

RFC 4790 : Internet Application Protocol Collation Registry

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 15 mars 2007

Date de publication du RFC : Mars 2007

<https://www.bortzmeyer.org/4790.html>

Dans la désormais longue série des RFC qui parlent d'internationalisation de l'Internet, voici un RFC décrivant un cadre dans lequel doivent s'inscrire les protocoles qui doivent chercher un identificateur parmi d'autres, ou bien trier des identificateurs. Le problème était trivial en ASCII mais la prise en compte de jeux de caractères plus étendus change la donne.

De nombreux protocoles Internet doivent manipuler des identificateurs (c'est le cas, par exemple, de Sieve ou de NAI). Ils doivent tester si deux identificateurs sont identiques (par exemple, un serveur IMAP doit, lorsqu'il reçoit la commande LOGIN `stéphane monmotdepasse` déterminer si c'est le même utilisateur que le `STEPHANE` qu'il a dans sa base). Ou bien ils doivent trier ces identificateurs (tiens, je n'ai pas trouvé de bon exemple, ici).

Avec le seul jeu de caractères US-ASCII, spécifier de tels tests était trivial. Il suffisait de dire si la comparaison était sensible à la casse ou pas. Avec les jeux de caractères modernes, adaptés à un Internet mondial, comme Unicode, tout est bien différent. Est-ce que `stéphane` est égal à `stephane`? Et à `STÉPHANE`? Chaque protocole va devoir répondre à cette question et chacun, jusqu'à présent, le faisait dans son coin.

C'est ce que veut changer notre RFC : désormais, tout protocole qui veut faire de la comparaison ou du tri de chaînes de caractères devra spécifier l'algorithme utilisé, en utilisant un format unique et en enregistrant cet algorithme dans le tout nouveau registre <<https://www.iana.org/assignments/collation/collation-index.html>> IANA. Notre RFC le peuple déjà avec quelques algorithmes simples comme `i;ascii-casemap` qui effectue une comparaison insensible à la casse, en n'utilisant que les caractères ASCII. L'espoir est que la plupart des algorithmes seront réutilisés, réduisant ainsi la charge de travail pour les programmeurs qui mettent en œuvre les RFC.