RFC 6778 : IETF Email List Archiving, Web-based Browsing and Search Tool Requirements

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 31 octobre 2012. Dernière mise à jour le 31 janvier 2014

Date de publication du RFC : Octobre 2012

https://www.bortzmeyer.org/6778.html

Ce RFC est le cahier des charges de l'outil d'accès aux archives des innombrables listes de diffusion de l'IETF. Le gros du travail de cette organisation repose sur ces listes de diffusion, dont la plupart sont publiques, archivées et accessibles via le Web. Cette masse d'information est un outil formidable pour comprendre les décisions de l'IETF et les choix techniques effectués. Mais son volume rend l'accès à l'information souvent difficile. L'IETF se vante de sa transparence (tout le travail est fait en public) mais avoir toutes les discussions accessibles ne suffit pas, si l'information est trop riche pour être analysée. D'où l'idée de développer une ensemble d'outils permettant d'accéder plus facilement à ce qu'on cherche. C'est désormais accessible en <https://mailarchive.ietf.org/>.

Avant, la recherche était particulièrement difficile si une discussion s'étendait sur une longue période et surtout si elle était répartie sur plusieurs listes. Imaginons un auteur d'un RFC, un président de groupe de travail, ou un simple participant, qui veut retrouver toutes les discussions qui ont concerné un "*Internet-draft*" donné. Il va trouver des messages dans la liste du groupe de travail mais aussi dans celles d'une ou plusieurs directions thématiques, et peut-être aussi dans la liste générale de l'IETF. Certains ont chez eux des copies de toutes ces listes (copies parfois incomplètes) pour utiliser des outils locaux. (Au fait, personnellement, je me sers surtout de grepmail <http://grepmail.sourceforge.net/> et je cherche encore <https://www.bortzmeyer.org/namazu.html> un outil qui indexerait les données.)

En l'absence de copie locale, notre participant IETF courageux va compter sur la force brute, ou bien va essayer un moteur de recherche généraliste (mais il est ennuyeux pour l'IETF de devoir compter sur un outil externe pour son propre travail).

Ce n'est pas qu'il n'existait aucun outil à l'IETF. Il y a des solutions proposées sur le site de l'IETF <http://www.ietf.org/> (l'option "*Email Archives Quick Search*"). Mais elles ne couvrent pas tous les besoins, loin de là.

Quels sont ces besoins ? La section 2 en donne la liste. D'abord, il faut une interface Web. Elle doit permettre de naviguer dans un fil de discussion donné, ou suivant la date. (Les outils d'archivage classique comme MHonArc <http://www.mhonarc.org/> permettent déjà tous cela.) Les fils doivent pouvoir être construits en suivant les en-têtes prévus à cet effet (References: et In-Reply-To:) mais aussi en se fiant au sujet pour le cas où les dits en-têtes auraient été massacrés en route, ce qui arrive.

Surtout, il faut une fonction de recherche. Elle doit fonctionner sur une liste, ou un ensemble de listes, ou toutes les listes. Elle doit permettre d'utiliser ces critères :

— Nom ou adresse de l'expéditeur,
— Intervalle de dates,
— Chaîne de caractères dans le sujet ou dans le corps du message,
— Chaîne de caractères dans un en-tête quelconque. Le plus important est bien sûr le Message-ID: lorsqu'on veut donner une référence stable d'un message (« ton idée a déjà été proposée dans <5082D93B.6000308@cnam.fr> »). Le but est de pouvoir faire circuler une référence comme http://datatracker.ietf.org/mlarchive/msg?id=5082D93B.6000308@cnam.fr.

Et de les combiner avec les opérateurs booléens classiques (AND, OR et NOT).

Il n'y a pas que la recherche dans la vie, on voudrait aussi évidemment avoir des URI stables pour chaque message (stables <http://www.w3.org/Provider/Style/URI.html> et beaux <https://www.bortzmeyer.org/beaux-urls.html>). Et cet URL devrait être mis dans le champ Archived-At: du message avant sa distribution (cf. RFC 5064 [1]).

À propos d'URI, comme indiqué dans l'exemple de recherche via un Message-ID: montré plus haut, notre RFC demande que les recherches soient représentables par un URI, qui puisse être partagé (et qui soit donc stable). Le RFC ne donne pas d'exemple mais cela aurait pu être quelque chose comme http://datatracker.ietf.org/mlarchive/search?subject=foobar&startdate=2002-05-01&endda (dans la version publiée, c'est en fait quelque chose comme https://mailarchive.ietf.org/arch/search/?q list=precis%2Cxmpp). Naturellement, « stabilité » fait référence à l'URI, pas au résultat, qui peut varier dans le temps (par exemple si des nouveaux messages arrivent).

La plupart des archives à l'IETF sont publiques et doivent être accessibles anonymement mais quelques unes nécessitent une autorisation. Dans ce cas, le cahier des charges demande que le système d'authentification utilisé soit celui de datatracker.ietf.org.

Difficulté supplémentaire, toutes les listes IETF ne sont pas gérées au même endroit et par le même logiciel. Les listes hébergées à l'IETF le sont avec mailman mais il existe aussi des listes hébergées ailleurs et notre cahier des charges demandent qu'elles puissent être intégrées dans ce système de recherche (par exemple en abonnant l'adresse du programme d'archivage à la liste).

Il y a aussi des cas où des archives de listes de diffusion deviennent accessibles alors qu'elles ne l'étaient pas avant et doivent être incorporées dans le corpus géré par le nouveau système. Le RFC demande qu'au moins le format d'entrée mbox (RFC 4155) soit accepté. Et qu'il serait souhaitable d'accepter aussi le maildir.

En sens inverse, le système doit permettre d'exporter les messages (un système ouvert ne l'est pas si on ne peut pas en sortir l'information pour la garder et l'étudier chez soi ; la fonction d'exportation

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc5064.txt

n'a pas qu'un intérêt technique, elle est aussi un moyen de garantir la pérénnité de l'information publique). On doit pouvoir exporter une liste entière, une liste pour un intervalle de temps donné et, dans le meilleur des cas, le résultat d'une recherche (une vue, pour employer le langage des SGBD). Là encore, l'exportation en mbox est impérative et celle en maildir souhaitée.

Voilà, c'est tout, l'appel d'offres formel de l'IAOC était en ligne <http://iaoc.ietf.org/documents/Mail-Archive-Tool-RFP-2012-06-12-03.pdf>. Plus rigolo pour le programmeur de cet ambitieux projet : la section 3 lui rappelle qu'il devrait se préparer à l'arrivée du courrier électronique internationalisé <https://www.bortzmeyer.org/courrier-entierement-internationalise.html>... Le produit final a été livré le 29 janvier 2014 et est disponible en <https://mailarchive.ietf.org/>.