Towards an end-to-end multilingual web
8th Multilingual Web Workshop, Riga 2015

Dennis Tan
Sr. Product Manager
IDN 101

A-Z 0-9 ‘-’

Internationalized Domain Names

Examples: nestlé.com 中文网.中国 日本.jp

IDNA2003

IDNA2008
Towards an end-to-end multilingual web
### Status of multilingual online content

Support for linguistic diversity by popular web services

<table>
<thead>
<tr>
<th>Name of service</th>
<th>Number of users</th>
<th>Languages supported</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>255 million active monthly</td>
<td>35+14</td>
<td>Network of 350 000 translators work through Twitter translation centre15</td>
</tr>
<tr>
<td>Google Translate</td>
<td></td>
<td>80</td>
<td>Statistical machine translation – based on patterns in large amounts of text, users are encourage users to contribute improved translations16</td>
</tr>
<tr>
<td>Facebook</td>
<td>1.3 billion active monthly</td>
<td>7317</td>
<td>Facebook also relies on a network of users who contribute translations16</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>21 million19</td>
<td>287</td>
<td>Number reflects languages for which official Wikipedias have been created20, 9 languages have over 1 million Wikipedia articles.</td>
</tr>
</tbody>
</table>

Internet Users in the World
Distribution by World Regions - 2014 Q2

Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 3,035,749,340 Internet users on June 30, 2014
Copyright © 2014, Miniwatts Marketing Group

Top 20 IDN registries: market share over time

Domain names: ASCII and IDN

IDNs are drivers of multilingualism

- IDNs enhance linguistic diversity in cyberspace
- Websites using IDNs are better predictors of language content than websites with an ASCII domain name

World report on Internationalised Domain Names 2014, Unesco, Verisign, Eurid
## Status of IDN support in popular browsers

<table>
<thead>
<tr>
<th></th>
<th>Google Chrome</th>
<th>Microsoft Internet Explorer 11</th>
<th>Mozilla Firefox</th>
<th>Opera</th>
<th>Safari</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can be forced to always show the IDN URL?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Decides whether to show the IDN URL as a whole or label by label?</td>
<td>Label by label</td>
<td>Label by label</td>
<td>Label by label</td>
<td>Based on the TLD only</td>
<td>Based on the script only</td>
</tr>
<tr>
<td>Contains a blacklist of characters that will prevent display of the IDN URL?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Has a configurable list that will allow display in specific languages?</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Has a whitelist of TLDs and will only show the IDN for TLDs in the list?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Has a whitelist of scripts and will only show the IDN for scripts in the list?</td>
<td>No</td>
<td>No</td>
<td>Yes, with algorithmic exceptions</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Allows for hybrid IDNs such as <a href="http://www.research.%D0%BE%D0%BD%D0%BB%D0%B0%D0%B9%D0%BD/">http://www.research.онлайн/</a>?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Source: World report on Internationalised Domain Names 2014, Unesco, Verisign, Eurid
Support for internationalized email addresses in top ten email clients across all platforms

福祥@邓福祥.com

<table>
<thead>
<tr>
<th>Market position</th>
<th>Client name</th>
<th>Share of market</th>
<th>Supports International Email Addresses?</th>
<th>Supports Sending of International Email?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Apple iPhone</td>
<td>26%</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Outlook</td>
<td>14%</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Apple iPad</td>
<td>12%</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Gmail</td>
<td>12%</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>Apple Mail</td>
<td>8%</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>Google Android</td>
<td>6%</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>Outlook.com</td>
<td>6%</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>Yahoo! Mail</td>
<td>5%</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>Windows Live Mail</td>
<td>2%</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>Windows Mail</td>
<td>2%</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Source: World report on Internationalised Domain Names 2014, Unesco, Verisign, Eurid
Universal Acceptance
Universal acceptance

1985
- .com
- .net
- .edu
- .org

2001
- .info
- .biz
- .coop
- .travel

2010
- .台湾
- .pf
- .泰国
- .한국

2013
- .shop
- .auto
- .doctor
- .点看

Software assumptions:
- TLD <= 3 characters
- ASCII only
- Hardcoded
A complete domain name must have one or more subdomain names and one top-level domain name.
A complete domain name must use dots (.) to separate domain names.
Domain names must use only alphanumeric characters and dashes (-).
Domain names must not begin or end with dashes (-).
Domain names must not have more than 63 characters.
The top-level domain name must be one of the predefined top-level domain names, like (com), (org), or (ca).

How to test domain name format?

In order to help your programming or testing tasks, FYIcenter.com has designed this online testing page for you to validate any given domain name using PHP regular expressions.

All you need to do is to enter a domain name in the form below and click the Start button.

```
Domain Name: rga.global
```

Test Result

The specified domain name has an **INVALID** format.

Other Online Testing Pages by FYIcenter.com

FYIcenter.com has prepared some simple but very interesting online testing pages that are useful for your programming and testing tasks:

- Test Credit Card Number Generator
- Credit Card Number Validator
- ...
- Show Me My Browser's Identification Information
- Show Me My IP Address and Host Name
- Domain Name Format Validator
- IP Address Name Format Validator
Universal acceptance

- Ubiquity

✓ Consistent
✓ Predictable
Universal acceptance

- Ubiquity
- Vicious circle

Image source: World Report on IDN Deployment 2013, Unesco, Verisign, Eurid
Universal acceptance

- Why is important?
  For the next generation of internet users
- Join the conversation
  icann.org/universalacceptance
- Be part of the solution

World Internet Penetration Rates by Geographic Regions - 2014 Q2

North America: 87.7%
Australia / Oceania: 72.9%
Europe: 70.5%
Latin America / Caribbean: 52.3%
Middle East: 48.3%
World, Avg.: 42.3%
Asia: 34.7%
Africa: 26.5%

Penetration Rates are based on a world population of 7,182,466,565 and 3,036,749,340 estimated internet users on June 30, 2014.
Copyright © 2014, Minitwatts Marketing Group
Confusables
What are confusable characters?

- Visually similar glyphs
  - Same script
    - Capital letter ‘O’, Latin number ‘0’
  - Different scripts
    - Cyrillic small letter ‘а’, Latin small letter ‘a’
    - Greek small letter ‘η’, Latin small letter ‘n’
    - Greek small letter ‘ρ’, Latin small letter ‘p’
The homograph issue

“... a way a malicious party may deceive computer users about what remote system they are communicating with, by exploiting the fact that many different characters look alike.”

Wikipedia.org
Preventing measures

Browser implementation techniques:

- Expose the ascii label (i.e. punycode) in all cases → bad user experience
  - Example:
    - user input = автомобилёк.com
    - Browser output = xn--80aaf1bkdcvh.com
- Show IDN if user language settings match IDN’s script, show ascii label otherwise → better user experience
  - Example:
    - User input = 汽车经销商.net
    - Browser output = 汽车经销商.net
Preventing measures

Registry techniques:

- IDNA2008 (RFC’s 5890, 5891, 5892, 5893)
  - Allowed code points in a domain name
- Unicode Technical Standard #39 (www.unicode.org/reports/tr39/)
  - Restriction-level detection
    - ASCII only
    - Single script
    - Highly restrictive (Latin/ Han, Hiragana, Katakana/Han, Bopomofo/ Han, Hangul)
    - Moderately restrictive (allow Latin with other, except Cyrillic and Greek)
    - Minimally restrictive (allow any mixture of scripts)
    - Unrestricted (allow any valid identifier, including symbols)
IDN conversion API and analysis tool

https://mctapi.verisign-grs.com

- A-label to U-label conversion and vice-versa
- Script analysis by label and code-point
- JSON format
- Available to public
[{
  "input":"ascii",
  "ascii":"xn--rsum-bsad.com",
  "unicode":null,
  "unicodeLabels":null,
  "codePoints":null,
  "scripts":null,
  "scriptCombination":null,
  "success":true
}]

[{
  "input":"unicode",
  "ascii":null,
  "unicode":null,
  "unicodeLabels":null,
  "codePoints":null,
  "scripts":null,
  "scriptCombination":null,
  "success":true
}]

[{
  "input":"unicode",
  "ascii":null,
  "unicode":null,
  "unicodeLabels":null,
  "codePoints":null,
  "scripts":null,
  "scriptCombination":null,
  "success":true
}]