L2/05-225

SECURITY ASPECTS IN IDN

Marcos Sanz
DENIC eG
sanz@denic.de

Abstract

- Prologue
- Parties
- UTR #36
- Definitions
- Recommendations
- Related problems
- Roadmap

Prologue

- In December 2002 RFC 3454 explicitly warns about the problems of "similar-looking characters" and suggests that "user applications can help disambiguate some similar-looking characters by showing the user when a string changes between scripts".
- In February 2005 xn--pypal-4ve.com is registered by The Shmoo Group.
- OMG, OMG, OMG

Interesting interested parties

- ICANN
  - Plans to update their IDN Committee's *Guidelines for the Implementation of IDNs*
  - http://www.icann.org/general/idn-guidelines-20jun03.htm
  - No activity judging
    - http://forum.icann.org/lists/idn-homograph/
    - http://forum.icann.org/lists/idn-discuss/
  - Workshop on IDN on the past 13th July
- ITU-T Study Group 17
  - Security, languages and telecommunication software
  - http://www.itu.int/ITU-T/studygroups/com17/index.asp
  - Meeting in October to discuss IDNs

Party! More parties...

- IETF, individual drafts:
  - "Suggested Practices for Registration of Internationalized Domain Names", draft-klensin-reg-guidelines-08.txt
    - Suggests applying JET Guidelines to alphabetic languages, sticking to one language tag per domain, variant tables and bundles.
  - "National and Local Characters for DNS Top Level Domain (TLD) Names", draft-klensin-idn-tld-05.txt
- IAB IDN Ad Hoc Committee
  - Initiated March 2005, haven't seen any output yet
- GAC, ALAC, NCUC...

Party! More parties...

- Unicode Consortium
  - Most prolific of all stakeholders
  - Undeniable expertise with Unicode Standard
  - Unicode Technical Report #36: "Unicode Security Considerations"

http://www.unicode.org/reports/tr36/

UTR #36

- It points out current problems with IDNA:
  - Too large a character repertoire
    - Symbols
    - Old fashioned characters
  - Not aligned with UAX #31: "Identifier and Pattern Syntax", http://www.unicode.org/reports/tr31/
    - No combining marks in the first position
  - Use of Unicode 3.2
    - Missing characters of language minorities
    - Normalization problems

Definitions on confusability

- **Visually confusable**: Two different strings whose appearance in common fonts in small sizes is sufficiently close to easily mistake.
- **Homographs**: Special kind of visually confusables. Two different strings that can always be represented by the same sequence of glyphs.

Visual spoofing is due to both, not only to the latter.

Definitions on confusability

- **Single script confusable**: Spoofing characters entirely within one script or using characters common across scripts (such as numbers).

a-b	ASCII
a‐b	U+0210 hyphen
dze	ASCII
dze	U+02A3 digraph
lOl	Expression of amusement
101	Binary 5

Definitions on confusability

- **Mixed script confusable**: Spoofing characters within more than one script and not a single script confusable.

paypal	ASCII
paypal	U+0430 Cyrillic
top	ASCII
top	U+03BF Greek

Definitions on confusability

- **Whole script confusable**: Mixed script confusables where each of the strings is entirely within one script.

caxap	Cyrillic
caxap	Latin
scope	Latin
scope	Cyrillic
BERT	Latin
□□□□	Cherokee

Other Bad Ideas

- **Bidirectional Spoofing**. IDNA and IRI specifications already require that:
  - Each label of a domain name must not mix RTL with LTR characters.
  - A label using RTL characters must start and end with RTL characters.

  But:

  http://سلام .دائم .com

  http:// دائم .a. سلام .com
- So better:
  - Avoid mixing RTL and LTR in a single domain name
  - Minimize the use of digits in host names and other IRI components containing RTL characters

Other Bad Ideas

- **Syntax Spoofing** examples directing us to bad.com

  http://example.com⁄x.bad.com
  (beware of U+2044 Fraction Slash)

  http://example.com?x.bad.com
  (beware of missing fonts as question marks)

  http://example.com----long-and-obscure-list-of-characters.bad.com
  (this one already on the wild)

Definition: Identifier Profile

- Identifiers: Special-purpose strings for identification
- UAX #31 permits definitions of *profiles* that add or remove characters to the specification
- **General Security Profile** excludes ca 60,000 characters
  - Not in modern use
  - Only used in specialized fields (liturgical, phonetical, mathematical...)
  - Ideographic characters not in the CJK core
    - 3 characters were explicitly allowed back because already in use by domain name registries.

Definition: Identifier Profile (II)

- **IDN Security Profile**, based on the general security profile. It provides a list of all and only those characters recommended for use in IDN:

  http://www.unicode.org/reports/tr36/data/idnchars.txt

  - Strict profile, defines characters on input/output
  - Lenient profile, more lenient on input than the strict profile

  It leaves 37,200 characters for use in IDN
  (not limited to Unicode 3.2)

Definition: Restriction Levels

1. **ASCII-Only**
2. **Highly Restrictive**
   - All characters from a single script except
     - Han + Hiragana + Katakana
     - Han + Bopomofo
     - Han + Hangul
   - No characters outside the Identifier Profile
3. **Moderately Restrictive**
   - Latin allowed with other scripts except
     - Cyrillic, Greek, Cherokee
4. **Minimally Restrictive**
   - Arbitrary mixture of scripts
5. **Unrestricted**
   - Allows characters outside the Identifier Profile

Definition on confusables

- Algorithms for confusable detection are defined
- Confusable data table in four flavours
  - Single-Script, Lowercase
  - Single-Script, Any-Case
  - Mixed-Script, Lowercase
  - Mixed-Script, Any-Case

http://www.unicode.org/reports/tr36/data/confusables.txt

Recommendations for ICANN

- Restricting domain names according to language is problematic:
  - Strings are sometimes language neutral
  - Languages are fluid
  - Foreign words
- "While the ICANN guidelines say 'top-level domain registries will [...] associate each registered IDN with one language or set of languages', that guidance is better interpreted as limiting to script rather than language".

Recommendations for users

- Use Good Software
- If registering domain names, care about the guidelines followed by the registry.
- Register confusables, if not automatically provided by the registry.
- Try to choose domain names that are less spoofable.

Recommendations for user agents

- Display the domain name in Nameprepped form
- If the domain name contains letters confusable with syntax characters, generate an alert.
- Let the user choose a Restriction Level and generate different kinds of alerts, if a domain name fails to satisfy it.
- Set default to Restriction Level 2
- Alert if the domain name is a whole-script or a mixed-script confusable.

PROBLEM 1: Core domains

- Highlighting the "core" domain to prevent syntax spoofs:

  http://example.com/x.**bad.com**

- **But:**
  - No formal definition of the concept
  - No explanation how to determine its position. Hardcoded lists?
  - There might be more than one "core"
  - It could be more dangerous to highlight the wrong core than not doing anything.

PROBLEM 2: Mixing scripts

- What's the problem with mixing scripts?
- There are lots of legitimate uses:
  - Ωmega, Teχ, Toys-Я-Us, HλLF-LIFE
  - IP□□□□, XML-документы
- Not mixing doesn't saves you from:
  - in-script spoofing
  - whole-script spoofing
- And remember, nothing will save you from *Conceptually Continuously Confused* (TM):
  - pay-pal.com
  - paypal-online.com
  - paypal24.com
  - ...

PROBLEM 3: Recommendation for registries

- "When a proposed domain name is confusable with an existing one, block it or avoid that another registrant registers it."
  - It's not current practice.
  - The determination of the registrant identity is not a trivial issue and one that domain name registries usually don't tackle with at all.
  - A name right usually also covers rights on graphical variants. Thus the domain name holder could, via appropriate existing dispute resolution mechanisms, always get those confusables, if need be.
  - The registry shouldn't try to compulsively satisfy registrants in a legally dubious/**risky** way.
    - "In a monopoly, discriminations are not allowed. If a registry is protecting a registrant from visual confusables, why not from conceptual confusables?"

Roadmap

- Cover other security areas, not directly related to IDNs: font spoofing, collation issues, private use characters..
- Move the Technical Report to a Technical Standard
  - Conformance to Unicode Standard does not imply conformance to any UTS
- Deliver input to ICANN for an update of their Guidelines for the Implementation of IDNs

Thank you
Any questions?