RFC 8752 : Report from the IAB Workshop on Exploring Synergy between Content Aggregation and the Publisher Ecosystem (ESCAPE)

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 1 mai 2020

Date de publication du RFC : Mars 2020

https://www.bortzmeyer.org/8752.html

Ce RFC est le compte-rendu d'un atelier de l'IAB qui s'est tenu en juillet 2019 au sujet du "*Web Packaging*", une proposition technique permettant de regrouper un ensemble de pages Web en un seul fichier, pouvant être distribué par des moyens non-Web, tout en étant authentifié. "*Web Packaging*" (ou "*WebPackage*") est un concentré de pas mal d'enjeux technico-politiques actuels.

Pour en savoir plus sur ce projet, vous pouvez consulter l'actuel dépôt du projet <https://github.com/WICG/webpackage>, géré par le WICG <https://wicg.io/>. (Il y avait un projet W3C à un moment mais qui a été abandonné <https://w3ctag.github.io/packaging-on-the-web/>). Vous pouvez commencer par ce document <https://github.com/WICG/webpackage/blob/master/explainer.md>. (Il y a aussi un brouillon à l'IETF <https://datatracker.ietf.org/doc/draft-yasskin-wpack-use-cases/>, et un nouveau groupe de travail, wpack <https://datatracker.ietf.org/wg/wpack/>.) Le projet avait été lancé à l'origine <https://blog.chromium.org/2018/05/the-state-of-web-at-google-io-2018.html> par Google.

La proposition a suscité pas mal de discussions, voire de contestations. N'est-ce pas encore un plan diabolique de Google pour entuber les webmestres ? L'IAB a donc organisé un atelier <https://www.iab.org/activities/workshops/escape-workshop/>, joliment nommé ESCAPE ("*Exploring Synergy between Content Aggregation and the Publisher Ecosystem*") au sujet du "*Web Packaging*". Cela permettait notamment de faire venir des gens qui sont plutôt du côté « création de contenus » (entreprises de presse, par exemple), et qui viennent rarement au W3C et jamais à l'IETF. Cet atelier s'est tenu à Herndon en juillet 2019. Il n'avait pas pour but de prendre des décisions, juste de discuter. Vous pouvez trouver les documents soumis par les participants sur la page de l'atelier <https://www.iab.org/activities/workshops/escape-workshop/>. Je vous recommande la lecture de ces soumissions.

Le principe de base de "*Web Packaging*" est de séparer l'écriture du contenu et sa distribution, tout en permettant de valider l'origine du contenu (l'annexe B du RFC décrit "*Web Packaging*" plus en détail). Ainsi, un des scénarios d'usage (section 2 du RFC) serait que le "*crawler*" de Google ramasse un paquetage de pages et de ressources sur un site Web, l'indexe, et puisse ensuite servir directement ce paquetage de pages et autres ressources au client qui a utilisé le moteur de recherche, sans renvoyer au site original. Et tout cela avec des garanties d'origine et d'authenticité, et en faisant afficher par le navigateur dans sa barre d'adresses l'URL original. Un autre usage possible serait la distribution de sites Web censurés, par des techniques pair-à-pair, tout en ayant des garanties sur l'origine (sur ce point particulier, voir aussi la section 3.3 du RFC). Notez que ces techniques font que le site original ne connait pas les téléchargements, ce qui peut être vu comme une bonne chose (vie privée) ou une mauvaise (statistiques pour le marketing). Et puis les inquiétudes vis-à-vis de "*Web Packaging*" ne viennent pas uniquement des problèmes pour avoir des statistiques. Des éditeurs ont dit lors de l'atelier qu'il étaient tout simplement inquiets des « copies incontrôlées ». En outre, l'argument de vie privée est à double tranchant : le site d'origine du contenu ne voit pas les téléchargements, mais un autre acteur, celui qui envoie le "*Web Packaging*" le voit.

Séparer création de contenu et distribution permet également la consultation hors-ligne, puisque un paquetage "*Web Packaging*" peut être auto-suffisant. Cela serait très pratique, par exemple pour Wikipédia. Actuellement, il existe des trucs plus ou moins pratiques (HTTrack…) mais le "*Web Packaging*" rendrait cette activité plus agréable. Notez que les participants à l'atelier ne se sont pas mis d'accord sur le caractère indispensable ou pas de la signature dans ce cas.

Potentiellement, un système comme "*Web Packaging*" pourrait également changer le monde du livre électronique : tout site Web pourrait être facilement « "*ebookisé*" ». Une amusante discussion à l'atelier a eu lieu sur l'intérêt des signatures. Comme souvent en cryptographie, les signatures ont une durée de validité limitée. Sept jours est proposé, par défaut, mais Moby Dick a été écrit il y a 61 000 jours.

Dernier scénario d'usage envisagé, l'archivage du Web <https://www.bortzmeyer.org/archive-du-web.html>. Ce n'est pas trivial, car il ne suffit pas de copier la page HTML, il faut garder toutes les ressources auxiliaires, d'où l'intérêt de "*Web Packaging*". Et la signature serait utile là aussi, pour vérifier que l'archive est sincère. (Voir aussi le RFC 7089 [1]).

La section 3 du RFC discute ensuite la difficile question de la relation entre les producteurs de contenu et les intermédiaires comme Google. Par exemple, si un producteur de contenu sous-traite la distribution du contenu à un CDN, il doit lui faire confiance pour ne pas modifier le contenu. "*Web Packaging*", avec son système de signature, résoudrait le problème. D'un autre côté, ça fait encore un format de plus dans lequel il faut distribuer le contenu, un coût pas forcément négligeable, et qui frappera de manière disproportionnée les petits producteurs, ou les moins pointus techniquement. Certains participants en ont profité pour râler contre AMP.

Comme toute nouvelle technique, "*Web Packaging*" pourrait mener à des déplacements de pouvoir dans l'écosystème du Web. Mais il est très difficile de prévoir ces effets (cf. RFC 5218). Est-ce que "*Web Packaging*" va favoriser les producteurs de contenu, les intermédiaires, les utilisateurs ? La section 4 du RFC explore la question. Par exemple, un des risques est la consolidation du pouvoir des gros intermédiaires. Si Facebook peut directement servir des paquetages Web, sans passer par le site original, les performances seront bien meilleures pour les paquetages déjà chargés par Facebook, qui gagnera donc en pouvoir. D'un autre côté, "*Web Packaging*" pourrait mener au résultat inverse : l'authentification des paquetages rendrait la confiance en l'intermédiaire inutile. (Personnellement, j'apprécie dans l'idée

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7089.txt

de "*Web Packaging*" que cela pourrait encourager le pair-à-pair, dont le RFC ne parle quasiment pas, en supprimant l'inquiétude quant à l'authenticité du contenu.)

La section 4 couvre d'autres questions soulevées par le concept de "*Web Packaging*". Par exemple, il ne permet pas facilement d'adapter le contenu à l'utilisateur puisque, au moment de fabriquer le paquetage, on ne sait pas qui le lira. Une solution possible serait, pour un même site Web, de produire plusieurs paquetages et de laisser l'utilisateur choisir, mais elle complexifie encore le travail des producteurs. (Personnellement, je pense que beaucoup de ces adaptations sont mauvaises, par exemple l'adaptation au navigateur Web dans l'espoir de contrôler plus étroitement l'apparence, et cela ne me chagrine donc pas trop si elles seraient plus difficiles.)

Et la sécurité ? Un mécanisme de distribution par paquetages Web signés envoyés par divers moyens serait un changement profond du mécanisme de sécurité du Web. Actuellement, ce mécanisme repose essentiellement sur TLS, via HTTPS (RFC 2818). Mais c'est très insuffisant ; TLS ne protège que le **canal**, pas les **données**. Si un site Web a des miroirs, HTTPS ne va pas protéger contre des miroirs malveillants ou piratés. Et, comme noté plus haut, si un contenu Web est archivé, et distribué, par exemple, par Internet Archive, comment s'assurer de son authenticité ? L'absence d'un mécanisme d'authentification des données ("*Object-based security*") est une des plus grosses faiblesses du Web (malgré des essais anciens mais jamais déployés comme celui du RFC 2660), et "*Web Packaging*", qui sépare la validation du contenu de sa distribution, pourrait contribuer à traiter le problème. Ceci dit, l'expérience de la sécurité sur l'Internet montre aussi que tout nouveau système amène de nouvelles vulnérabilités et il faudra donc être prudent.

Et la vie privée ? De toute façon, quand on récupère un contenu, qu'il soit sous forme de paquetages ou de pages Web classiques, on donne au serveur des informations (et HTTP est terriblement bavard, il n'y a pas que l'adresse IP qui est transmise). Il semble qu'au moins, "*Web Packaging*" n'aggrave pas les nombreux problèmes du Web. (Personnellement, je pense qu'il pourrait même les limiter mais l'analyse exacte est compliquée. À l'heure actuelle, si vous suivez un lien depuis Facebook, le site Web d'origine et Facebook sont au courant. Avec "*Web Packaging*" seul Facebook le saurait. Est-ce un progrès ?)

La technologie AMP, très controversée a souvent été mentionnée pendant l'atelier. Elle n'a pas de rapport direct avec "*Web Packaging*" mais elle sort de la même société, et est souvent présentée comme faisant partie d'un même groupe de technologies modernes. La section 5 du RFC discute donc des problèmes spécifiques à AMP.

Je n'ai pas regardé les outils existants pour faire du "*Web Packaging*" donc ce site n'est pas encore sous forme de paquetage.