RFC 7995 : PDF Format for RFCs

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 16 décembre 2016

Date de publication du RFC : Décembre 2016

https://www.bortzmeyer.org/7995.html

Parmi les nombreux formats de publication des RFC prévus suite à l'adoption du nouveau format (RFC 7990 [1]), ce RFC décrit l'utilisation de PDF. On pourra donc, en suivant ces règles, avoir une jolie version papier des RFC.

Actuellement, les RFC peuvent bien sûr être imprimés mais c'est un peu tristoune (cf. annexe A). Avec le nouveau cadre de gestion des RFC, le format canonique du RFC sera du XML (RFC 7991), à partir duquel seront produits automatiquement (par des outils qui ne sont pas encore développés) divers formats. HTML sera sans doute le principal pour une publication en ligne (RFC 7992), mais il y a des partisans de PDF, surtout pour l'impression sur le bon vieux papier. Ce RFC 7995 décrit donc l'utilisation de PDF comme format de sortie pour les RFC. À noter que ce format, créé et piloté par une entreprise privée n'est pas à proprement parler un « format Internet » et est sans doute moins connu des participants à l'IETF que ne l'est HTML.

La norme PDF est 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>.

Première question (section 2 de notre RFC), quelle version de PDF choisir ? PDF a évolué dans le temps, chaque version ajoutant de nouvelles fonctions. C'est aujourd'hui un format très complexe, difficile à mettre en œuvre complètement. C'est pour cela qu'il existe des profils de PDF, restreignant ce format pour des usages spécifiques. Ainsi, PDF/X est conçu pour l'échange de fichiers avec l'imprimeur. Pour les RFC, documents souvent normatifs, et à longue durée de vie, les exigences principales sont l'accessibilité et la stabilité. C'est ce que fournissent les profils PDF/UA (accessibilité) et PDF/A-3 (archivage à long terme).

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7990.txt

Pour les RFC, les profils choisis sont la version 1.7 de PDF, suffisamment ancienne pour être gérée par la plupart des logiciels, le profil PDF/A-3 pour limiter le nombre de fonctions à gérer, et PDF/UA pour l'accessibilité.

La section 3 de notre RFC détaille ensuite toutes les exigences particulières des RFC, pour le format de sortie PDF. Je ne les commente pas toutes ici, seulement celles qui me semblent importantes. Par exemple, la délicate question des polices. PDF permet d'inclure une police dans le document, ou bien de se référer simplement à une police par son nom. Dans ce dernier cas, si le logiciel qui lit le PDF ne trouve pas exactement cette police, il se rabat sur une police « proche », avec un résultat qui n'est pas forcément satisfaisant. De toute façon, PDF/A, le profil « archivage » impose d'inclure la police utilisée, pour éviter de dépendre de logiciels futurs. À noter que cela peut impliquer de payer : peu de polices sont gratuites pour l'auteur. L'annexe C.4 discute des polices acceptables. Il y a les gratuites, mais sans support Unicode complet, comme Source Sans Pro, Source Serif Pro ou Source Code Pro. Bien meilleure du point de vue de la couverture Unicode, mais payante, est Skolar <https://www.rosettatype.com/Skolar>. L'idéal sera peut-être la nouvelle police Noto <https://www.google.com/get/noto/>. Les RFC ayant maintenant le droit d'utiliser des caractères non-ASCII, mais avec des restrictions (cf. RFC 7997), il est possible que des caractères soient refusés par le "*RFC Editor*" uniquement parce qu'ils ne sont pas présents dans les polices utilisées.

Le choix des caractéristiques des polices (chasse fixe ou variable, empattement ou pas) devra suivre les mêmes règles que pour HTML et CSS, règles qui figurent dans le RFC 7993. À propos de HTML, notons d'ailleurs que notre RFC sur PDF demande que le PDF ressemble visuellement autant que possible au document HTML. Si vous écrivez un logiciel qui produit le PDF pour un RFC et que vous hésitez sur tel ou tel aspect graphique, consultez le RFC 7992 sur la sortie HTML.

Parmi les autres exigences pour la production de PDF, notre RFC demande qu'on évite les césures.

PDF permet de mettre des liens hypertextes. L'intérêt est faible puisque PDF est surtout utilisé pour le papier (si on regarde sur un écran PDF n'a aucun avantage par rapport au format bien plus ouvert qu'est HTML), mais le RFC prévoit quand même cette possibilité. Il y aura donc des liens, à la fois externes (vers des URL, y compris vers d'autres RFC et, dans ce cas, le RFC 7322 requiert que cela soit vers la page d'information officielle du "*RFC Editor*") et internes (une section du RFC référençant une autre). Les liens internes sont parfois produits automatiquement (par exemple depuis la table des matières vers les sections du texte).

Un problème délicat est celui de la façon dont le texte est stocké dans le document PDF. PDF permet en effet plusieurs façons de réaliser ce stockage. Elles donnent le même résultat visuel mais se comportent différemment pour des fonctions comme la recherche de texte. Ainsi, le mot « IETF » peut être stocké comme une image, comme quatre lettres positionnées indépendamment, ou comme un mot unique. Le stockage en image posera évidemment des problèmes aux outils comme pdftotext (mais ce n'est pas forcément grave pour les RFC, on a toujours le source XML) ou aux outils de synthèse vocale, nécessaires aux malvoyants. Pour la recherche de texte, la solution du mot unique est certainement meilleure, même si elle ne permet pas une typographie aussi subtile. Mais il y a aussi le placement des phrases. La phrase « "*The IETF supports the Internet*" » peut être stockée comme cinq mots différents stockés indépendamment (y compris dans un ordre différent de celui de la phrase) et positionnés ensuite, ou bien comme un objet unique.

Notre RFC recommande d'au moins garder les mots dans l'ordre du texte (PDF/UA l'impose).

Pour les images, si le source XML contenait à la fois de l'art ASCII et du SVG, notre RFC impose bien sûr qu'on utilise le SVG pour produire le PDF. Le texte alternatif aux images, indispensable pour

l'accessibilité <http://www.w3.org/TR/WCAG20-TECHS/PDF1.html>, doit être mis dans le PDF (dans la propriété /Alt).

Les métadonnées (noms des auteurs, date, etc) sont très utiles pour l'indexation et la recherche et doivent donc être mises dans le PDF. PDF a plusieurs façons d'embarquer des métadonnées, et la recommandation est d'utiliser XMP.

Parmi les zillions de fonctions de PDF, il peut agir en container d'autres fichiers (oui, comme tar ou AVI). Tous les logiciels PDF ne savent pas extraire ces fichiers embarqués dans le PDF mais c'est une fonction utile, au cas où. Le RFC recommande donc que des fichiers utiles soient ainsi embarqués : le source XML du RFC, les codes sources (dans les éléments <sourcecode> du RFC), les images (dont les sources SVG)...

Dernier point, les éventuelles signatures. Pour l'instant, il n'y a pas de mécanisme standard pour signer les RFC et en garantir l'authenticité mais, lorsque ce sera le cas, PDF permettra d'inclure une signature dans le fichier produit. (Cette fonction existe dans PDF depuis longtemps.)

Le RFC contient aussi trois annexes intéressantes. L'annexe A est un historique de la relation compliquée entre les RFC et PDF. Depuis longtemps, une version PostScript du RFC était acceptée par le "*RFC Editor*" et publiée, même si très peu d'auteurs en ont profité. Cela concernait surtout les RFC ayant des images ou des formules mathématiques comme les RFC 1119 ou RFC 1142. Le PDF produit par le "*RFC Editor*" pour tous les RFC (ou par) n'était, lui, qu'une simple « impression » du RFC en texte brut.

L'annexe B rappelle ce que doit faire un bon logiciel de production de contenu imprimé, avec découpage en pages. C'est plus dur que cela n'en a l'air, car il faut gérer les veuves et les orphelines, ne pas couper juste après le titre d'une section, ne pas couper les dessins en art ASCII, placer les tableaux intelligemment, etc.

Enfin, l'annexe C décrit une partie des outils disponibles pour le producteur de documents PDF. Si les logiciels de visualisation sont nombreux, il faut noter que tous n'ont pas la totalité des fonctions qu'utilise le format de sortie des RFC (par exemple les liens hypertexte). Du côté des imprimantes (le papier étant le but final de la plupart des documents PDF), certaines savent même gérer le PDF directement (dans les autres cas, ce sera au logiciel de visualisation de produire le format attendu par l'imprimante, souvent PostScript).

Et pour produire le PDF à partir du XML des RFC ? Une solution possible, puisqu'il existe une feuille de style XSLT (disponible en ligne <http://greenbytes.de/tech/webdav/rfc2629xslt/rfc2629xslt.html#output.pdf.fop>) est de produire du FO qui sera ensuite transformé en PDF, par exemple avec FOP (je n'ai personnellement eu que des expériences décevantes avec FO). Mais il existe plein de bibliothèques qui produisent du PDF, et qui pourraient être utilisées.

Comme notre RFC impose l'utilisation de profils de PDF comme PDF/A, un outil important est le logiciel de vérification qui s'assure que le résultat est bien conforme aux exigences de ce profil. Pour l'instant, il semble qu'il n'existe pas grand'chose dans ce domaine. Il faudra donc compter sur l'outil de production de PDF pour qu'il fasse un travail correct.