L2/25-069

Factors used in determining the Identifier_Type of characters

Last updated: March 3, 2025

Introduction

Not all characters that are members of scripts defined as "Recommended" in UTS39 are equally widely used. Some of them may be rare or even unfamiliar to most users of the script, or, upon being encountered, would most likely not be recognized for themselves, but as a font variation of a more familiar character. They thus fail to serve the purpose of identifiers in providing a "useful mnemonic".

UTS39 defines "Recommended" scripts as those that are in "widespread everyday common use". In principle, this concept should carry through when defining character subsets inside these scripts, but that needs a bit more precision to define what is and isn't common use for the purpose of identifiers.

Among subsets of characters from a recommended script that should not be allowed, there are some categories for which the code points are ineligible for identifiers on technical grounds. Their Identifier_Type assignments are directly derived from other properties and therefore do not enter here (such as Deprecated or Not_XID[1]). Some others are considered for the purposes of default identifiers to be problematic enough to need some explicit decision for inclusion.

The three Identifier_Types, **Technical**, **Obsolete** and **Uncommon_Use** are the primary categories that break down characters that are not allowed, though they are members of a Recommended script.

Several considerations come into play when defining factors to be used in assigning one of these identifier types. The first is that while it may be more easily possible to affirmatively decide whether a character is "Obsolete" or encoded for "Technical" or other specialized uses, it is much harder to come to a reasonable cutoff for "Uncommon_Use".

In some cases, as for ideographs, there's simply a sliding scale with no hard cutoff. In other cases, different communities may regard different characters as uncommon, and the cutoff is determined by the list of supported orthographies.

It may be useful to turn the problem on its head by affirmatively looking for characters that are **not** Uncommon_Use; by looking for characters for which there is evidence of "living people using these letters in their day-to-day orthographies and accepting them in identifiers would help those people".

Language and Orthographies

This leads to a mix of "objective" factors combined with more subjective review by experts with linguistic or other expertise in the identifier domain, such as registry operators with respect to internationalized domain names in the different regions/countries.

[1] Validity of programming language identifiers is governed by properties, such as XID_Start and XID_Continue.

The goal should be to exclude support for orthographies that have been abandoned or are only being used by languages that have been relegated to a secondary status in the speakers' daily life. A European example would be Bavarian, which can technically be classified as its own language, though it is primarily a spoken language, with high German as the common written language in Bavaria and Austria. Even though there are millions of speakers and there exist orthographies for it that use some accented Latin characters, those characters should not be considered in common use, because the language is not written for "everyday" purposes, but only to transcribe/archive.

Similar considerations apply to some of the liturgical languages, or uses of specialized orthographies for them.

But at the same time, the goal should be to make sure to not exclude smaller languages, especially ones where there is data on them being actively written, or in some cases recognized as official languages in some jurisdiction.

Language Status as Proxy

There aren't any solid data on script usage that can be plugged in to satisfy some objective criteria. A multi-year effort by ICANN to define a suitably restricted yet useful subset of characters for use in IDNs in the DNS Root Zone and Second Level ended up substituting a measure of language vitality documented in the EGIDS scale [EGIDS].[2] Other information may be available on whether a language is threatened or endangered [Glottolog].

EGIDS levels address the status of and support for a language but not for any associated orthography. However, they can be used as a proxy in making an initial and approximate cutoff for which languages to support, but it is still necessary to establish that an orthography exists for that language in the given script and that it is stable and one that is actively used for everyday purposes.

With very few exceptions, writing systems for languages of EGIDS level 4 or better would qualify. The few exceptions are writing systems for which an argument can be made that they are not in general use or not in the preferred script. Languages that are classified as 5, on the other hand, might bear investigation to get more direct information on the level of use of any associated orthography. Sometimes, only the lack of institutionalized instruction marks the difference between these two levels, but the language and writing system are otherwise in vigorous use and stable.

If there are credible suggestions that a language uses a different script in preference to the "native" one, or that a language is used only orally, it should be excluded from consideration for selecting the recommended subset for that script, independent of any EGIDS level.

Ultimately, the goal is to support those orthographies used by living languages that are in active use for "everyday purposes".

[2] The main source of EGIDS classification is no longer freely available. However, several other sources list languages with this classification, such as [Glottlog](https://glottolog.org) or WikiData. Glottolog also contains other measures of endangerment of languages.

Direct Evidence of Online Use

Direct evidence of the type and range of online use may help make the decision of whether a character is or isn't in common use and whether the usage is "general" or perhaps only "incidental" or "specialized".

Online usage has the advantage of being accessible to observation.

In addition, there are available corpora[3] that can be consulted to determine whether some character, while used with the language, is common enough to be required for identifiers.

When considering other evidence such as the existence of a Wikipedia in the writing system in question, care must be taken to evaluate whether that represents a genuine effort or is more of a demonstration project, feasibility study or an expression of enthusiasm for the language by the creators. The number of contributors and their activity level may be a useful data point.

When considering evidence on the use of the writing system online, care must be taken to distinguish websites that are concerned with the language itself (learning guides, dictionaries) from websites that use the writing system for other purposes from daily life, including administrative and commercial interactions, or heavy use on social media. From an identifier perspective, any use of the writing system to label contributions or contributors is a stronger indication of suitability for identifiers than just ordinary text.

In addition, any availability of data on the level of literacy, population size, and whether the orthography is the preferred one for the language community can supplement evidence of online use.

Established Practice

There should also be a goal of identifying and vetting existing identifier practices. Particularly for IDNs, some domain registries have taken great pains to curate their repertoire based on the best information available to them, while others have been content to simply copy the PVALID subset of one or more Unicode blocks wholesale. Aligning the set of recommended characters for identifiers with existing identifier practices of the former kind can be valuable. For example, after some discussion, UTC is adopting the combination of existing, locally deployed Han subsets [L2/25-031], instead of defining a Unicode-specific collection.

At this point, the published and implemented definition of a recommended set of characters for the DNS Root Zone [RZ-LGR] and a set of reference subsets for the Second Level [RefLGR] should be treated like other "registry practice" from Unicode's perspective, and therefore the goal would be to minimize discrepancies to those motivated by differences based on the nature of Unicode default identifiers compared to IDNs.

[3] For example, the University of Leipzig makes available a large set of corpora for languages using the Latin script.

Alignment with existing practice should best be understood as helping determine a better default assignment of Identifier_Type, as opposed to a constraint on which identifier type values can be assigned based on substantiated and documented evidence. If the latter meets the criteria established here, it should always be sufficient to override any Identifier_Type in favor of Recommended, or to adjust any assignment of Uncommon_Use, Technical or Obsolete to better fit the facts.

Other types of identifiers

While IDNs are an important use case and define a deliberately conservative set of identifiers, there are other types of identifiers, including user handles, that should be considered in defining the Identifier_Type property for use with Unicode default identifiers.

For example, IDNs are limited to lowercase in IDNA2008, but that should not prevent default identifiers from containing uppercase characters. Their Identifier_Type should match that of their lowercase equivalents. IDNs are in NFC, but it has been the practice to not limit the listing of Identifier_Type to composed characters. However, combining characters may be assigned Uncommon_Use, even if they occur in canonical decompositions of Recommended characters. Implementations working in NFC may wish to treat them as such.

There may be other instances where the definition of default identifier could end up being a bit broader than what established IDN practice suggests, so the goal is not to limit default identifiers to those suitable for IDNs, but to minimize the deviation with justification.

When to use which Identifier_Type

The following identifier types may be assigned singly or in combination. The values are based on best available information and may be updated when new information becomes available. Multiple classifications may be possible, particularly where a character is of one type in a commonly used writing system and of different type in another context.

- **Uncommon_Use** should focus on usage of the character in orthographies for living languages and is assigned where this character or the associated orthographies are not in common use. Uncommon_Use can also represent the absence of confirmed or credible data for a level of usage that would correspond to "common everyday use" for an orthography in widespread use. Uncommon_Use is the default for newly encoded characters for all scripts other than Excluded scripts, unless a sufficient level of usage can be confirmed for one or more specific orthographies at the time of encoding.
- **Obsolete** should focus on the degree to which a character is in common modern use. If a writing system or orthography has fallen out of use, or a character is no longer used in a given context, that could make classification as Obsolete the appropriate choice. A character can become obsolete in the context of a writing system; it is not required that the entire writing system have fallen out of use. A character may be Obsolete in the context of a widely used writing system, but also part of an orthography where it is in Uncommon_Use.
- **Technical** should focus on the purpose of use. If a character is limited to particular types of texts or forms part of a notation without concurring everyday use, then it would be appropriate to

categorize it as Technical. Technical uses can comprise for liturgical purposes, poetry, phonetic notation and so on. A character may have a common technical use but also be used in one or more orthographies at a level that is marked by Uncommon_Use, or it could be a Obsolete as a character for general use.

- **Inclusion** should focus on punctuation, or characters that look like punctuation, and that should not be automatically included in identifiers but may be appropriate in specific identifier environments. Usually, the reason for not allowing the character is that it can be confused with syntax characters in the given environment.

The other Identifier Types are largely, if not fully determined by a character's other property values and are therefore automatically assigned.

Documentation

Finally, assignments of Identifier_Type values **Obsolete**, **Technical**, **Inclusion**, and, for new characters, overriding the default of Uncommon_Use in favor of **Recommended** should be documented on the character level, citing at least one source that determined the choice of assignment, although the reviewers may have consulted additional material.

References

[EGIDS] Lewis and Simons, EGIDS: Expanded Graded Intergenerational Disruption Scale,” documented in [Glottolog] and summarized here: https://en.wikipedia.org/wiki/Expanded_Graded_Intergenerational_Disruption_Scale_(EGIDS)

[Glottolog] Glottolog 5.1 edited by Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian, https://glottolog.org

[RefLGR] ICANN, “Second-Level Reference Label Generation Rules”, https://www.icann.org/resources/pages/second-level-lgr-2015-06-21-en

[RefLGR-Overview] ICANN, “Reference Label Generation Rules (LGR) for the Second Level — Overview and Summary”, https://www.icann.org/sites/default/files/packages/lgr/lgr-second-level-overview-summary-25oct24-en.pdf

[RZ-LGR] ICANN, “Root Zone Label Generation Rules”, https://www.icann.org/resources/pages/root-zone-lgr-2015-06-21-en