

WG-LTRU rechartering EITF Last Call

J-F C. Morfin

<http://bc47.com/wgltru-recharter-lc.pdf>

The IETF WG-LTRU has the responsibility of documenting the IETF "langtags" (BCP47) IETF Charters are reviewed every year. This review is subject to comments by the participants to the IETF community.

Proposed new description of the IETF Working Group on langtags

The primary language subtags allowed by the current IANA Language Subtag Registry are limited to ISO 639-1 and ISO 639-2 in the same way as they were limited by RFC 3066. This covers approximately 500 possible language subtags. ISO 639-3, scheduled for approval towards the end of 2006, will make available around 7000 more code elements for identifying languages.

The working group will prepare an update to the Language Subtag Registry procedures to allow the use of 3-letter code elements from ISO 639-3 as primary language subtags or extended language subtags as appropriate. The working group will also deliver means to update the current IANA Language Subtag Registry with the newly available subtags.

The working group will examine, and if necessary clarify or adjust, procedures and guidelines with respect to extended language subtags and variant subtags. Use cases include the identification of signed languages, transliterations, and transcriptions. The working group will also consider how the draft work on ISO 639-6 may relate to the future development of language tags.

The working group will clarify text where necessary. It may also make adjustments to the registration process and the form of the registry if this is deemed appropriate based on ongoing registration and operational experience. These adjustments and clarifications are not expected to delay the progress of the work.

Work on the drafts is planned to start before ISO 639-3 is fully approved and published. However, the WG will not finalize the drafts before ISO 639-3 is fully approved.

1. what is what ?

This Charter is ambiguous as it now calls for enabling "extended languages", but does not allude to what is a language, a primary language, an extended language, etc., distinctions which are ignored by ISO. The current debate at the WG-LTRU shows this is a key topic.

- One of the authors of RFC 3066 Bis said that what "is different is that these registrations will need to be evaluated for 'lang-or-extlang-ness', which is something we do not have today" (Addison Phillips). Actually this is something no one has today.
- This means that the IETF MUST enter the business of qualifying language name from the ISO language codes. The comments provided by the author of this proposition (who did not use it for his own ISO 639-3 standard) show the risk of many technical, cultural, political, or legal controversies over the discriminatory nature of the matter at hand.

I note that this wording is pure hazard. Legacy texts spoke of "language subtags". The first subtag to qualify a language was to be the "primary" language-subtag, and the others the "extended" language-subtags. The English text permitted to read this as "primary language" and "extended language" subtags.

2. To "clarify" non documented procedures and guidelines

The Charter does not discuss the responsibility sharing with the ietf-languages@iana.org mailing list. This is important when the only existing selection algorithm the Reviewer has at hand is the availability of a UNICODE CLDR locale file (cf. current Charter: "The RFC 3066 standard for language tags has been widely adopted in various protocols and text formats, including [W3C] HTML, XML, and CLDR, as the best means of identifying languages and language preferences.").

This means that the WG should therefore "clarify" the procedures and guidelines that are non-existent. Some Guidance should be provided, in the Charter or via an IAB RFC like in the OPES case.

3. To build over a lack of experience

The proposed Charter seems to believe that the current registration procedures respects the BCP 47 rules. The only current experience at hand is that the IESG does not respect these rules. Today, there is no actual experience of the way in which the currently approved documents are to be applied.

However, there certainly are reasons why the IESG considers the `ietf-languages@alvestrand.no` mailing list, and both the RFC 3066 Language Tags Reviewer and its owner Harald Alvestrand, as the RFC 3066 Bis `ietf-languages@iana.org` mailing list and its single Moderator/Language Subtags Reviewer). This reasons are only known to the IESG, the IESG should then give the proper indications to address this issue.

4. Consistence with ISO 639 series

The charter does not consider ISO 639-4, while ISO 639-4 establishes the common rules for the whole ISO 639 series, nor ISO 639-5, which complements ISO 639-3.

The charter believes that ISO 639-6 can only concern the long-term future of the Internet, while its delay is currently an urgent and tangible problem for several types of applications.

The Charter should only stipulate that the WG-LTRU is to consistently adapt the state of the art ISO 639 series, ISO 3166 series and UN.49, ISO 15984 (scripts), ISO 15879 (locale files), to the IETF language tagging system.

5. Archeolinguistic

The charter quotes signed languages, transliterations, and transcriptions, but only through use cases. It does not consider modern issues concerning normed, typed and printed, computable, networked, and translated languages, or computer languages. It does not consider the impact of text/word processors. It does not even consider the most usual architext issue like the HTML script and the content being in different language and scripts.

There are four reasons for language tags: (1) terminology, to understand the concepts under the discourse and possibly translate it, (2) information, to store and retrieve knowledge, and for human exchanges as (3) an author and as (4) an audience. The Charter should ask the WG-LTRU detail how langtags address them.

6. Multimedia, computer languages, and data interchanges not supported

RFC 3066 was multimodal in not documenting the mode. For an unknown reason, RFC 3066 Bis is unimodal (script only). Yet, it does not support computer languages such JavaScript, Perl, C, etc. and do not support interchange language such as the IETF protocols. This is very concerning as the Internet is multimodal. Even applications such as HTML, XML or CLDR are to support multimodal streams. The present state of the art makes that data interchanges and computer commands can be supported by human languages.

The WG should therefore organise a full multimodal (as ISO 639 is) support of every language following its main format repartition : languages (main and extended subtags covering referents and variants), modes (written [scripts], typed [keyboards], voice [accents], signs and icons - main and extended [styles]), and regions (ISO 3166 and UN.49 - main and extended local subtags).

7. Authors and relational spaces are not supported

The charter only considers the attribute of the text, not the attributes of the author and of the audience(s). When I send a mail on the IETF mailing lists, its text is in Franglish (one could tag as "en-fra-latn-fr") while the IETF targeted audience is "en-Latn". However, I would have tagged my texts as "en-latn-fr". These three usages are legitimate, but are not documented.

I note that up to now texts are classed and archived either by author or by relational space. The language of the text itself is only considered when dealing with translators.

8. Interoperability framework

The charter does not consider interoperability with any other existing or future non-IETF language code and practice.

ISO 639-4 will call for ISO 11179 (Metadata Registry) compatibility, or conformance. This seems required and should be addressed in two steps. First, a WG-MDR should document how such a conformity should be supported with a distributed and multilingual network architecture (the is the purpose of my MDRS project - multilingual distributed referential system). IANA Registries are actually updated directories. OSI Registries do not change values: they register new values with a date. They are like a CSV, they can document a dated obsolete situation.

When this is documented, the WG-LTRU should document how interoperability would be supported along with the RFCs of this WG-MDR.

9. BCP 47 Internal consistency

The charter does not consider the impact of its related work on the filtering/matching part of BCP 47.

10. BCP 47 Running code

Co-Chair documented irt. the filtering document of BCP 47 that RFC 3066 Bis format conformance could not be checked without calling on the IANA registry. Parser development seems to show that WG-LTRU Charter requested validity check is not easy to achieve. New documents may impose additional constraints. The WG-LTRU should publish a parser approved logic.

11. Intent or wording confusion?

The distinction is unclear between “fully approved and published” for starting the work and only “fully approved” for finalisation. What is intended?

12. IETF language support structure

The Charter MUST clarify who the IETF Language doctrine, strategy, and policy leader (cf. RFC 3935) is.

Languages are not application options of the ASCII Internet. Such a limited vision is obsolete and has already delayed IDNA deployment. This leads to a fragmentation <http://info.intgovforum.org/yoppy.php?poj=53> that we all clearly want to avoid.

This is why

- either, the WG-LTRU Charter must discuss an IETF/UNICODE MoU, delegating the management of the IETF Language area and the hosting of the Languages Registries, in the same way the Names and Numbers are delegated to ICANN.
- or, the IESG must discuss the creation of an IETF Multilingual Internet area. Its first priority will be to document what is a language in the IETF context and an internet architecture to support languages. This will provide the WG-LTRU the necessary elements to guide the ietf-languages@iana.org Moderator and Languages Subtag Registries Reviewer.

13. Workability of the Produced documents

Recent debate over the filtering document has shown that the current BCP 47 solutions cannot be fully deployed due to the load they would create on the IANA server. The organisation of the dissemination of the initial 820+ pages planned Registry should be part of the Charter. With indication of the update cycles (an equivalent to the TTL, permitting users to know that they have the current data from the IANA registry [we

eventually talk of a larger system than the DNS]).

Due to its size, the Charter should document the way the new Registry should be presented to the IESG for approval.

14. Well-formedness

"en-western' is invalid but not unusable. Maybe we should tighten up the restrictions in 3066ter." "What *is* discrepant is that nothing in 3066bis prevents a tagger from using a tag like "en-1901", even though that tag is not valid.". These quotes are part of recent exchanges. The rechartering should:

- include a review of the RFC 3066 Bis restrictions and document their possible flaws and the limits of further restrictions.
- provide a complete example list of well-formed and ill-formed tags (this list has been initiated) to permit the test of the well-formedness and validity checkers.

The WG should produce a review of the risks of errors should the checkers not be in position to access the IANA registry. Since the resulting document is a BCP the WG should report on the existing open use libraries and their experienced problems, in order to inform developers.

15. Privacy violations

The language tags in their present version publicly disclose information on the accompanying text and therefore disclose information on their author and on their readers they might want to keep private. This is particularly the case of "retro-meta-spam" or "dynamic cookie". This consists in spamming metadata in pages or mails. When the reader clicks on the next page or answers the mail, he/she signals that he/she is able to read the language. "Tell me the language you speak and I will know a lot on your culture, race, religion".

The WG-LTRU should be given the task to propose ways to crypt or hide the language/referent, mode/style and country/cultural actual information contained in langtag substags and their extensions.

16. Scalability

The Charter does not provide guidance about the support of idiolects (individual language) which will most probably become a technical common issue during the timeframe of the WG. The same the Charter does not provide guidance concerning non-geography related language (diaspora, trade idiolect, denier "EU" support). The Charter does not target the scalability required by the Internet architecture (RFC 1958).

17. multilingualisation

The charter should underline the need for the multilingualisation (localisation of the globalisation) of the WG-LTRU produced registry.