Guidelines on deployment support for domain names and email addresses containing Non-ASCII characters in a software

This document contains technical recommendations for developers of software designed to support or already supporting Non-ASCII domain names and/or email addresses.

The document seeks to offer developers the data they need for implementing Universal Acceptance (UA) principles into their software products taking into consideration the specific features of Non-ASCII domain names and email addresses.

The Universal Acceptance Steering Group (UASG) defines the following five Universal Acceptance criteria:

1. Accept – Accepting occurs whenever a domain name or an email address is received as a string of characters from a user interface, from a file, or from an API used by a software application.

2. Validate – Validation is intended to ensure that the entered information is valid.

3. Store – Whether information is properly stored and retrieved from a database (or file).

4. Process – Processing is intended to ensure that domain names and email addresses are properly used by an application to perform its activity.

5. Display – Whether domain names and email addresses are properly rendered within a user interface.

This document was drafted in keeping with these five UA criteria and sets forth recommendations as detailed below:

I. For supporting domain names containing Non-ASCII characters

1. Accepting domain names

1.1. Both GUI users and APIs must be able to transfer domain names to the software as both U-labels[1] (Unicode) and A-labels (Punycode).

1.2. If a user transfers the domain name through GUI as an A-label[2], turning it into a U-label using data preprocessing is recommended.

2. Validation

2.1. The following requirements have to be met when registering domain names containing Non-ASCII characters:

2.1.1. All internationalized domain names must follow IDNA 2008.

2.1.2. Mapping the domain name to lower case and carrying out NFC[3] normalization is recommended before validating the domain name.

2.1.3. Usually the list of the acceptable symbols in domain names is available in IDN Tables of the relevant TLD or, if absent, from the Terms and conditions of domain name registration in the relevant TLD.

2.1.4. We advise against using symbols other than those listed in clause 2.1.3 in a single domain name label.

We also advise taking precautions to prevent homoglyph attacks (a deception technique exploiting similarities of characters).

2.1.5. Recommended that domain names should begin and end with a digit or a letter; they shouldn't end with a hyphen.

2.1.6. Before checking the domain name length, it has to be turned from a U-label into an A-label.

2.1.7. A domain name can have the length of a second-level and lower-level domain name, from 1 to 63 octets. The length of an internationalized domain name in symbols must be measured in A-label.

2.1.8. A domain name's total length shall not exceed 255 characters in A-label.

2.2. For existing domain names, check its delegation with a DNS query.

Additional checks can be carried out in specific cases to verify whether a top-level domain exists. A regularly updated IANA TLD list[4] shall be used for this purpose.

3. Storage

3.1. A domain name must be stored in a database or in files in Unicode (UTF-8).

3.2. A domain name can be stored as an A-label in addition to an U-label, but labels have to be matched whenever one of the labels is changed.

4. Processing

4.1. All operations with domain names shall be carried out in Unicode (UTF-8).

4.2. When searching a data set that includes domain names, we advise ensuring that each label is presented both as an A-label and a U-label for better search performance.

5. Display

5.1. Domain names shall be displayed using graphical user interface in Unicode (UTF-8).

5.2. A domain name can be displayed as an A-label only when supplementing a domain name as a U-label.

5.3. When displaying domain name-related errors, highlight the domain name label related to the error.

II. For supporting email addresses containing Non-ASCII characters

1. Accepting email addresses

1.1. The domain in the email address shall be presented as per I.1 above.

1.2. The local part in an email address must be presented in Unicode (UTF-8) only.

2. Validation

2.1. The domain part[5] of an email address shall be validated as per I.2 above.

2.2. The local part[6] of an internationalized email address shall follow the EAI standard.

2.3. The following norms should be followed when creating (registering) a new email address:

2.3.1. Mapping the email address to lower case and carrying out NFC normalization is recommended before validating the email address. If necessary, NFKC[7] normalization can be used.

2.3.2. The local part in an email address can contain letters of alphabets, special symbols, and digits.

2.3.3. It is recommended that mixing letters of different alphabets should be prohibited in the local part of the email address, except the cases when the writing system of a nation or a territory uses several alphabets at the same time.

2.3.4. We advise to avoid using special symbols in the local part of an email address, except the dot (.), underscore (_) and hyphen (-). In some cases, usage of the plus (+) can be accepted.

2.3.5. We advise against starting and ending the local part of an email with any special symbols or having two special symbols in a row.

2.3.6. The length of the local part in an email address must be between 1 and 64 characters.

3. Storage

3.1. The domain part of an email address must be stored as per I.3 above.

3.2. The local part of an email address must be stored in Unicode (UTF-8).

4. Processing

4.1. All operations with email addresses should be carried out in Unicode (UTF-8).

4.2. When searching a data set that includes email addresses, we advise ensuring that each label in the domain part is presented both as an A-label and a U-label for better search performance.

5. Display

5.1. The domain part of an email address shall be displayed as per I.5 above.

5.2. The local part of an email address shall be displayed in Unicode (UTF-8).

5.3. When displaying errors related to email addresses, we advise highlighting the local part in an email address, if it contains an error, or the relevant domain name label, if the error concerns the domain part.

III. Sources

This document is based on the following materials:

- Introduction of Universal Acceptance,

https://uasg.tech/wp-content/uploads/documents/UASG007-en-digital.pdf

- Email Address Internationalization – Technical Perspective,

  https://uasg.tech/wp-content/uploads/documents/UASG019B-en-digital.pdf

- Universal Acceptance Readiness Framework,

  https://uasg.tech/wp-content/uploads/documents/UASG026-en-digital.pdf

- RFC 5321 - Simple Mail Transfer Protocol,

  https://tools.ietf.org/html/rfc5321

- RFC 5322 - Internet Message Format,

  https://tools.ietf.org/html/rfc5322

- RFC 1035 - Domain Names - Implementation and Specification,

  https://tools.ietf.org/html/rfc1035

- RFC 3492 - Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications,

  https://www.ietf.org/rfc/rfc3492.txt

- Internationalized Domain Names for Applications standard

https://www.ietf.org/rfc/rfc5890.txt

https://www.ietf.org/rfc/rfc5891.txt

https://www.ietf.org/rfc/rfc5892.txt

https://www.ietf.org/rfc/rfc5893.txt

https://www.ietf.org/rfc/rfc5894.txt

https://www.ietf.org/rfc/rfc5895.txt

- EAI standard for internationalized email

  https://tools.ietf.org/html/rfc6530

  https://tools.ietf.org/html/rfc6531

  https://tools.ietf.org/html/rfc6532

  https://tools.ietf.org/html/rfc6533

  https://tools.ietf.org/html/rfc6855

  https://tools.ietf.org/html/rfc6856

- Unicode Consortium, https://home.unicode.org/
- Universal Acceptance Steering Group (UASG), https://uasg.tech/
- Поддерживаю.РФ – a project to develop an ecosystem supporting domain names and email addresses in Cyrillic script (in Russian) https://поддерживаю.рф/
- Unicode normalization forms http://www.unicode.org/reports/tr15/
- Unicode Character Code Charts https://www.unicode.org/charts/
- IDN tables for TLDs https://www.iana.org/domains/idn-tables

[1] U-label is an internationalized domain name containing Unicode characters.

[2] A-labels are ASCII-compatible (ACE) internationalized domain name labels. A-labels always start with an ACE “xn--” prefix. An A-label can be turned into a U-label and vice-versa without losing any information.

[3] Unicode normalization forms - http://www.unicode.org/reports/tr15/

[4] http://data.iana.org/TLD/tlds-alpha-by-domain.txt

[5] Domain part of an email address is the domain name after the @sign.

[6] The local part of an email address is the email address before the @ sign.

[7] Unicode normalization forms - http://www.unicode.org/reports/tr15/