RFC 8118 : The application/pdf Media Type

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 18 mars 2017

Date de publication du RFC : Mars 2017

https://www.bortzmeyer.org/8118.html

Le format PDF, largement utilisé sur l'Internet, n'a sans doute pas besoin d'être présenté ici. De toute façon, ce nouveau RFC ne prétend pas décrire PDF, juste le type de contenu `application/pdf`. Ce RFC remplace l'ancien RFC 3778 [1], notamment pour tenir compte du fait qu'officiellement, PDF n'est plus une spécification Adobe mais une norme ISO, 32000-1 :2008.

Donc, si vous envoyez des documents PDF via l'Internet, que ce soit par courrier ou par le Web, vous êtes censé les étiqueter avec le type MIME `application/pdf` (le type de premier niveau `applicaton/` indiquant que c'est un format binaire, non utilisable en dehors des applications spécialisées). Ce type a été enregistré à l'IANA <https://www.iana.org/assignments/media-types/application/pdf> (section 8 du RFC).

PDF avait été conçu pour le monde du papier (les commerciaux d'Adobe répétaient dans les années 90 que PDF permettait d'avoir « le même rendu partout » ce qui n'a pas de sens sur écran, où tous les écrans sont différents), ce qui se retrouve dans de nombreux concepts archaïques de PDF comme le découpage en pages. Un document PDF est un « rendu final », typiquement non modifiable, avec du texte utilisant différentes polices, des images…PDF permet également de représenter des liens hypertexte, une table des matières…On peut même inclure du JavaScript pour avoir des documents interactifs. PDF permet également le chiffrement et la signature, et a un mécanisme (en fait, plusieurs) pour placer des métadonnées, XMP. Bref, PDF est un format très complexe, ce qui explique les nombreuses failles de sécurité rencontrées par les programmes qui lisent du PDF.

La norme PDF est désormais déposée à l'ISO (ISO 32000-1) mais l'archaïque ISO ne distribue toujours pas librement ces documents. Si on veut apprendre PDF, il faut donc le télécharger sur le site d'Adobe <http://www.adobe.com/devnet/pdf/pdf_reference.html>.

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc3778.txt`

Pour les protocoles où il y a une notion d'identificateur de fragment (comme les URI, où cet identificateur figure après le croisillon), PDF permet d'indiquer une partie d'un document. Cela fera partie de la future norme ISO, mais c'était déjà dans l'ancien RFC 3778. Cet identificateur prend la forme d'un ou plusieurs couples clé=valeur, où la clé est, par exemple, `page=N` (pour aller à la page n° N), `comment=ID` (aller à l'endroit marqué par l'annotation ID), `zoom=S` (agrandir d'un facteur S), `search=MOT` (aller à la première occurrence de MOT)...(Je n'ai pas réussi à faire fonctionner ces identificateurs de fragments avec le lecteur PDF inclus dans Chrome. Quelqu'un connait un logiciel où ça marche ?)

PDF a également des sous-ensembles. La norme est riche, bien trop riche, et il est donc utile de la restreindre. Il y a eu plusieurs de ces sous-ensembles de PDF normalisés (voir sections 2 et 4 du RFC). Ainsi, PDF/A, sous-ensemble de PDF pour l'archivage à long terme (ISO 19005-3 :2012), limite les possibilités de PDF, pour augmenter la probabilité que le document soit toujours lisible dans 50 ou 100 ans. Par exemple, JavaScript y est interdit. PDF/X (ISO 15930-8 :2008), lui, vise le cas où on envoie un fichier à un imprimeur. Il restreint également les possibilités de PDF, pour accroitre les chances que l'impression donne exactement le résultat attendu. Enfin, PDF/UA (ISO 14289-1 :2014) vise l'accessibilité, en insistant sur une structuration sémantique (et non pas fondée sur l'apparence visuelle) du document. Tous ces sous-ensembles s'étiquettent avec le même type `application/pdf`. Ils ne sont pas mutuellement exclusifs : un document PDF peut être à la fois PDF/A et PDF/UA, par exemple.

Il existe d'innombrables mises en œuvre de PDF, sur toutes les plate-formes possible. Celle que j'utilise le plus sur Unix est Evince.

Un mot sur la sécurité (section 7 du RFC). On l'a dit, PDF est un format (trop) complexe, ce qui a des conséquences pour la sécurité. Comme l'impose la section 4.6 du RFC 6838, notre RFC inclut donc une analyse des risques. (Celle du RFC 3778 était trop limitée.) Notamment, PDF présente les risques suivants :

— Les scripts inclus, écrits en JavaScript,
— Le chargement de fichiers extérieurs, et les liens hypertexte vers l'extérieur,
— Les fichiers inclus, qui peuvent être absolument n'importe quoi, et qui viennent avec leurs propres dangers (sans compter le risque de leur exportation vers le système de fichiers local).

Et c'est sans compter sur les risques plus génériques, comme la complexité de l'analyseur. Il y a eu de nombreuses failles de sécurité dans les lecteurs PDF (au hasard, par exemple CVE-2011-3332 <`http://www.kb.cert.org/vuls/id/225833`> ou bien CVE-2013-3553 <`https://technet.microsoft.com/en-us/library/security/msvr13-007.aspx`>). La revue de sécurité à l'IETF avait d'ailleurs indiqué <`https://datatracker.ietf.org/doc/review-hardy-pdf-mime-02-secdir-lc-hallam-bak`> que les premières versions du futur RFC étaient encore trop légères sur ce point, et demandait un mécanisme pour mieux étiqueter les contenus « dangereux ».

Vous avez peut-être noté (lien « Version PDF de cette page » en bas) que tous les articles de ce blog ont une version PDF, produite via LaTeX (mais elle n'est pas toujours complète, notamment pour les caractères Unicode). Une autre solution pour obtenir des PDF de mes articles est d'imprimer dans un fichier, depuis le navigateur.

La section 2 du RFC rappelle l'histoire de PDF. La première version date de 1993. PDF a été un très grand succès et est largement utilisé aujourd'hui. Si on google `filetype:pdf`, on trouve « Environ 2 500 000 000 résultats » (valeur évidemment très approximative, le chiffre rond indiquant que Google n'a peut-être pas tout compté) . Si PDF a été créé et reste largement contrôlé par Adobe, il en existe une version ISO, la norme 32000-1, qui date de 2008 (pas de mise à jour depuis, bien qu'une révision soit attendue en 2017). ISO 32000-1 :2008 est identique à la version PDF 1.7 d'Adobe.

Normalement, les anciens lecteurs PDF doivent pouvoir lire les versions plus récentes, évidemment sans tenir compte des nouveautés (section 5 du RFC).

Quels sont les changements depuis l'ancienne version, celle du RFC 3778 ? La principale est que le "*change controller*", l'organisation qui gère la norme et peut donc demander des modifications au registre IANA est désormais l'ISO et non plus Adobe. Les autres changements sont :

— Une mise à jour de la partie historique,
— Une mise à jour de la partie sur les sous-ensembles de PDF, qui étaient moins nombreux autrefois,
— Une section sécurité bien plus détaillée.