d'autres RFC avaient fait d'autre choix et l'exemple le plus connu était HTTP qui spécifiait que, lorsqu'un serveur envoie des données sans indiquer `charset`, c'est qu'elles sont en ISO 8859-1 (RFC 2616, section 3.7.1, le RFC 7231 a depuis changé cela), comme les en-têtes de la réponse HTTP. En pratique, le serveur HTTP prudent indiquera le "*charset*" (par exemple `Content-Type: text/html; charset=UTF-8`) plutôt que de compter sur cette règle mal connue.

Pire, certains types ont un mécanisme inclus dans les données pour trouver le "*charset*", notamment `text/html` et `text/xml`. Bref, il est difficile à un programme de compter sur la valeur par défaut. Ce RFC formalise en fait un usage déjà très répandu : être **explicite**.

La nouvelle règle est donc « pas de valeur par défaut générique pour tous les `text/*`, chaque sous-type de `text/*` peut définir sa propre valeur par défaut, y compris imposer qu'une valeur explicite soit indiquée. » Pour des raisons de compatibilité avec l'existant, les sous-types existants ne sont pas

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

modifiés, la nouvelle règle s'appliquera aux nouveaux sous-types. Notre RFC recommande en outre de choisir de ne **pas** avoir de valeur par défaut et, soit d'indiquer le "*charset*" dans le contenu (comme le fait XML avec son `<?xml version="1.0" encoding="iso-8859-1"?>`), soit de rendre obligatoire le paramètre `charset=` (indication explicite du "*charset*").

Les nouveaux enregistrements de sous-types MIME de `text/*` ne devraient donc plus indiquer de valeur par défaut pour le "*charset*". Si c'est absolument nécessaire, le RFC recommande que cette valeur soit UTF-8 (on est en 2012 et ASCII ne devrait plus être considéré comme le jeu de caractères par défaut).

On l'a dit, ce RFC ne change pas les sous-types existants et, par exemple, le "*charset*" par défaut du sous-type `text/plain` reste donc, lui, en ASCII.

Une alternative à l'étiquetage des données a toujours été d'essayer de deviner le "*charset*" par diverses heuristiques. Le RFC rappelle que cette méthode est dangereuse et déconseillée. Autre piège, le conflit entre l'information extérieure au contenu et celle incluse dans le contenu. Que faut-il faire avec un document XML servi en HTTP comme `text/html; charset=iso-8859-1` et commençant par `<?xml version="1.0" encoding="utf-8"?>` ? Le RFC recommande de se fier à l'information la plus interne (ici, la déclaration XML, qui dit que le document est en UTF-8), probablement la plus correcte.

Ce RFC 6657 est désormais celui cité à l'IANA `<https://www.iana.org/cgi-bin/mediatypes.pl>` pour les nouveaux enregistrements dans le registre des sous-types `<https://www.iana.org/assignments/media-types/text/>` de `text/`.