**Proposal -** How to proceed with attesting Latin letters in MSR2 using available sources, starting from language alphabets side

1. Description of working material

* [02-REP-LanguageAZ-EGIDS-1-4.docx](https://docs.google.com/document/d/1mOIzm5EfpFJmMZmxsrloKdQsqRo8q9hYRcCmw_fxXFo/edit?usp=sharing) contains 180 languages using Latin script with EGIDS = 1 to 4. In the same table 10 languages are marked processed by members of GP during testing the time necessary to process one language. These languages are marked with the name of person who has processed the language and a comment “Should be processed again with Final Principles for inclusion and exclusion”  
    
  Mirjana has tried to convert the word table to excel, to facilitate the use, but during conversion some links to omniglot site are lost, and for now we need these links. Conversion should be investigated in future.
* [02-REP-LatinPart-MSR2-Attestation](https://docs.google.com/spreadsheets/d/1fFe-uo5Gbh6nxf3c8LIoP_UvOpGN9lH-8RoPXAdeq8c/edit?usp=sharing).
  + **First sheet** (Latin part of MSR2 with Attestation) contains old table made by Chris Dillon with MSR2 code points starting with code point 00DF. Lower code points are plane ASCII and they are not interested for our purpose. Letters with “CLDR” in column Ref are attested.
  + **Second sheet** (Code Points not in MSR2) contains letters found in processed languages but not included in MSR2
  + **Third sheet** (To be investigated) contains everything else found in processed languages (for example digrafs, letters composed with one letter and one or more marks, ...). Column Ref contains the initials of the member who processed the language.
  + **Third, Forth and Fifth sheet** are inherited from Chris Dillon

1. Every member willing to participate in the proposed test for collecting material for discussion on Principles for inclusion and exclusion should choose 3 to 5 languages (as Meikal proposed) and try to process them using these instructions**. When someone chooses the languages she or she should put her/his name in the column “Processed by” in the table** [02-REP-LanguageAZ-EGIDS-1-4.docx](https://docs.google.com/document/d/1mOIzm5EfpFJmMZmxsrloKdQsqRo8q9hYRcCmw_fxXFo/edit?usp=sharing)**.**
2. The second column of the table [02-REP-LanguageAZ-EGIDS-1-4.docx](https://docs.google.com/document/d/1mOIzm5EfpFJmMZmxsrloKdQsqRo8q9hYRcCmw_fxXFo/edit?usp=sharing). has a link to the omniglot entry for this language. If some of the links don’t work, this case should be notified in the column Comment for the selected language. Mirjana will correct the link. Another language could be chosen.
3. On the omniglot page for a specific language, usually, the alphabet for the selected language is given. The alphabet is given as image. If there is a need to get character from the alphabet, some other source must be found. On the same page, at the end, there are other links that can be used for further identification of characters used in alphabet. Usually, the Wikipedia link proves to be very useful.   
   **Notice**: only lowercase letters should be considered.
4. All plain ASCII charaters should be skipped. The investigation of other character should be started with copying the character to the clipboard. When the character is copied to the clipboard following site should be used to get UNICODE representation(code) of the character

<http://unicode.org/cldr/utility/character.jsp>

If it is difficult to find UNICODE code point for specific character in the alphabet, this case should be noted in the third sheet (To be investigated). Minimum set of data entered should be

* 1. Glyph (if the source allows copy/paste function)
  2. Language using the code point
  3. Ref – Initials of the member who processed the language
  4. Comment on the sources which provided relevant information
  5. Any additional information that was found

1. After finding the UNICODE code point for the letter, every letter in the specific alphabet should be investigated and checked against [02-REP-LatinPart-MSR2-Attestation](https://docs.google.com/spreadsheets/d/1fFe-uo5Gbh6nxf3c8LIoP_UvOpGN9lH-8RoPXAdeq8c/edit?usp=sharing). using the UNICODE code point.

There are three possible cases:

1. If the letter is previously marked with CLDR in column “Ref” in the table [02-REP-LatinPart-MSR2-Attestation](https://docs.google.com/spreadsheets/d/1fFe-uo5Gbh6nxf3c8LIoP_UvOpGN9lH-8RoPXAdeq8c/edit?usp=sharing). , this letter has to be skipped, it is attested.
2. If any letter (without CLDR in column Ref) is found, the row in the table must be filled with available information. For this test, the initials of the person who has processed the letter should be put in the column “Ref”.
3. If any new letter, which is not in the table [02-REP-LatinPart-MSR2-Attestation](https://docs.google.com/spreadsheets/d/1fFe-uo5Gbh6nxf3c8LIoP_UvOpGN9lH-8RoPXAdeq8c/edit?usp=sharing). would be found, it must be added in the Second sheet (Code points not in MSR2) with all available information. Also, the initials of the person who has processed the letter should be put in the column “Ref”.
4. All combinations of the letters and marks which do not fit in previous cases should be written in Third Sheet (To be investigated)
5. Minimum set of information on specific new letter that must be filled in the Second sheet (Code points not in MSR2) are:

* Unicode code
* Glyph (if the source allows copy/paste function)
* Example of language using the code point
* UNICODE character name
* Comment on the sources which provided relevant information

1. When someone finishes the processing of the language she/he should put “Finished” in the column “Processed by”, after the name written in step 2, in the table [02-REP-LanguageAZ-EGIDS-1-4.docx](https://docs.google.com/document/d/1mOIzm5EfpFJmMZmxsrloKdQsqRo8q9hYRcCmw_fxXFo/edit?usp=sharing)