Requirements and Process for Changing Script Status for Identifier Use

Asmus Freytag

January 7, 2023 – (supersedes L2/23-054) (REV 6 of the proposal)

Script Status

UTS#39 defines Identifier type Recommended as characters in "widespread common everyday use". Formally, the definition is based on membership of the character in a Recommended script in UAX#31 (with some exceptions). *Recommended* scripts are therefore in "widespread common everyday use", while other scripts with less active modern use might be classed as *Limited_Use*. There is a third class of scripts, *Excluded*, which covers scripts that are practically without living native users; that represent notational systems; or that are otherwise unsuitable for identifiers.

These characterizations are not permanent in every case; they are intended to track actual use of a given script, including any significant changes in usage over time. The definition of Recommended script is used as input to other specifications outside of the Unicode Standard, such as the Label Generation Rules for the DNS Root Zone (see "Root Zone LGR" under https://icann.org/idn for details).

Because of such dependencies, it is advisable to use a very deliberate process when adjusting the status of a script in UAX#31 (and therefore the Identifier_Type of its member characters). Such a process must first and foremost establish, for example, whether in the context of identifiers the usage for a Limited_Use script has changed sufficiently so that it fits the requirements of being in "widespread common everyday use", or whether relevant usage has effectively ceased, making it a candidate for adjusting the status to Excluded

This calls for a clearer understanding of the criteria that determine whether a script is considered Recommended, Limited_Use or Excluded.

Evidence Supporting Script Status Assignments

The following discusses acquiring, organizing and weighing evidence of script usage with the aim of assigning or modifying a script's status as **Recommended, Limited_Use** or **Excluded** in the context of identifiers. The process of assigning a status should not be seen as automatic or mechanical, and especially for Limited_Use scripts, there's rarely a single item of evidence that is determinative; instead, all of the evidence should be considered together and in context.

The purpose for an identifier is not as much coverage of one or more specific orthographies and documents, but allowing useful mnemonics —including certain kinds of non-words – that can be used as labels. Unlike when a script is first proposed for encoding, what is important is not whether a script is found in documents, past or present, but whether there is a community that is actively conducting its daily business in that script, in settings where online identifiers are of common concern. In addition, any

reclassification of its identifier type would take place after the script has been implemented and is already available for use.

For identifiers therefore, any documented "active online use" of a script should be weighed fairly high. In some ways, depending on how extensive such use is, it can be both a necessary and a sufficient condition.

If a script is only found to be used online in specialized settings, as opposed to everyday ones (such as social media, or administrative use), then this would indicate the absence of a user community focused on conducting their ordinary business in this script. In contrast, given strong positive evidence of such use, the status of the script should reflect the extent of such use, which is best considered as a combination of pervasiveness of use coupled with size of the user community (considered together, not separately).

Suggested Types of Evidence to be Considered

Evidence for active online use would include everyday online use of the script, for example in social media, including titles, description and comments on videos or images; an actively maintained Wikipedia in the script; online news, particularly if independently produced; as well as commercial, administrative or governmental websites, and so on, especially interactive ones, such as order forms or application for standard services.

A useful type of evidence for "widespread everyday common use" is whether a script is used for the principal language of instruction. This includes evidence of online instructional materials or publication of math, history, science or other textbooks required in a large number of primary and secondary schools in that script.

Online search would readily find a variety of entries for common search terms. If, on the contrary, the script is primarily used in the preservation of cultural heritage with day-to-day activities of the user community conducted in other scripts, then that would argue against making a change in classification at this time.

In principle, a small user community alone does not disqualify; for example, where a script is used as the primary or exclusive script in a country or region. For these users, even if the community is small, the script is clearly in "widespread, common everyday use". This is to be seen in contrast to scripts that are used as alternative to a dominant script for the same language. In the latter case, there are a number of factors that weigh in favor or against the proposition that the script is in "widespread everyday common use".

Beyond observation of online use, there is little reliable and direct information on script use by various populations. This is particularly true for most of the scripts currently considered Limited_Use. Some conclusions about likely usage levels for a script can be derived from available census data on the principal languages for which that script is used, and factors such as literacy levels and or use of alternate scripts for the language.

Such data is available for individual languages and their user community, but also the degree to which the language is in active use (for example: Ethnologue) and being actively transmitted to the next generation of speakers (see Expanded Graded Intergenerational Disruption Scale or EGIDS). With some care, like factoring literacy data and the effect of competing scripts, such data can be a useful proxy for some of the information that may not yet have been compiled on the script level.

When a script is a customary written form for a language with an EGIDS[1] level of 0 to 4 it could be assumed to be in "widespread everyday common use" — particularly if no alternate scripts serve the same language communities in day-to-day contexts. Where there is another script in more widespread use for that language, particularly if that script is Recommended, it would undercut that presumption.

In such cases, or in cases where there is no cohesive community that uses a script exclusively for its language, or the language is not the primary or exclusive one used for day-to-day activities, any reclassification would have to be based on detailed further arguments supporting the conclusion that the script is in "widespread, everyday, common use" at this time.

Proposals for Reclassifying a Script

The way a script is used, and the number of users may change over time or more information about these aspects may have become available. Either of these reasons may prompt a proposal for formally reclassifying a script for identifier purposes. To be useful, such a proposal needs to present evidence.

In addition to presenting evidence of everyday non-specialist online use, a proposal to reclassify a script should provide supportive evidence based on the status of the languages, the size of their user communities, literacy levels and any alternate scripts used in the same communities.

Submission Requirements

A submission must identify the script, list it's current status as well as the proposed adjusted status in UAX#31. The proposal must be accompanied by information on the submitter, and experts or native users of the script ready and willing to help settle any questions raised in review.

Any proposal must be accompanied by a thorough summary as well as detailed citation of the available evidence, paying particular attention to the types of evidence and data suggested in the section of Evidence to be Considered, above.

The submitted proposal must contain the necessary argumentation, explaining how the evidence argues for the intended outcome, but also accounting for the quality and comprehensiveness of the adduced data. All corroborating evidence must be supplied with the proposal, or openly available online at a stable location.

Any data available on confusables within the script and with other scripts should be supplied.

---

[1] EGIDS: https://www.ethnosproject.org/expanded-graded-intergenerational-disruption-scale/

Review

The Unicode Script Ad-Hoc and the Properties (SAH) and Algorithms Group (PAG) are jointly tasked with reviewing all submissions, but the PAG will act as point of contact. Those deemed incomplete, or with unpersuasive evidence will be returned. In all other cases the reviewers will evaluate the evidence and, on that basis, arrive at a recommendation to be presented to the UTC either arguing for or against making and adjustment.

The number of scripts is bounded, and the number of candidates for a possible re-evaluation is even more limited. Rather than approaching this review as based on a series of inflexible rules, it is probably best to recognize that each will by nature result from a case-by-case decision. The task therefore is to acquire available data and to find whether they support a persuasive rationale for why that script should now be considered in "widespread common everyday use".

As Recommended scripts are particularly sensitive in the context of identifiers, it is imperative that confusables can be mitigated. Until such mitigation is available, Recommended status will be on hold.

Decision

The final decision rests with the Unicode Technical Committee, which will publish its conclusion in the meeting minutes.