Inventer un meilleur système de nommage : pas si facile

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 19 juin 2011. Dernière mise à jour le 22 juin 2011

https://www.bortzmeyer.org/no-free-lunch.html

On voit apparaître en ce moment des tas de projets de faire un système de nommage meilleur que l'actuel système utilisé dans l'Internet, c'est-à-dire la combinaison de noms de domaines arborescents (comme munzer.bortzmeyer.org) et du DNS pour la résolution de ces noms en autres données, comme les adresses IP. Comme je connais deux ou trois choses sur le DNS, on me demande parfois mon avis sur ces projets. Il est souvent difficile de les analyser car ils se donnent rarement la peine d'expliciter leur cahier des charges : quel est le but exact ? Quel est le problème qu'ils essaient de résoudre ? Ce n'est jamais dit. À la place, ces projets se contentent en général de dire que le DNS est mauvais, sans préciser en quoi. Il y a une bonne raison pour cette absence fréquente de cahier des charges précis : on ne peut pas tout avoir à la fois. Le système actuel optimise certains critères, un autre système pourrait choisir d'en optimiser d'autres, mais aucun système ne pourrait être parfait et tout optimiser à la fois. Comme on n'a pas de succès médiatique en proposant un système qui est simplement meilleur sur certains points et moins bon sur d'autres, la plupart des projets choisissent de passer discrètement sous silence leur cahier des charges.

Impossible de faire une liste complète de tous ces projets, et il en apparait de nouveaux régulièrement. Un des plus beaux exemples d'esbrouffe médiatique était le projet DNS-P2P <https://www.bortzmeyer.org/dns-p2p.html> qui n'a produit en tout et pour but qu'un seul message de 140 caractères <https://twitter.com/#!/brokep/status/8779363872935936> et plus rien ensuite, mais qui a réussi à générer énormément de buzz. Parmi les projets plus récents, et peut-être plus sérieux, on trouve INS <http://changaco.net/ins/> ou Namecoin <http://forum.bitcoin.org/?topic=6017.0>.

Ces projets suivent en général le même cheminement : constatation de problèmes bien réels avec le système actuel (par exemple la censure par les autorités <http://www.justice.gov/iso/opa/ag/speeches/2010/ag-speech-101129.html> ou bien l'arrachage d'un nom de domaine à sa titulaire par les requins de l'appropriation intellectuelle), annonce spectaculaire d'un nouveau système rempli d'adjectifs (acentré, pair-à-pair, libre, décentralisé), compréhension, dès qu'on commence à réflechir cinq minutes, que le problème est plus compliqué qu'il n'en avait l'air sur Facebook, abandon. Pour essayer de limiter ce gâchis, j'ai fait cet article pour expliquer quelles sont les propriétés souhaitables d'un système de nommage et pourquoi il est difficile, voir impossible, de les concilier.

Voici une liste, que j'espère exhaustive, des propriétés qu'on souhaiterait avoir pour notre beau système de nommage :

— **Identificateurs parlants**. Tout le monde préfère `www.rue89.com` à `BE25 EAD6 1B1D CFE9 B9C2 0CD1 4136 4797 97D6 D246`. Pour certains, « identificateurs parlants » peut aussi amener à souhaiter des identificateurs structurés, permettant l'analyse (par exemple, `rue89` = domaine enregistré, `com` = TLD) et de reconnaître le fait qu'il y a un rapport entre `foo.example.com` et `bar.example.com`.
— **Identificateurs uniques au niveau mondial**. On n'a certainement pas envie de changer ses signets ou ses cartes de visite lorsqu'on passe de France en Corée. De même, si on fait de la publicité pour `fr.wikipedia.org`, on n'a certainement pas envie de dire « sauf si vous êtes chez le FAI Untel, auquel cas c'est `fr.wp.encyclo` ou si vous utilisez Namecoin, auquel cas c'est `fr/wikipedia` ». On veut que le même nom marche partout et à coup sûr (propriété que ne fournissent pas les moteurs de recherche.).
— **Identificateurs stables**. La disparition d'un URL est une des plaies du Web. On veut évidemment qu'un identificateur, donné comme référence dans un livre ou un article scientifique, soit toujours valable dix ans après.
— **Identificateurs sûrs**. Le terme est un peu flou. Disons qu'on voudrait que les mécanismes d'avitaillement et de résolution des identificateurs ne puissent pas être subvertis trop facilement par un méchant. (Comme peut l'être le DNS avec la faille Kaminsky <`https://www.bortzmeyer.org/comment-fonctionne-la-faille-kaminsky.html`> ou comme le sont les noms de domaine dans les pays où il n'y a pas de sécurité juridique du titulaire <`http://www.domainesinfo.fr/chronique/232/cedric-manara-proteger-ses-noms-de-domaine-quels-moyens-juridiq`
`php`>.)
— **Identificateurs résolvables**. Dans la plupart des cas, on ne s'intéresse pas à l'identificateur pour lui-même, on veut l'utiliser pour obtenir une autre information (une adresse IP, par exemple, pour pouvoir s'y connecter). Il faut donc un mécanisme de résolution, pas juste d'avitaillement (d'enregistrement). Ce point est délicat parce que, d'une certaine façon, **tout** type d'identificateur est résolvable. Il suffit de tout mettre dans une DHT, par exemple (en oubliant les problèmes de sécurité, cruciaux avec les DHT). Ou, au contraire de l'approche pair-à-pair de la DHT, on peut passer par un serveur Web qui fait des recherches dans une base centrale et envoie un résultat. Donc, quand je dis « identificateur résolvable », je veux dire, « de manière raisonnable » (oui, c'est très flou mais c'est clair pour moi qu'un serveur Web centralisé n'est pas une solution raisonnable).
— **Identificateurs enregistrables facilement, pas cher et sans possibilité de refus arbitraire**. Idéalement, on voudrait un système d'enregistrement « pair-à-pair » c'est-à-dire où il n'existe pas d'autorité jouant un rôle particulier. L'expérience prouve en effet que ces autorités tendent toujours à abuser de leur pouvoir.

Or, et c'est le point important, **on ne peut pas avoir toutes ces propriétés à la fois**. Par exemple, si on veut des identificateurs parlants comme `milka.fr` pour une personne prénommée Milka, on va susciter des jalousies et on court les risques de se le faire prendre par une grosse société ayant beaucoup d'avocats. Ces identificateurs ne seront pas stables. Autre problème de stabilité : si un identificateur est parlant, il risque d'y avoir des pressions pour le modifier, si le mot acquiert un autre sens, ou si on change d'avis (un URL comme `http://example.org/monblog/jean-michel-michu-est-un-clown` posera un problème de stabilité si vous souhaitez adoucir le ton plus tard...) Des identificateurs numériques arbitraires comme `1f8efda3-df57-4fd4-b755-8808a874dd38` ne suscitent pas de convoitises, ne risquent pas de devoir être modifiés, mais ne sont plus parlants... De même, pour avoir des noms enregistrables en pair-à-pair complet, la seule méthode réaliste semble être de les tirer au hasard dans un espace de grande taille (pour éviter tout risque de collisions), ce qui les rendra très peu parlants.

Regardons quelques exemples de familles d'identificateurs et voyons leurs propriétés :

— Les noms de domaine sont uniques au niveau mondial, parlants, relativement stables <`https://www.bortzmeyer.org/pourquoi-le-dns.html`>, mais pas assez en raison des convoitises qu'ils suscitent et de l'absence de sécurité juridique pour les titulaires. Grâce au DNS (RFC 1034 [1]), ils sont facilement résolvables et, grâce à DNSSEC, cette résolution peut être assez sûre.

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc1034.txt`

Aussi bien pour l'unicité que pour la sécurisation avec DNSSEC, ils sont enregistrés via une autorité, le registre, dont le contrôle fait régulièrement l'objet de conflits.

— Les URI (RFC 3986) sont de tellement de sortes différentes qu'on ne peut pas formuler d'opinion générale sur eux. Pour les URL, une de leurs composantes essentielles est un nom de domaine et ils héritent donc des propriétés des noms de domaine.
— Les clés PGP du RFC 9580 (ou, d'une manière générale, tous les identificateurs qui sont des clés cryptographiques publiques comme les identificateurs HIP <https://www.bortzmeyer.org/hip-resume.html> ou bien les clés SSH du RFC 4251), ne sont pas parlants du tout (suite de chiffres très longue), sont quasi-uniques (la probabilité de collision est infime), bien que créés sans registre (tirage au sort dans un espace de très grande taille), et fournissent une grande sécurité : prouver qu'on est bien le titulaire d'un identificateur est facile, en signant avec la clé privée, et personne ne peut vous prendre votre identificateur, même pas l'État. Ils ne sont pas résolvables directement, mais peut-être demain les mettra-t-on tous dans une DHT qui permettra de trouver facilement les données correspondants (c'est encore un sujet de recherche). Ils sont très stables, sauf percée de la cryptanalyse.
— Les UUID du RFC 9562 sont également quasi-uniques, générés sans registre, et peu parlants. Mais ils n'ont pas la sécurité des clés cryptographiques.
— Les ISBN sont stables, peu parlants, nécessitent un registre, et ne sont pas résolvables à l'heure actuelle. Quelqu'un connait-il un service sur le Web qui prend un ISBN et donne les métadonnées du livre comme son titre et le nom de l'auteur ? Ce service doit évidemment fonctionner avec des livres qui ne sont plus en vente. Worldcat <http://xisbn.worldcat.org/xisbnadmin/doc/> ne convient pas, il ne renvoie que la langue et la date de publication sur tous les ISBN que j'ai testé. Quant à OpenLibrary <http://openlibrary.org/dev/docs>, il ne semble connaître aucun livre en français.
— Ils ne sont pas généralement considérés comme de vrais identificateurs mais il faut quand même dire un mot des mots-clés utilisés dans une recherche, par exemple via Google. En effet, beaucoup de gens les utilisent comme identificateurs et, au lieu de dire « Allez en http://www.sarkozy.fr/ » disent « tapez Sarkozy dans Google ». Cette méthode est très mauvaise <https://www.bortzmeyer.org/identificateur-vs-moteur-de-recherche.html> car ces « identificateurs » sont parlants, résolvables (via le moteur de recherche) mais n'ont aucune stabilité. Aujourd'hui, c'est tel site qui répond à telle recherche, demain ce sera un autre.
— Les adresses IP sont très peu stables (la plupart du temps, elles dépendent du FAI), peu parlantes, nécessitent un registre (à l'exception des ULA du RFC 4193) mais sont très facilement résolvables via les tables de routage distribuées par des protocoles comme BGP.
— Les identificateurs EUI-64 (adresses Ethernet, adresses MAC, etc) sont peu parlants, attribués par un registre (deux niveaux de registre, l'IEEE et l'entreprise), uniques mais non résolvables. Et ils n'offrent aucune sécurité (sur Linux, vous pouvez changer votre adresses MAC avec `ifconfig`.)
— Les DOI ne sont pas parlants, dépendent d'un registre, n'ont pas de stabilité particulière (la stabilité ne dépend pas que de la technique mais aussi de pratiques sociales : des identificateurs comme les DOI qui dépendent entièrement d'une société privée ont un avenir très incertain) et ne sont pas résolvables (au début de la bulle DOI, il était question de les résoudre avec le protocole Handle du RFC 3650 mais ce projet n'a jamais débouché et tous les résolveurs de DOI aujourd'hui dépendent de l'URL d'un serveur centralisé, donc du DNS).
— Les "*tags*" du RFC 4151 sont assez parlants (au choix de leur auteur), uniques, et extrêmement stables. Celui de cet article est tag:bortzmeyer.org,2006-02:Blog/no-free-lunch. Comme ils incluent une date (février 2006 pour les "*tags*" de ce blog, reflétant le début de leur utilisation), l'absence de stabilité des noms de domaine sous-jacents n'est pas un problème. Le registre du .com ou bien le gouvernement du registre (celui des États-Unis) peut me prendre mon nom de domaine <http://www.guardian.co.uk/technology/2011/jul/03/us-anti-piracy-extradition-p en .com mais le "*tag*" formé avec celui-ci restera à moi. Par contre, les "*tags*" n'offrent aucune résolvabilité. En pratique, ils ne sont utilisés que pour fournir des identificateurs uniques et stables dans la syndication, pour que le lecteur de flux de syndication puisse savoir s'il a déjà vu un article.

Comme le montre cette liste, il n'existe pas d'identificateur idéal, qui aurait toutes les propriétés souhaitables (cf. RFC 1737 pour un exemple de cahier des charges pour des identificateurs idéaux). Le verrons-nous apparaître dans le futur, grâce aux progrès de la recherche fondamentale ? Peut-être mais je ne crois

pas, le problème est trop fondamental. Bien que cela n'ait pas été démontré mathématiquement, je pense que faire un système qui ait toutes ces propriétés, c'est comme de violer le premier ou le second principe de la thermodynamique. Lorsque quelqu'un arrive avec une telle proposition, il existe une infime possibilité qu'il soit un génie qui ait découvert une nouvelle voie. Mais il est bien plus probable qu'il soit un escroc ou tout simplement un ignorant qui n'a pas fait la moindre recherche avant de concevoir son système.

Dans l'état actuel de l'art, il faut donc rejeter tout projet qui ne dit pas clairement quelles propriétés il veut. Si les auteurs du projet ne veulent pas lister explicitement les propriétés de leur système de nommage, c'est parce qu'ils ont peur de montrer que leur système n'est pas idéal et ne fait pas tout.

Ce problème de l'impossibilité de tout optimiser à la fois est souvent présenté sous le nom de triangle de Zooko (par exemple dans un excellent texte de Dan Kaminsky <http://dankaminsky.com/2011/01/05/djb-ccc/#zooko>). Mais je trouve que le triangle de Zooko oublie plusieurs propriétés importantes donc j'ai préféré faire ma liste de propriétés.

Merci à O. Marce pour ses remarques. Parmi les bonnes lectures sur le sujet, je recommande l'article d'Emmanuelle Bermès, « Des identifiants pérennes pour les ressources numériques ; L'expérience de la BnF <http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf> ». Il y a aussi mon exposé sur les identificateurs <https://www.bortzmeyer.org/identificateurs.html>. Enfin, j'ai fait également un exposé sur ce problème, dont les transparents sont en ligne <https://www.bortzmeyer.org/nommage-beurre.html>.