

# NIST tool helps Internet master top-level domains

**At the request of a worldwide Internet organization, a computer scientist at the National Institute of Standards and Technology developed an algorithm that may guide applicants in proposing new “top-level domains”—the last part of an Internet address, such as .com, that people type in navigating the Web.**

As new top-level domains are added to the familiar .com, .info and .net, the algorithm\* checks whether the newly proposed name is confusingly similar to existing ones by looking for visual likenesses in its appearance. Having visually distinct top-level domain names may help avoid confusion in navigating the ever-expanding Internet and combat fraud, by reducing the potential to create malicious look-alikes: .COM with a zero instead of .COM, for instance.

Later this year, the Internet Corporation for Assigned Names and Numbers (ICANN) plans to launch the process for proposing a new round of “generic” top-level domains (gTLDs), strings such as .net, .gov and .org meant to indicate organizations or interests. In preparing for newly proposed gTLDs, ICANN reached out to various algorithm developers, including NIST’s Paul E. Black, as among those engaged to “provide an open, objective, and predictable mechanism for assessing the degree of visual confusion” in gTLDs.

Black’s algorithm compares a proposed gTLD with other TLDs and generates a score based on their visual similarities. For example, the domain .COM scores an 88 percent visual similarity with the familiar .COM. The resulting scores may help indicate whether the newly proposed domain name looks too much like existing ones.

To make its assessments, the algorithm rates the degree of similarity between pairs of alpha-numeric characters. Some pairs, such as the numeral “1” and its dead-ringer, the lowercase letter “l,” are assigned the highest scores for visual similarity while other pairs, such as “h” and “n”, are given lower scores. The algorithm takes other considerations into account, for example how certain pairs of letters, like “c” and “l,” can join to look like a third letter (“d”), as in the case of “close” and “dose.”

Employing these scores and considerations, the algorithm computes the “cost” of transforming one string of characters into another, such as “opel” into “apple.” Lower cost means higher visual similarity. The algorithm then adjusts for the relative lengths of the two strings (different lengths increase their distinctiveness) and converts the final cost into a percent similarity.

ICANN is considering future enhancements to the algorithm, such as having it check for visual confusion between existing domains and future planned Internet top-level domain names in scripts such as Cyrillic.

Source: National Institute of Standards and Technology

*This document is subject to copyright. Apart from any fair dealing for the purpose of private study, research, no part may be reproduced without the written permission. The content is provided for information purposes only.*