# RFC 8746 : Concise Binary Object Representation (CBOR) Tags for Typed Arrays

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 29 février 2020

Date de publication du RFC : Février 2020

https://www.bortzmeyer.org/8746.html

---

Ce nouveau RFC étend le format de fichiers CBOR (normalisé dans le RFC 8949 [1]) pour représenter des tableaux de données numériques, et des tableaux multidimensionnels.

Le format CBOR est en effet extensible par des **étiquettes** ("*tags*") numériques qui indiquent le type de la donnée qui suit. Aux étiquettes définies dans la norme originale, le RFC 8949, ce nouveau RFC ajoute donc des étiquettes pour des types de tableaux plus avancés, en plus du type tableau de base de CBOR (qui a le type majeur 4, cf. RFC 8949, et dont les données ne sont pas forcément toutes de même type).

Le type de données « tableau de données numériques » est utile pour les calculs sur de grandes quantités de données numériques, et bénéficie de mises en œuvres adaptées puisque, opérant sur des données de même type, contrairement aux tableaux CBOR classiques, on peut optimiser la lecture des données. Pour comprendre l'utilité de ce type, on peut lire « "*TypedArray Objects*" <http://www.ecma-international.org/ecma-262/6.0/#sec-typedarray-objects> » (la spécification de ces tableaux dans la norme ECMA de JavaScript, langage dont CBOR reprend le modèle de données) et « "*JavaScript typed arrays*" <https://developer.mozilla.org/en-US/docs/Web/JavaScript/Typed_arrays> » (la mise en œuvre dans Firefox).

La section 2 spécifie ce type de tableaux dans CBOR. Un tableau typé ("*typed array*") est composé de données numériques de même type. La représentation des nombres (par exemple entiers ou flottants) est indiquée par l'étiquette. En effet, il n'y a pas de représentation canonique des nombres dans un tableau typé (contrairement aux types numériques de CBOR, types majeurs 0, 1 et 7) puisque le but de

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc8949.txt

ces tableaux est de permettre la lecture et l'écriture rapides de grandes quantités de données. En stockant les données sous diverses formes, on permet de se passer d'opérations de conversion.

Il n'y a pas moins de 24 étiquettes (désormais enregistrées dans le registre IANA des étiquettes CBOR <https://www.iana.org/assignments/cbor-tags/cbor-tags.xml>) pour représenter toutes les possibilités. (Ce nombre important, les étiquettes étant codées en général sur un seul octet, a suscité des discussions dans le groupe de travail, mais se justifie par le caractère très courant de ces tableaux numériques. Voir la section 4 du RFC.) Par exemple, l'étiquette 64 désigne un tableau typé d'octets (`uint8`), l'étiquette 70 un tableau typé d'entiers de 32 bits non signés et petit-boutiens (`uint32`), l'étiquette 82 un tableau typé de flottants IEEE 754 de 64 bits gros-boutiens, etc. (CBOR est normalement gros-boutien, comme tous les protocoles et formats Internet, cf. section 4 du RFC.) Les étiquettes ne sont pas attribuées arbitrairement, chaque nombre utilisé comme étiquette encode les différents choix possibles dans les bits qui le composent. Par exemple, le quatrième bit de l'étiquette indique si les nombres sont des entiers ou bien des flottants (cf. section 2.1 du RFC pour les détails).

Le tableau typé est ensuite représenté par une simple chaîne d'octets CBOR ("*byte string*", type majeur 2). Une mise en œuvre générique de CBOR peut ne pas connaitre ces nouvelles étiquettes, et considérera donc le tableau typé comme une bête suite d'octets.

La section 3 de notre RFC décrit ensuite les autres types de tableaux avancés. D'abord, les tableaux multidimensionnels (section 3.1). Ils sont représentés par un tableau qui contient deux tableaux unidimensionnels. Le premier indique les tailles des différentes dimensions du tableau multidimensionnel, le second contient les données. Deux étiquettes, 40 et 1040, sont réservées, pour différencier les tableaux en ligne d'abord ou en colonne d'abord. Par exemple, un tableau de deux lignes et trois colonnes, stocké en ligne d'abord, sera représenté par deux tableaux unidimensionnels, le premier comportant deux valeurs, 2 et 3, le second comportant les six valeurs, d'abord la première ligne, puis la seconde.

Les autres tableaux sont les tableaux homogènes (étiquette 41), en section 3.2. C'est le tableau unidimensionnel classique de CBOR, excepté que tous ses élements sont du même type, ce qui peut être pratique au décodage, pour les langages fortement typés. Mais attention : comme rappelé par la section 7 du RFC, consacrée à la sécurité, le décodeur doit être prudent avec des données inconnues, elles ont pu être produites par un programme malveillant ou bogué, et donc non conformes à la promesse d'homogénéité du tableau.

La section 5 de notre RFC donne les valeurs des nouvelles étiquettes dans le langage de schéma CDDL (RFC 8610).