RFC 8153 : Digital Preservation Considerations for the RFC Series

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 25 avril 2017

Date de publication du RFC : Avril 2017

https://www.bortzmeyer.org/8153.html

La préservation, sur le long terme, des documents qui ne sont jamais passés par une forme papier, est un défi important de notre époque. Nous pouvons relire aujourd'hui toute la correspondance du ministère des affaires étrangères de Louix XV, pourrons-nous, dans un siècle ou deux, relire les documents numériques du vingtième siècle ? Pourrons-nous relire les RFC ? C'est le but de ce document que d'explorer les pistes permettant de donner aux RFC une meilleure chance de conservation.

Le "*RFC Editor*" (RFC 8728 [1]) est à la fois l'éditeur et l'archiviste des RFC. Les deux fonctions sont souvent contradictoires : l'éditeur voudrait utiliser les derniers gadgets pour publier des jolis trucs (multimédia, par exemple, ou contenus exécutables), l'archiviste est prudent et conservateur et voudrait des technologies simples. L'éditeur doit produire des documents clairs et lisibles. L'archiviste doit les conserver, et pour une durée potentiellement bien plus longue que les modes technologiques, durée qui peut atteindre des siècles (on est ravis, aujourd'hui, quand on retrouve les textes de lois d'un royaume depuis longtemps oublié, au fin fond de la Mésopotamie, même quand ces lois ont depuis longtemps cessé d'être applicables).

Notez que des organisations comme l'IETF produisent plein de documents (les discussions sur les listes de diffusion, par exemple), mais que ce RFC se focalise sur la préservation des RFC.

Historiquement, les RFC étaient en texte seul. Ce format avait des tas d'avantages <https://www.bortzmeyer.org/rfc-en-texte-brut.html>. Simple, et auto-documenté (la seule spécification nécessaire pour le comprendre était ASCII), il convenait bien à l'archivage. Régulièrement, des naïfs demandaient que le "*RFC Editor*" passe à un format « plus moderne », en général une mode passagère, oubliée quelques années après. Le texte seul a donc tenu très longtemps, et à juste titre.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc8728.txt

Mais la roue de l'histoire a fini par tourner et le RFC 6949 a pris acte du fait qu'on n'allait pas rester en texte seul éternellement. Le format officiel des RFC, décrit dans le RFC 7990 est désormais fondé sur XML, avec divers enrichissements comme le jeu de caractères Unicode (RFC 7997) ou les images en SVG (RFC 7996). Cela fait peser une pression plus forte sur l'archivage : si on est certain de pouvoir relire le texte brut en ASCII dans cent ans, qu'en est-il d'images SVG ? L'ancien système d'archivage des RFC ne va donc a priori pas suffire. (Le XML lui-même est relativement auto-documenté. Si on met des documents XML sous les yeux d'un programmeur qui n'en a jamais vu, il pourra probablement rétro-ingénierer l'essentiel. Ce n'est pas forcément le cas des vocabulaires qui utilisent XML, notamment le compliqué SVG.)

Le nouveau système d'archivage suivra le cadre conceptuel d'OAIS (norme ISO 14721 <https://www.iso.org/standard/57284.html>, disponible en ligne <http://public.ccsds.org/publications/archive/650x0m2.pdf>). Sur OAIS, on peut lire la bonne introduction d'Emmanuelle Bermes <https://figoblog.org/2005/11/29/1089/>. Il faut notamment distinguer deux tâches (section 1.1 de notre RFC) :

— Préservation des bits : archiver un fichier informatique et pouvoir le ressortir des dizaines d'années après, au bit près. Cela se fait, par exemple, en recopiant régulièrement le fichier sur de nouveaux supports physiques, et en vérifiant via une somme de contrôle ou une signature que rien n'a changé. Des classiques sauvegardes, vérifiées régulièrement, suffisent donc.
— Préservation du contenu : il ne suffit plus de stocker et de restituer les bits, il faut aussi présenter le contenu à l'utilisateur. Avoir une copie parfaite des bits d'un fichier WordPerfect de 1990 ne sert pas à grand'chose s'il n'existe plus aucun logiciel capable de lire le Wordperfect sur les machines et systèmes d'exploitation modernes. Assurer la préservation du contenu est plus complexe, et il existe plusieurs solutions, par exemple de garder une description du format (pour qu'on puisse toujours développer un outil de lecture), et/ou garder non seulement les fichiers mais aussi les outils de lecture, **et** tout l'environnement qui permet de les faire fonctionner.

Ceci dit, ce problème d'archivage à long terme des fichiers numériques n'est ni nouveau, ni spécifique aux RFC. Il a été largement étudié par de nombreuses organisations. On peut citer la BNF <http://www.bnf.fr/fr/professionnels/innov_num_preservation_numerique.html>, le projet LIFE en Grande-Bretagne <http://www.life.ac.uk/3/docs/Hole_pasig_v1.pdf>, ou l'étude du cycle de vie faite à la Bibliothèque du Congrès <http://blogs.loc.gov/digitalpreservation/2012/02/life-cycle-models-for-digital-stewardship/>. Des processus pour maintenir sur le long terme les fichiers, avec recopies régulières et nombreuses vérifications, existent.

Les RFC bénéficient depuis un certain temps d'un mécanisme similaire de préservation des bits : les métadonnées (indispensables pour retrouver un document) sont créées et enregistrées. Les fichiers sont recopiés d'un ordinateur à l'autre au fur et à mesure que les anciennes technologies de stockage deviennent dépassées. En outre, depuis 2010, tous les RFC sont imprimés sur du papier, pour avoir « ceinture et bretelles ». Les RFC plus anciens que 2010 subissent également parfois ce traitement, mais il existe des trous (RFC perdus, ou, tout simplement, problèmes avec le droit d'auteur, avant que cette question ne soit explicitement traitée, cf. RFC 8179).

Cette copie papier s'est avérée utile au moins une fois, quand 800 RFC ont dû être été re-saisis à la main, suite à une panne informatique (et une insuffisance des sauvegardes). Un petit détail amusant au passage : le "*RFC Editor*" à une époque acceptait des documents qui n'étaient pas des RFC, et qu'il faut aussi sauvegarder, voir l'histoire antique des RFC <http://www.rfc-editor.org/history.html>.

Il n'y a pas actuellement de sauvegarde de l'environnement logiciel utilisé pour lire les RFC, l'idée est que cela n'est pas nécessaire : on pourra toujours lire du texte brut en ASCII dans cent ans (la preuve est qu'on n'a pas de mal à relire le RFC 1, vieux de quarante-huit ans). Le processus de sauvegarde préserve les bits, et on considère que la préservation du contenu ne pose pas de problème, avec un

format aussi simple. (Par exemple, l'impression sur le papier ne garde pas les hyperliens mais ce n'est pas un problème puiqu'il n'y en a pas dans le format texte brut.)

Mais, puisque les RFC vont bientôt quitter ce format traditionnel et migrer vers un format plus riche, il faut reconsidérer la question. La section 2 de notre RFC explore en détail les conséquences de cette migration sur chaque étape du cycle de vie. Il faut désormais se pencher sur la préservation des contenus, pas seulement des bits.

Certaines caractéristiques du cycle de vie des RFC facilitent l'archivage. Ainsi, les RFC sont immuables. Même en cas d'erreur dans un RFC, il n'y a jamais de changement (au maximum, on publie un nouveau RFC, comme cela avait été fait pour le RFC 7396). Il n'y a donc pas à sauvegarder des versions successives. D'autres caractéristiques du cycle de vie des RFC compliquent l'archivage. Ainsi, ce n'est pas le "*RFC Editor*" qui décide d'approuver ou pas un RFC (RFC 5741). Il n'a donc pas le pouvoir de refuser un document selon ses critères à lui.

Le "*RFC Editor*" maintient une base de données (qui n'est pas directement accessible de l'extérieur) des RFC, avec évidemment les métadonnées associées (état, auteurs, date, DOI, liens vers les éventuels errata puisqu'on ne corrige jamais un RFC, etc). Les pages d'information sur les RFC sont automatiquement tirées de cette base (par exemple , pour ce RFC).

Les RFC citent, dans la bibliographie à la fin, des références dont certaines sont normatives (nécessaires pour comprendre le RFC, les autres étant juste « pour en savoir plus »). Idéalement, les documents ainsi référencés devraient également être archivés (s'ils ne sont pas eux-même des RFC) mais ce n'est pas le cas. Notre RFC suggère que l'utilisation de Perma.cc serait peut-être une bonne solution. C'est un mécanisme d'archivage des données extérieures, maintenu par groupe de bibliothèques juridiques de diverses universités. Pour un exemple, voici la sauvegarde Perma.cc <https://perma.cc/E7QG-TG98> (https://perma.cc/E7QG-TG98) de mon article sur le hackathon de l'IETF <https://www.bortzmeyer.org/hackathon-ietf.html>.

Dans un processus d'archivage, une étape importante est la normalisation, qui va supprimer les détails considérés comme non pertinents. Elle va permettre la préservation du contenu, en évitant de garder des variantes qui ne font que compliquer la tâche des logiciels. Par exemple, bien que XML permette d'utiliser le jeu de caractères de son choix (en l'indiquant dans la déclaration, tout au début), une bonne normalisation va tout passer en UTF-8, simplifiant la tâche du programmeur qui devra un jour, écrire ou maintenir un logiciel de lecture du XML lorsque ce format sera à moitié oublié.

Or, au cours de l'histoire des RFC, le "*RFC Editor*" a reçu plein de formats différents, y compris des RFC uniquement sur papier. Aujourd'hui, il y a au moins le format texte brut, et parfois d'autres <http://www.rfc-editor.org/old/rfc-online-2000.html>.

Maintenant qu'il existe un format canonique officiel (celui du RFC 7991), quelles solutions pour assurer la préservation du contenu ?

— "*Best effort*", préserver les bits et espérer (ou compter sur les émulateurs, ce qui se fait beaucoup dans le monde du jeu vidéo vintage),
— Préserver un format conçu pour l'archivage (PDF/A-3 étant un candidat évident - voir le RFC 7995, d'autant plus que le XML original peut être embarqué dans le document PDF),
— Préserver le XML et tous les outils, production, test, visualisation, etc. (Ce que les mathématiciens ou les programmeurs en langages fonctionnels appeleraient une fermeture.)

La première solution, celle qui est utilisée aujourd'hui, n'est plus réaliste depuis le passage au nouveau format. Elle doit être abandonnée. La deuxième solution sauvegarde l'information dans le document, mais pas le document lui-même (et c'est embêtant que le format archivé ne soit pas le format canonique, mais uniquement un des rendus). Et l'avenir de PDF/A-3 est incertain, on n'a pas encore essayé de le lire trente ans après, et les promesses du marketing doivent être considérées avec prudence (d'autant plus qu'il y a toujours peu d'outils PDF/A, par exemple aucun logiciel pour vérifier qu'un document PDF est bien conforme à ce profil restrictif). Pour la troisième solution, cela permettrait de refaire un rendu des RFC de temps en temps, adapté aux outils qui seront modernes à ce moment. Mais c'est aussi la solution la plus chère. Si on imagine un futur où XML, HTML et PDF sont des lointains souvenirs du passé, on imagine ce que cela serait d'avoir préservé un environnement d'exécution complet, les navigateurs, les bibliothèques dont ils dépendent, le système d'exploitation et même le matériel sur lequel il tourne !

Une solution plus légère serait de faire (par exemple tous les ans) un tour d'horizon des techniques existantes et de voir s'il est encore possible, et facile, de visualiser les RFC archivés. Si ce n'est pas le cas, on pourra alors se lancer dans la tâche de regénérer des versions lisibles.

Au passage, il existe déjà des logiciels qui peuvent faciliter certains de ces activités (le RFC cite le logiciel libre de gestion d'archives ArchiveMatica <https://www.archivematica.org/>).

Après cette analyse, la section 3 de notre RFC va aux recommandations : l'idée est de sauvegarder le format canonique (XML), un fichier PDF/A-3, le futur outil xml2rfc et au moins deux lecteurs PDF (qui doivent être capables d'extraire le XML embarqué). Les tâches immédiates sont donc :

— Produire les PDF/A-3 (à l'heure de la publication de ce RFC, l'outil n'est pas encore développé) avec le XML à l'intérieur, et l'archiver,
— Archiver le format canonique (texte seul pour les vieux RFC, XML pour les nouveaux),
— Archiver les versions majeures des outils, notamment xml2rfc,
— Archiver deux lecteurs PDF,
— Avoir des partenariats avec différentes institutions compétentes pour assurer la sauvegarde des bits (c'est déjà le cas avec la bibliothèque nationale suédoise <https://www.ietf.org/blog/2017/01/a-new-rfc-archive/>). Un guide d'évaluation de ce ces archives est ISO 16363.

La version papier, par contre, ne sera plus archivée.

Conclusion (section 4) : les RFC sont des documents importants, qui ont un intérêt pour les générations futures, et qui valent la peine qu'on fasse des efforts pour les préserver sur le long terme.