RFC 6129 : The 'application/tei+xml' mediatype

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 12 février 2011

Date de publication du RFC : Février 2011

https://www.bortzmeyer.org/6129.html

Petit à petit, de nombreux formats XML obtiennent un type MIME, comme indiqué dans le RFC 7303 [1] et RFC 6839. Ces types permettent d'identifier un contenu de manière précise sur le réseau, et ont la forme `application/XXX+xml` où XXX indique le format précis. Ici, notre RFC concerne le type `application/tei+xml`, pour le format TEI, un format de documents très utilisé dans le monde des bibliothèques et dans les sciences humaines en général (le secteur de l'étude des documents anciens est un utilisateur très visible de TEI, mais ce n'est pas le seul).

En effet, TEI ("*Text Encoding and Interchange*") est un format utilisé depuis longtemps, par les musées, les bibliothèques, et des chercheurs, par exemple pour encoder un manuscrit qu'on a réussi à déchiffrer. Nous sommes donc ici à l'intersection du monde des "*geeks*" et de celui d'Umberto Eco. Si vous voulez voir des exemples de documents TEI, on en trouve un peu partout, par exemple une version d'≪ Alice au pays des merveilles ≫ issue du Projet Gutenberg et TEIsée <http://pgtei.pglaf.org/>. L'original est en <http://www.gnutenberg.de/pgtei/0.5/examples/alice/alice.tei> mais j'en ai fait une version légèrement modifiée (en ligne sur https://www.bortzmeyer.org/files/alice.tei) pour en retirer ce qui était spécifique de Gutenberg.

Comme tout bon fichier XML, les documents TEI ont un "*namespace*" XML, ici `http://www.tei-c.org/ns/1.0`. Plusieurs éléments XML peuvent être utilisés pour être la racine d'un document TEI, mais les plus communs sont `<TEI>` et `<teiCorpus>`. On les trouve souvent sur l'Internet avec des extensions de fichier `.tei` et `.teiCorpus`. (À noter qu'une recherche Google des fichiers se terminant par `.tei` ne donne pas le résultat attendu <http://www.google.com/search?q=filetype%3Atei> en raison du nombre de chinois nommés Tei...)

Notre RFC est très court. TEI lui-même est spécifié dans des documents du consortium TEI <http://www.tei-c.org/Vault/P5/1.8.0/doc/tei-p5-doc/en/html/> (le format est du Relax NG et disponible en format XML <http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng> ou en format compact <http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rnc>) et ce RFC ne fait que spécifier le type MIME et donner quelques conseils. Armé de ces schémas, on peut valider avec rnv <http://www.davidashen.net/rnv.html> :

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7303.txt

```
% rnv tei_all.rnc alice.tei
alice.tei
```

Par contre, la validation avec xmllint échoue en raison d'une boucle sans fin <https://bugzilla.gnome.org/show_bug.cgi?id=626779>.

Enfin, le consortium TEI fournit divers outils <http://www.tei-c.org/Tools/> permettant, par exemple, la traduction vers LaTeX ou HTML. À noter que le journal de ce consortium est sur revues.org <http://jtei.revues.org/>.

À noter que la traditionnelle section sur la sécurité couvre des points inhabituels dans un RFC. Par exemple, les documents encodés en TEI peuvent être tellement anciens qu'ils sont montés dans le domaine public mais parfois ils sont au contraire encore plombés par des règles d'appropriation intellectuelle. Le format TEI permet de spécifier ces règles, avec une grande granularité (élément `availability`). Ces spécifications peuvent être utilisées pour mettre en œuvre un système de menottes numériques ou, plus simplement, pour permettre l'application du droit moral de l'auteur (citer correctement les auteurs de chaque partie d'un document TEI).

D'autre part, TEI a déjà été utilisé pour encoder des documents ayant un rôle légal, et qui peuvent être encore applicables des siècles plus tard. Il peut donc être nécessaire de prendre des précautions pour vérifier l'authenticité et l'intégrité de ces documents (TEI lui-même ne le fait pas).

Enfin, les documents encodés peuvent être confidentiels et, là encore, cette confidentialité peut devoir être respectée sur de longues périodes (des dizaines d'années, si des personnes physiques sont mentionnées). Regardez donc bien les règles pour les archives de votre pays avant de distribuer des documents TEI.

Merci à Laurent Romary et Sebastian Rahtz pour leurs relectures et pour m'avoir envoyé un autre exemple de document TEI, un encodage du « Conte de Noël » de Dickens. Le fichier (en ligne sur `https://www.bortzmeyer.org/files/carol.zip`) contient le source TEI et on peut en tirer automatiquement une version EPUB, qui est dans le fichier (en ligne sur `https://www.bortzmeyer.org/files/3256.epub`).