

This Domain Name is Greek to Me: An Introduction to Internationalized Domain Names for Investigators

By *nwaters*
Created 2010-07-06 09:41
[Jonathan D. Abolins](#)

An Analogy...

Consider a telephone directory for a city such as New York. The peoples' names and addresses are linked to their respective telephone numbers. Up to now, the listings are all in Latin characters.¹ The publishers decide to allow people to have entries in their preferred languages. Now, Ms. Ivana Ivanova has a listing in Russian Cyrillic, Mr. Achmed Husseyin has a listing in Arabic script, and so on. While the underlying telephone system hasn't changed much, the way people can look up people and their telephone numbers has greatly changed. Some people will find it easier looking up people and businesses sharing the same language. Other people will be confused because they cannot decipher the new entries

Internationalized Domain Names (IDNs)

Something like the above analogy has been happening to the Internet. In recent years, the Internet Corporation for Assigned Names and Numbers (ICANN) has been establishing a system for "Internationalized Domain Names" (IDNs).² Unicode allows the IDNs to use a wide variety of character sets from the world's languages.

For several years, non-Latin domain names could be registered under Latin character top level domain names (TLDs) such as .com or .net. This May, four countries—Egypt, Saudi Arabia, the United Arab Emirates, and the Russia Federation—started registering domains under country code TLDs (ccTLDs) in their native scripts.³ So, now it is possible to have domain names that, other than the dot ("."), are totally in Arabic or Cyrillic characters. This is only the beginning. More countries will follow suit.

This will make the Internet accessible to billions of people whose native languages do not use Latin characters and who are not readily able to switch between their native script and Latin characters on their computer systems. They'll be able to access local Internet resources and services in their local language.

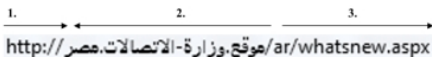
Many other people, including some investigators, might never run into an IDN. But for those encountering an IDN for the first time, the IDNs can present several challenges. The good news is that not all that much has changed at the "bits and bytes" and IP address level. But at other levels, there are some big changes with challenges. We'll look at these challenges and ways of dealing with them.

First Challenge: Recognizing non-Latin IDNs as Domain Names

We've tended to associate domain names with recognizable Latin character TLDs like .com or ccTLDs like .uk or .ru. Some IDNs will have a "traditional" TLD or ccTLD. So, `москва.com`, despite its Cyrillic text, can be recognized as a possible domain name (or, perhaps, a reference to a DOS/Win executable file). But the new non-Latin ccTLD IDNs, such as `سجل.مصر` or `правительство.рф`, aren't so recognizable as domain names.

If translators working with investigators are not aware of IDNs, they may miss important Internet clues. The translators might translate the parts of the IDNs as words without realizing that they might be domain names. For example, would we recognize the translation "register.Egypt" for "سجل.مصر" as a possible domain name?

By the way, locating the ccTLD is different if the IDN uses a character set for a language that's read right to left like Arabic or Hebrew. The ccTLD will be the leftmost label in the domain name. Figure 1 shows how a URL with an Arabic IDN can have the reading direction change several times.



1. → ← 2. → ← 3. → ←
`http://سجل.مصر/وزارة-الاتصالات.مصر/ar/whatsnew.aspx`

Figure 1: Example of a URL with an Arabic IDN and how the reading directions of the text can change as the character sets change.

Second Challenge: Will Our Network Tools Work with IDN?

Input of non-Latin IDNs via keyboard requires a special setup. Character map tools can allow us to "build" the Unicode text. But with both of these methods, we have to know the language's writing system to make sure we're entering the correct characters. If the IDN is already available in electronic form, we can "copy and paste" the text.

No matter how we input the non-Latin IDNs, many of the commonly used network tools were not designed to work with Unicode. Figure 2 shows what happens when we try a nslookup of a Russian IDN using its Unicode text.⁴

```
jabolins@pc7:~$ nslookup правительство.рф
Server:      192.168.1.1
Address:     192.168.1.1#53

** server can't find правительство.рф: NXDOMAIN

jabolins@pc7:~$
```

Figure 2. Example of an unsuccessful nslookup of a Russian IDN using its Unicode text.

Converting Unicode to ASCII Compatible Encoding

Fortunately, there are many tools for converting between Unicode and ASCII Compatible Encoding (ACE) or punycode.⁵ ACE is used by DNS and many other network tools to handle Unicode. The punycode conversion of a Unicode string will render the non-Latin character labels of a server name into ACE; Latin character labels stay the same. Two online Unicode→Punycode converters I've found quite handy are the ones at <http://www.idnstuff.com/> and <http://idnaconv.phymail.de/>. (Make sure the browser can handle Unicode UTF-8 encoding.)

Examples of Unicode IDNs and their punycodes:

- `москва.com` <-> `xn--80adxhks.com`
- `سجل.مصر` <-> `xn--rgbn6c.xn--wgbh1c`
- `правительство.рф` <-> `xn--80aealotwbjpid2k.xn--p1ai`

Once the Unicode IDN is converted, its punycode can be used with the network tools. Figure 3 shows an example of a successful nslookup on the punycode for the same IDN attempted in Figure 2.

```

jabolins@pc7:~$ nslookup xn--80aealotwbjpid2k.xn--plai
Server:      192.168.1.1
Address:     192.168.1.1#53

Non-authoritative answer:
Name:   xn--80aealotwbjpid2k.xn--plai
Address: 95.173.135.62
jabolins@pc7:~$

```

Figure 3. Example of a successful nslookup of the same Russian IDN using its punycode value.

Whois Lookups for IDNs

If the IDN has a Latin character TLD, commonly used whois tools should work with the IDN's punycode. But whois lookups for the new non-Latin character ccTLDs may have a harder time finding the right whois database. Command line whois searches don't readily work. AllWhois.com, which automatically locates the appropriate whois database server and returns the information for the queried domain, doesn't seem to work either.

One option is to use nslookup or other tools to find an IP address and, then, do a whois lookup of the IP address. Example: The IP address for سجل مصر (punycode: xn--rgrbn6c.xn--wgbh1c) is 81.21.97.106. A whois of that IP address points us to the African Internet Numbers Registry's whois database at whois.afinic.net.

Another option is using the Internet Assigned Numbers Authority's (IANA) Root Zone Database of ccTLDs at <http://www.iana.org/domains/root/db/> to find the right whois database. Find the ccTLD on that page and follow the link for the ccTLD to the IANA Delegation Record for the ccTLD. If there is a whois link for the ccTLD, it will be under the "Subdomain Information". If no whois link is available, we will have contact information for the ccTLD administrators.

Homograph Attacks: Real Risks or Hype?

Various people have been suggesting that IDN will open the doors wide open for homographic attacks using look-alike characters from different character sets.^{6,7} In 2005, Shmoo.com registered, as a proof-of-concept, a domain name (punycode: xn--pypal-4ve.com) that looked like paypal.com.⁸ But other people, like Michele Neylon on the CircleID site, point out that ICANN and others working on the Internet infrastructure are well aware of the risks and have been taking steps to reduce the opportunity for such attacks.⁹

So far, there have been no widespread homograph attacks reported. The Anti-Phishing Working Group's survey of phishing trends in 2009 found the last time they saw a case where a phishing campaign used a homograph IDN was in January 2009.¹⁰ The Group believes that rarity of homograph attacks in phishing may be due to the scammers having plenty of success without using look-alike domains and due to some browsers, such as Firefox, rendering URLs with non-Latin characters into punycode.

Closing

This article is a mere introduction to IDN issues which can affect investigations. There is very little IDN guidance yet available for investigators. I am hoping to help start to fill that gap and would love to hear from readers dealing with IDN or other multilingual/international issues. I will be periodically posting updated information online at <http://www.meydaonline.com/idn/>. You can contact me at Jon.Abolins@gmail.com.

Notes

1. For the purposes of this article, "Latin characters" will mean the basic alphanumeric characters found on a US or UK English keyboard or in basic ASCII.
2. ICANN. "Internationalized Domain Names." ICANN. 27 June 2010. <http://www.icann.org/en/topics/idn/>
3. Dam, Tina. "IDN ccTLDs – The First Four." 13 May 2010, ICANN. 27 June 2010. <http://blog.icann.org/2010/05/idn-cclds-%E2%80%93-the-first-four/>.
4. Also systems can vary in their ability to properly render Unicode. In my tests on Windows 7 and Ubuntu 10.4 Linux systems, the Linux bash shell environment rendered Unicode without a problem. But the Windows cmd environment could not render the Unicode properly. How to fix these problems is beyond the scope of this article. Fortunately, ASCII Compatible Encoding/Punycode makes much of this Unicode rendition problem moot.
5. Internet Engineering Task Force. "RFC 3492: Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)." March 2003, IETF. 27 June 2010. <http://tools.ietf.org/html/rfc3492>.
6. Gabrilovich, Evgeniy and Gontmakher, Alex. "The Homograph Attack." 28 July 2006, Technion – Israel Institute of Technology. 27 June 2010. <http://www.cs.technion.ac.il/~gabr/publications/papers/homograph.html>
7. April, Ben. "Can IDN Usage Open a Can of Unicode Worms?" 4 January 2010, TrendLabs Malware Blog. 27 June 2010. <http://blog.trendmicro.com/can-idn-use-open-a-can-of-unicode-worms/>.
8. Johanson, Eric. "The State of Homograph Attacks." 11 February 2005, The Shmoo Group. 27 June 2010. <http://www.shmoo.com/idn/homograph.txt>
9. Neylon, Michele. "IDN Scaremongering: Mashable and Times Online Screw Up." 4 January 2010, CircleID. 27 June 2010. http://www.circleid.com/posts/idn_scaremongering_mashable_and_times_online_screw_up/
10. Aaron, Greg and Rasmussen, Rod. "Global Phishing Survey: Trends and Domain Name Use 2H2009." May 2010, The Anti-Phishing Working Group. 27 June 2010. http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2009.pdf

Jonathan D. Abolins is an Analyst for the Meyda Online: Information Security & Networked World Studies project. He started this one-person research effort to examine the security implications of emerging technologies and their uses. He blogs at <http://jabolins.LiveJournal.com> and frequently posts on Twitter at <http://twitter.com/jabolins>.

Source URL: <http://www.dfinews.com/article/domain-name-greek-me-introduction-internationalized-domain-names-investigators>