

Nouvelle version d'IDN

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 22 août 2010. Dernière mise à jour le 2 novembre 2010

<https://www.bortzmeyer.org/idnabis.html>

Une série de cinq RFC vient de sortir, représentant la nouvelle version de la norme IDN, permettant d'utiliser des noms de domaine en Unicode. Cette nouvelle version est officiellement nommée « IDNA 2008 » mais n'a pas respecté le calendrier original, qui était complètement irréaliste. Il vaut donc mieux l'appeler « IDNabis ».

IDNabis marque des changements importants dans les concepts sous-jacents (indépendance par rapport à la version d'Unicode, détermination de la liste des caractères autorisés selon un algorithme et non plus selon une table, suppression de l'étape de canonicalisation obligatoire, etc), mais les conséquences pratiques pour les utilisateurs seront faibles. L'écrasante majorité des noms de domaines légaux selon IDNA 1 le seront toujours avec IDNabis, leur encodage en Punycode (RFC 3492¹) reste le même (donc, [Caractère Unicode non montré²] [Caractère Unicode non montré] [Caractère Unicode non montré] [Caractère Unicode non montré] sera toujours représenté sur le câble par xn--pgbs0dh et maçonnerie par xn--maonnerie-r3a), un certain nombre de chaînes de caractères seront désormais autorisées mais elles concernaient surtout des écritures peu répandues. D'autres, qui étaient autorisés (mais rarement utilisés) sont désormais interdites.

La liste des RFC qui forment IDNabis comprend :

- RFC 5890, « *Internationalized Domain Names for Applications (IDNA) : Definitions and Document Framework* », donne les définitions des termes essentiels comme les nouveaux "U-label" (forme Unicode d'un nom légal) et "A-label" (forme Punycode d'un nom légal),
- RFC 5894, « *Internationalized Domain Names for Applications (IDNA) : Background, Explanation, and Rationale* », explique et justifie (fort mal, selon moi), le projet IDNabis et ses concepts; ce RFC n'est pas une norme, il n'est là que pour information,
- RFC 5891, « *Internationalized Domain Names in Applications (IDNA) : Protocol* », est le cœur de la nouvelle norme, il décrit le protocole utilisé,

1. Pour voir le RFC de numéro NNN, <https://www.ietf.org/rfc/rfcNNN.txt>, par exemple <https://www.ietf.org/rfc/rfc3492.txt>

2. Car trop difficile à faire afficher par L^AT_EX

- RFC 5892, « *"The Unicode code points and IDNA"* », spécifie l'algorithme utilisé pour déterminer si un caractère est légal en IDNAbis, illégale ou bien si sa légalité dépend du contexte,
- RFC 5893, « *"Right-to-left scripts for IDNA"* », expose les règles pour les noms de domaine dont une partie s'écrit de droite à gauche (par exemple en hébreu),
- RFC 3492, « *"Punycode : A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)"* », faisait partie de IDNA 1 mais est inchangé pour IDNAbis.
- RFC 5895, « *"Mapping Characters for Internationalized Domain Names in Applications (IDNA) 2008"* », RFC non normatif (car contesté) sur la canonicalisation des noms de domaines. (Sur ce sujet, on peut aussi consulter le TR46 du consortium Unicode <<http://www.unicode.org/reports/tr46/>>.)

IDNAbis est bien plus complexe que IDNA 1. Ce dernier ne comportait que trois RFC, le RFC 3490, décrivant le protocole, le RFC 3491 qui normalisait nameprep, l'algorithme de canonicalisation (cette étape a été abandonnée dans IDNAbis), et le RFC 3492, sur Punycode, le seul RFC survivant de IDNA 1.

Quels sont les changements par rapport à IDNA 1 ? La description la plus complète des changements figure dans l'annexe A du RFC 5891. Pour la résumer :

- Le protocole est désormais indépendant de la version d'Unicode : tout changement dans Unicode est automatiquement disponible.
- Les caractères de ponctuation et les symboles sont désormais presque tous exclus.
- Il n'y a plus d'étape de normalisation standard. Chaque application est désormais libre d'effectuer la correspondance entre ce qu'a tapé ou sélectionné l'utilisateur et l'IDN envoyé sur le réseau.
- Le modèle de sélection des caractères autorisés est passé de « entièrement manuel, caractère par caractère » à « essentiellement algorithmique, fondé sur les propriétés Unicode - avec un peu d'exceptions manuellement ajoutées ». C'est ce qui permet l'indépendance par rapport aux versions d'Unicode.

Mais IDNAbis reste largement compatible avec l'ancien IDN (même principe de fonctionnement, même Punycode, même préfixe xn--, beaucoup de règles communes). En pratique, les utilisateurs, et même les registres de noms verront peu de différences.

Quelles étaient les motivations pour créer un IDNAbis seulement quelques années après le premier ? Il y en avait plusieurs, pas toutes avouables. De fortes pressions de l'ICANN s'étaient exercées, notamment pour avoir un prétexte pour retarder l'introduction des IDN dans la racine (devenue effective en mai 2010). Il y avait aussi toute une campagne de FUD concernant un soi-disant <<https://www.bortzmeyer.org/idn-et-phishing.html>> rôle des IDN dans le hameçonnage. Très présente au début du projet, cette motivation, souvent répétée en des termes sensationnalistes (comme la répétition du terme ridicule de « caractères dangereux ») a été sérieusement édulcorée au fur et à mesure du travail du groupe idnabis <<http://tools.ietf.org/wg/idnabis>> de l'IETF (voir par exemple le compte-rendu de la réunion IETF de Philadelphie en mars 2008 <<http://www.ietf.org/proceedings/71/minutes/idn.txt>>). Aujourd'hui, il n'en reste plus gère de trace dans les RFC.

D'autres motivations étaient plus consensuelles, comme le souhait d'avoir un IDNA indépendant de la version d'Unicode. Par exemple, IDNA 1 était lié à Unicode 3.2 et les écritures enregistrées par le consortium Unicode **après** la sortie de la 3.2 (comme le tfinagh) étaient donc interdites d'IDN. Ce point est désormais réglé.

Tout cela ne signifie pas que le résultat final fasse l'unanimité et, pour un bon résumé des questions qu'IDNAbis laisse ouvertes, on peut consulter la FAQ du consortium Unicode <<http://unicode.org/faq/idn.html>>. Pour le point de vue des promoteurs d'IDNAbis, voir le RFC 5894.

Il ne semble pas exister encore beaucoup d'implémentations de IDNAbis mais ce n'est pas forcément dramatique : les différences pratiques entre les deux versions sont suffisamment faibles pour que, pour la plupart des caractères, utiliser une des nombreuses bibliothèques mettant en œuvre l'ancienne version soit suffisant. Sinon, la première mise en œuvre d'IDNAbis en logiciel libre semble être celle de Verisign

<<http://www.verisign.com/domain-name-services/current-registrars/internationalized-domain-index.html>> et la seconde la GNU libidn <<http://josefsson.org/libidn2/>> dans sa version 2.

Voici les transparents d'un exposé que j'ai fait sur IDNAbis :

- PDF, version adaptée à la visualisation sur l'écran (en ligne sur <https://www.bortzmeyer.org/files/idnabis-SHOW.pdf>),
- PDF, version adaptée à l'impression (en ligne sur <https://www.bortzmeyer.org/files/idnabis-PRINT.pdf>),
- Source en LaTeX/Beamer (en ligne sur <https://www.bortzmeyer.org/files/idnabis.tex>).