IP first round of feedback on Latin LGR (on Proposal of 20180525)

DATE: 2018-06-22

# Overview

The proposed Repertoire (dated 2018-05025) gives a multitude of code points for discussion, as below. But the proposal cannot be considered even formally complete without an accompanying XML file, which was not available until just before the IP completed this feedback and will therefore be considered separately.

# Conclusion

The IP submits the following set of preliminary comments without any claim of completeness, but in the hope the Latin GP finds them useful in continuing the discussion and developing a more formally complete proposal.

Attached is the original document with some mark-up. This should also be consulted, as not all marked-up issues are listed in the summary below.

# Comments on main document (.docx)

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| 1.0 | The code points from the two Unicode blocks, Latin-1 and Latin-Extended-A, appear in many “basic” sets of Latin characters. | It is essential to document the reasons why any omissions have been made from these blocks.  This general point stands, while specific kinds of cases are discussed in more detail below. |
| 1.1 | Some code points are used in languages or under circumstances that fall short of the criteria generally used. | The LGP might summarize its negative findings and omissions - at least for the two Unicode blocks here. For other Latin blocks, that do not contain code points that have been encoded well before Unicode, a more summary treatment may suffice. |
| 1.2 | It is also worth comparing the decisions that have been made for Latin code points in the Second-Level reference LGRs.  The details for specific languages are available at https://www.icann.org/resources/pages/second-level-lgr-2015-06-21-en | It is perfectly acceptable for Root Zone Latin to be held to a more restrictive standard than the Second-level. But reasons must be stated. |
| 1.3 | There is prima facie evidence for exclusion of CPs where there is no evidence for use in languages beyond EGIDS 1-4:  e.g. 48. U+0115 in Bavarian (5) ref. [111, 112]. | EGIDS level ≥5 languages generally should not serve as requirement for new code points. |
| 1.3.1 | Considering this in the context of other uses of the breve, the precomposed characters including the Breve mark over vowels, are mainly used in pedagogic or academic analysis of classical Latin verse. | Even though tfrequently used in that field, this is for linguistic analysis (comparable to items in the International Phonetic Alphabet), and therefore not one included in repertoires by the Procedure. |
| 1.4 | However, even in languages with EGIDS ≤ 5 there may be strong evidence of use in **everyday writing** whether established or on the rise (That is: the language can be seen as an emerging written language). | There may be exceptions to placing the boundary between EGIDS 4 and 5 if the language is written by a good percentage, and/or used in business. (These are both indicators that it may be appropriate for TLDs) |
| 1.4.1 | Also, in highly literate and multilingual communities (e.g. NL) dialectal languages may be classed surprisingly high on EGIDS: e.g. Ethnologue lists West Frisian as (2), where Ethnologue also mentions that "literacy level is low". | Whatever the assessment of West Frisian, everyday written use remains the criterion |
| 1.5 | Overall, a language should not be used to justify inclusions if:  a) it is not written/not usually written/ not commonly used beyond orally; b) if written, it does not have a stable orthography (even though the main language in the culture has had a stable orthography forever) then. | On EGIDS level ≥5, some of the details of written use do matter: for example, "educational support" ought to go beyond keeping alive a dead language as part of "cultural heritage"; evidence of ***everyday written use*** isrequired. |
| 1.6 | It is also possible that EGIDS status is assigned mistakenly. | e.g. West Frisian is listed by the GP proposal as EGIDS (1), whereas Ethnologue list it as (2). This is pointed out in the attached annotated .docx proposal. |
| 2.0 | Potentially Unjustified exclusions:  The IP noted that the following code points were excluded:   * From Latin-1   U+00FF LATIN SMALL LETTER Y WITH DIAERESIS   * From Latin Extended-A   U+0109 LATIN SMALL LETTER C WITH CIRCUMFLEX  U+011D LATIN SMALL LETTER G WITH CIRCUMFLEX  U+0125 LATIN SMALL LETTER H WITH CIRCUMFLEX  U+012D LATIN SMALL LETTER I WITH BREVE  U+0135 LATIN SMALL LETTER J WITH CIRCUMFLEX  U+014F LATIN SMALL LETTER O WITH BREVE  U+0157 LATIN SMALL LETTER R WITH CEDILLA  U+0163 LATIN SMALL LETTER T WITH CEDILLA  U+016D LATIN SMALL LETTER U WITH BREVE  U+0177 LATIN SMALL LETTER Y WITH CIRCUMFLEX | The IP does not necessarily request that all these code points should be part of RZ-LGR, but at minimum their exclusion should be discussed in some detail in the proposal.  In general, please offer documentation or justification of the decisions on the listed code points.  All decisions by the GP should always be fully documented..  The IP will expect a detailed explanation for exclusion of characters that seem to be part of popular writing systems (French) or very common recent invented language (Esperanto). Of all the invented languages, this one may arguably have enough (written) use as L2 to at least warrant an evaluation. |
| 2.1 | Re French:  00FF y with diaeresis - 2nd level reference LGR has it for French - it's rare but the only "Latin-1" code point not covered.  U+00FF is typically part of the French writing system (thus its presence in Latin-1), but is relatively rare in usage. It is mostly used in proper names (cities, historic persons, etc…). | [Interesting enough, today while visiting a museum in Quimper, France, one member of IP saw a usage in common name (this case was paÿs). It was on an old inscription, therefore a bit historic, perhaps before spelling stabilization. But it shows that excluding in the French context might be hasty]. |
| 2.2 | Re Esperanto:  Concerning the excluded code points from Latin Extended-A, the vast majority are used in Esperanto. While Esperanto is a ‘recent’ invented language (EGIDS 3), it by far the most popular of its kind, and while not a primary language for probably any one, it is practiced by millions of users. | We should expect, given the size of the user community, some explanation of their decision. |
| 2.3 | Other Specific causes of interest (as examples):  U+0157 (Livonian) and  U+0177 (Welsh)  may be thought problematic, but the cases are very different.  U+0163 (T with CEDILLA) is different again. |  |
| 2.3.1 | U+0157, r with cedilla, was used in Livonian, an extinct language of Latvia. It is listed in the- 2nd level reference LGR for Latvian.  Of Livonian, Ethnologue says: "no known L1 speakers. Last speaker died in 2013. Ethnic population: 200."  (It was included in Latvian in the- 2nd level reference LGR because several references still listed it, but it may be OK to not perpetuate that in the RZ. |  |
| 2.3.2 | U+ 0177 (Small Letter Y with Circumflex) is used in Welsh precisely on a par with U+0175 (Small Letter W with Circumflex).  It just happens that the former is also used in Chichewa (EGIDS 3). But since Welsh is EGIDS 2, both characters should be included.  U+0163 LATIN SMALL LETTER T WITH CEDILLA is part of a dis-unification process with U+021B LATIN SMALL LETTER T WITH COMMA BELOW and therefore may still exist in Romanian corpus. |  |
| 2.3.3 | U+0163 (T with CEDILLA) is an interesting case because it is related to now Romanian U+021B LATIN SMALL LETTER T WITH COMMA BELOW, the latter was added as a dis-unification from U+0163. It is likely that U+0163 will be present in Romanian corpora.  T with Cedilla exists as part of the General Alphabet of Cameroon Languages, in the Kabyle dialect of the Berber language, and possibly elsewhere.  It may be fine to exclude U+0163, but more care is needed. Given its possible variant situation with U+021B it is not a decision we can reverse easily.  And more from wiki (<https://en.wikipedia.org/wiki/Cedilla> ) concerning U+0163:  (The MSR only omits "kra" from the Latin Extended-A block, but Unicode is quite clear that some are not used for languages, but for transliteration, e.g. T with Cedilla is supposedly for Semitic transliteration). |  |
| 3 | Code points outside the Maximal Repertoire  Repertoire: code point U+1E13  Repertoire: code point U+1E4B  Repertoire: code point U+1E71   These are all listed as Venda (1), a statutory language of South Africa with L1 1mio+. There are no good data on how widely it is written, though Wikipedia has 250 entries. | please supply documentation. |
| 4 | Section 5.4 ‘Code points proposed to be added to MSR-3’. | This section needs to be revised to take into account the actual content of MSR-3 and the conclusion that was achieved after discussion between the Latin GP and IP. |
| 5 | ‘COMBINING MACRON BELOW’. U+0331  It may be unwise to use this combining mark in domain name because it will look like an underline and therefore disappear in the common display of IRI which are typically underlined.  At minimum it will likely be confused with the form w/o the combining mark.  Applications differ in where they place an underline, but some will overprint the combining mark:  C:\Users\asmusf\AppData\Local\Temp\ogpcgbdkellpclgb.png | Please reconsider  U+0331 |
| 6 | African writing systems using Latin based characters.  This area needs an extensive development to describe in detail the repertoire requirement.  The Latin GP may consult the Arabic LGR proposal which had to address a similar issue for Arabic based writing systems in use in Africa. | In particular, raw EGIDS numbers can serve only as a rough guide; where there's any reasonable doubts, details about actual current use and the degree to which the language is written in *everyday use* is to be adduced.  At some point IP will have to accept expert opinions, but with a good dose of external review. |
|  |  |  |

# Comments on LGR specification (.xml)

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
|  | An XML was recently submitted by Latin GP but IP has not been able to consider it. | IP will review that file not only for agreement with the table, but for all the required points of information and provide additional feedback under separate cover. |

# Comments on Test Labels

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
|  | As yet there is no test label file. | Please supply. |
|  |  |  |