RFC Format BoF

Wednesday, 6 March 2014 @ 15:20-16:20

Agenda

- Background
- The RFC Format Design Team
- The I-Ds
- Non-ASCII in RFCs
- Next Steps
- Q&A

Background

The format announcement in May 2013 indicated several things:

- the canonical format we are exploring for RFCs is XML
- four publication formats will be created from that XML: HTML, EPUB, text and PDF
- non-ASCII characters would be allowed in a controlled fashion

http://www.rfc-editor.org/pipermail/rfc-interest/2013-May/005584.html

RFC Format Design Team

- An RFC format design team was put together during IETF 87 in Berlin to clear up the details implied by those statements

https://www.rfc-editor.org/rse/wiki/doku.php?id=design:design-team

Many thanks to Nevil Brownlee (ISE), Tony Hansen, Joe Hildebrand, Paul Hoffman, Ted Lemon, Julian Reschke, Adam Roach, Alice Russo, Robert Sparks (Tools Team liaison), and Dave Thaler for their participation

The I-Ds

The 'XML2RFC' version 2 Vocabulary - https://datatracker.ietf.org/doc/draft-reschke-xml2rfc/

The 'XML2RFC' version 3 Vocabulary - https://datatracker.ietf.org/doc/draft-hoffman-xml2rfc/

The Use of Non-ASCII Characters in RFCs - https://datatracker.ietf.org/doc/draft-flanagan-nonascii/

HyperText Markup Language Request For Comments Format - https://datatracker.ietf.org/doc/draft-hildebrand-html-rfc/

SVG Drawings for RFCs: SVG 1.2 RFC - https://datatracker.ietf.org/doc/draft-brownlee-svg-rfc/

PDF for an RFC Series Output Document Format–
In Progress

Non-ASCII in RFCs

Note that the tools used to post I-Ds and create RFCs are not ready to allow this new guidance

Requirements

- Searches of an index need to be able to find multiple ways of writing an author's name;
- People whose system do not have the fonts needed to display a particular RFC need to be able to read the non-canonical HTML, text, or PDF RFC correctly.

Note on language

- The language of the RFC Series is English. The use of non-ASCII characters is expected to be more the exception within the body of the text than the rule.

Person Names

Non-ASCII characters in names are allowed, but in this case an additional ASCII-only variant is required.

Header

CURRENT:

Network Working Group
Request for Comments: 2611
BCP: 33
Category: Best Current Practice

L. Daigle
Thinking Cat Enterprises
D. van Gulik
ISIS/CEO, JRC Ispra
R. Iannella
DSTC Pty Ltd
P. Faltstrom
Tele2/Swipnet
June 1999

PROPOSED/NEW

Network Working Group
Request for Comments: 2611
BCP: 33
Category: Best Current Practice

L. Daigle
Thinking Cat Enterprises
D. van Gulik
ISIS/CEO, JRC Ispra
R. Iannella
DSTC Pty Ltd
P. Fältström (P. Faltstrom)
Tele2/Swipnet
June 1999

Acknowledgements

CURRENT:

7. Acknowledgements

The following people contributed significant text to early versions of this draft: Patrik Faltstrom, William Chan, and Fred Baker.

PROPOSED/NEW:

7. Acknowledgements

The following people contributed significant text to early versions of this draft: **Patrik Fältström (Patrik Faltstrom), 陈智昌 (William Chan)**, and Fred Baker.

Body Text

Where required for implementation and understanding of content:

- Require identifying the codepoint (e.g. U+0394)
- Encourage using the actual character (e.g., Δ) so reader can more easily see what the character is, if they can render it (U+ form etc isn't human intelligible per se)
- Allow official character names like "Greek Capital Letter Delta”

Body Text – example

(color and boldface highlight examples – their use is not part of the proposal for non-ASCII text)

- CURRENT (draft-ietf-precis-framework) :
  However, the problem is made more serious by introducing the full range of Unicode code points into protocol strings. For example, the characters U+13DA U+13A2 U+13B5 U+13AC U+13A2 U +13AC U+13D2 from the Cherokee block look similar to the ASCII characters "STPETER" as they might look when presented using a "creative" font family.

- PROPOSED/NEW:
  However, the problem is made more serious by introducing the full range of Unicode code points into protocol strings. For example, the characters U+13DA U+13A2 U+13B5 U+13AC U+13A2 U +13AC U+13D2 **(STᎵETER)** from the Cherokee block look similar to the ASCII characters "STPETER" as they might look when presented using a "creative" font family.

- ALSO ACCEPTABLE:
  However, the problem is made more serious by introducing the full range of Unicode code points into protocol strings. For example, the characters **"STᎵETER" (U+13DA U+13A2 U+13B5 U+13AC U+13A2 U+13AC U+13D2)** from the Cherokee block look similar to the ASCII characters "STPETER" as they might look when presented using a "creative" font family.

Body Text – example

(color and boldface highlight examples – their use is not part of the proposal for non-ASCII text)

- Current (RFC 6912)
- The IAB recommendations do, however, leave some issues open that need to be addressed. It is not clear that all code points permitted under IDNA2008 that have a General_Category of Lo or Lm are appropriate for a zone such as the root zone. To take but one example, the code point U+02BC (MODIFIER LETTER APOSTROPHE) has a General_Category of Lm. In practically every rendering (and we are unaware of an exception), U+02BC is indistinguishable from U+2019 (RIGHT SINGLE QUOTATION MARK), which has a General_Category of Pf (Final_Punctuation). U+02BC will also be read by large numbers of people as being the same character as U+0027 (APOSTROPHE), which has a General_Category of Po (Other_Punctuation), and some computer systems may treat U+02BC as U+0027. U+02BC is PROTOCOL VALID (PVALID) under IDNA2008 (see the IDNA Code Points document [RFC5892]), whereas both other code points are DISALLOWED. So, to begin with, it is plain that not every code point with a General_Category of Ll, Lo, or Lm is consistent with the type of conservatism principle discussed in Section 4.1 below or the previous IAB recommendations.

- Proposed/New(RFC 6912)

The IAB recommendations do, however, leave some issues open that need to be addressed. It is not clear that all code points permitted under IDNA2008 that have a General_Category of Lo or Lm are appropriate for a zone such as the root zone. To take but one example, the code point **U+02BC (ʼ, MODIFIER LETTER APOSTROPHE)** has a General_Category of Lm. In practically every rendering (and we are unaware of an exception), U+02BC is indistinguishable from **U+2019 (’, RIGHT SINGLE QUOTATION MARK)**, which has a General_Category of Pf (Final_Punctuation). U+02BC will also be read by large numbers of people as being the same character as **U+0027 (', APOSTROPHE)**, which has a General_Category of Po (Other_Punctuation), and some computer systems may treat U+02BC as U+0027. U+02BC is PROTOCOL VALID (PVALID) under IDNA2008 (see the IDNA Code Points document [RFC5892]), whereas both other code points are DISALLOWED. So, to begin with, it is plain that not every code point with a General_Category of Ll, Lo, or Lm is consistent with the type of conservatism principle discussed in Section 4.1 below or the previous IAB recommendations.

The “use” versus “mention” case

- The mention, as opposed to the use, of non-ASCII characters requires Unicode identifiers, encourages the use of the non-ASCII characters, and allows the use Unicode character names.
- The distinction is between an occasion in which a word appears in a sentence in order to convey the meaning of that word (use), and an occasion in which a word appears in order to make some point about that word itself (mention).
- Spelling will follow the guidance in the Merriam-Webster dictionary. Other uses, including preference for spelling (like the _New Yorker_ stylesheet coöperate and so on) are not allowed.

Use:

CATEGORY	EXAMPLES
naïve	300

Mention:

CATEGORY	EXAMPLES
Latin	naïve (U+0063 U+0061 U+00EF U+0076 U+0065)

Non-ASCII Characters in Artworks

- In artwork, non-ASCII characters (such as box-drawing characters) cannot be used for creating structure, but they are allowed in examples of non-ASCII text.
- Additional guidance TBD

Bibliographic Text

The reference entry must be in English; whatever subfields are present MUST be available in ASCII. As long as good sense is used, they MAY also include non-ASCII characters at author discretion. This applies to both normative and informative references.

Examples of bibliographic entry

RFC 5832

Current

```
[GOST3410]    "Information technology.  Cryptographic data
               security.  Signature and verification processes of
               [electronic] digital signature.", GOST R 34.10-2001,
               Gosudarstvennyi Standard of Russian Federation,
               Government Committee of Russia for Standards, 2001.
               (In Russian)
```

Allowable addition to the above citation

"Информационная технология. Криптографическая защита информации. Процессы формирования и проверки электронной цифровой подписи ", GOST R 34.10-2001, Государственный стандарт Российской Федерации, 2001.

Examples of bibliographic entry

RFC 5992

Current

```
[Kert86]        Kertom, G., "Sami-Russian and Russian-Sami
                dictionary: textbook for primary school pupils",
                Leningrad: Prosveshchenie Leningrad Department,
                1986.  Published in Russian, no authoritative
                translation is known.
```

Allowable addition to the above citation

Kertom, G., "Сами-русский и русско-саамский словарь: Учебник для учащихся начальной школы", Ленинградский Департамент Просвящения, 1986.

Format Next Steps

- For all drafts:
  - Incorporate feedback received during IETF 89 (as appropriate)
  - Finish drafts and start the formal publication process (which includes more community review) before IETF 90
- Create a Statement of Work to write the specs and start the community review process during IETF 90
  - Specification(s) for tool(s) that create XML, HTML, PDF, and TXT documents, following the requirements and profiles defined in the drafts/RFCs produced by the RFC Format Design Team
- Start development of tools based on well-defined specs by IETF 91

Q&A

- When we run out of time, please take the discussion to the list:

rfc-interest@rfc-editor.org

<https://www.rfc-editor.org/mailman/listinfo/rfc-interest>