Integration Panel:

Maximal Starting Repertoire — MSR-5 Overview and Rationale

REVISION – APRIL 06, 2021 – PUBLIC COMMENT

Table of Contents

1 Overview

This document describes the Maximal Starting Repertoire (MSR) for the Label Generation Rules (LGR) described in "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels" [Procedure]. The Procedure defines a two-stage process, in which community based Generation Panels (GP) propose LGRs specific to a given script, which are then reviewed and integrated by the Integration Panel (IP). To provide a starting point for the GPs the IP is tasked with establishing the maximal set of code points as well as the default whole label variant evaluation rules for the Root Zone. Collectively these are the MSR.

The repertoire covers scripts that are in widespread, everyday common use and thus essential for mnemonic labels for TLDs. Within each script, only code points supported by IDNA 2008 and appropriate for the Root Zone are selected. In particular, included code points are in widespread, everyday common use for each script, as opposed to code points only used in historic documents, or other specialized or limited uses. This document gives the detailed rationale used by the Integration Panel (IP) in defining the MSR, and also gives guidance to the Generation Panels (GPs) on how to use the MSR in generating proposed LGRs.

The reader of this document is assumed to be familiar with the [Procedure][1], particularly the parts that describe the role of the IP and the tasks and expectations on the GPs. Relevant parts of the [Procedure] are repeated in this document, but the [Procedure] as a whole is the formal framework on which the MSR is based.

This document is part of the fifth version of the MSR (MSR-5), which supersedes the all previous versions of the Maximal Starting Repertoire and extends the repertoire of [MSR-4] by adding once code point each to the Latin and Devanagari scripts and 2 code points to the Arabic script. MSR-5 makes no updates to the repertoire of any of the other scripts already found in MSR-4. The detail changes in repertoire from MSR-4 are documented below in Section 8. The full content of MSR-5 is specified in a set of files as described in the next section.

2 Maximal Starting Repertoire (MSR-5)

2.1 Files

2.1.1 Overview

MSR-5 is provided as a collection of files that are self-contained and supersede the files from previous version(s). The current document, "Maximal Starting Repertoire - MSR-5-Overview and Rationale-2021-04-06" provides background on the content and development of this version of the MSR, including a discussion of the methodology used and rationale for specific design decisions. It also provides

[1] References to documents cited are provided at the end.

additional guidance to Generation Panels on using the MSR as basis for their LGR proposals to the Integration Panel.

2.1.2 Normative Definition

The normative definition of MSR-5 is provided as an XML. The MSR is expressed using a standard format defined in "Representing Label Generation Rulesets in XML" [RFC7940]. At its core, the MSR consists of a list of code points defining the repertoire plus a set of Whole Label Evaluation (WLE) rules defining the default rules for the root zone. Each code point in the file is annotated with the Unicode version in which it was first assigned, and the scripts in which it is used. CJK Unified ideographs are further annotated with the source sets from which they were entered into the MSR.

The XML format defined in [RFC7940] supports the specification of variants and their disposition; however, these features are unused in the MSR. The Generation Panels are expected to use the full XML format when submitting their script-specific LGR proposals to the Integration Panel (including variant information where applicable).

Each code point in the MSR is tagged with one or more Unicode Script Identifiers[2] prefixed with “sc:” following the convention in [RFC7940] for designating Unicode property values. These tag values match the Unicode Script Extension property (“scx”); where that property value consists of multiple script IDs, the tag will have multiple values, for example, for U+3006 IDEOGRAPHIC CLOSING MARK the tags are “sc:Hani sc:Hira sc:Kana”. (For details on the Unicode Script Property see [UAX24]).

2.1.3 HTML Presentation

A mechanically generated, non-normative HTML presentation of the MSR is provided for ease of review. This presentation is augmented by summary data, as well as data extracted from the Unicode Character Database [UCD], such as the character name and Unihan source information for CJK Unified Ideograph code points.

2.1.4 Code Charts

A non-normative PDF file MSR-5-Annotated-non-CJK-Tables-20190125 shows the repertoire for the majority of scripts in the MSR presented in the form of marked up tables in a format similar to that used for character code charts in the Unicode Standard.

- Code cells with yellow highlighting are part of the MSR.
- Code cells without highlighting (that is, white cells) are for code points that are not PVALID in IDNA 2008 [RFC5892] or [IDNAREG].
- Code cells with blue highlighting are ***excluded*** from the MSR because of the Letter Principle.
- Code cells with pink highlighting are explicitly ***excluded*** from the MSR based on a specific rationale for exclusion, usually based on typical or intended usage.. This also applies to code points defined in Unicode 12.0 or later, but not included in the MSR (see Section 3.6).

[2] Unicode Script Identifiers are a subset of those defined in [ISO 15924]. Composite scripts, like “Jpan” would correspond to three element scripts each designated with a Unicode Script Identifier: Hani, Hira and Kana.

- Excluded code points are shown for reference for any block of code points that also contains part of the repertoire of the MSR, however, any blocks consisting *entirely of excluded code points* are suppressed from the listing, as are any other blocks not containing any part of the MSR repertoire. (See Section 5.17)
- There is no special highlighting for code points that were added to MSR-5. Note that MSR-5 adds no new scripts, the only changes to the repertoire are additions for existing scripts discussed in Section 5 and listed in Section 0.

The tabular listing of Unicode character names contains additional information about certain code points; for excluded code points listed, a shorthand notation for the principal rationale leading to the exclusion of the code point is provided. See the cover page of the file.

Because of size, the tables showing repertoires for Han ideographs MSR-5 Annotated Han Tables 2021-04-06 (PDF) and Hangul syllables MSR 5 Annotated Hangul Tables 2021-04-06 (PDF) are broken off into separate PDF files. For these files, a lack of highlighting (white cells) represents code points excluded from the MSR. The contents of the Hangul tables are unchanged from MSR-1. For CJK Unified ideographs, the file repeats the source information published in the latest available version of the Unicode Standard.

2.2 Determining the Contents of the MSR

The [Procedure] contains a number of explicit and implicit prescriptions on how to define the maximal repertoire. A key aspect is the adherence to the set of Principles[3] defined in [IABCP]. While these principles apply to the overall process of defining the integrated LGR for the root zone (RZ-LGR), they suggest a certain approach for the Integration Panel to follow in developing the MSR.

The following sections describe the approach taken by the IP in determining the status of a script or code point, in accordance with the [Procedure]. Whenever the IP determined that there is some uncertainty in establishing the status of any code points or scripts, the IP uses the following guidelines in deciding whether or not to include them in the MSR.

IP's Approach for Determining the Contents of the MSR

- *Script-level determination*: a script will only be included in the MSR if the Integration Panel has conclusively determined that a script is appropriate for the root zone.

- *Character-level determination*: If a script has been included in the MSR:

  - All its code points will be included in the MSR for detailed review by the GP except for those that the Integration Panel has conclusively determined to be inappropriate for the root zone.

[3] With exception of the letter principle the applicability of [IABCP] is not restricted to the Root Zone. Likewise, the MSR, augmented by any code points excluded solely because of the letter principle, may serve as the starting point for LGR development in other zones.

- If, while integrating the LGR proposals, the Integration Panel cannot conclusively determine that a code point is appropriate for the root zone (based on an LGR proposal and the proposal's justification, in the light of any expert advice), the code point will not be accepted.

Because the MSR is a framework for the GPs to do their work, any pre-emptive removal of code points is done with the intent of limiting the remaining code points so that the GPs can focus on code points that are relevant. If the IP has any uncertainty about the usage status of any individual code point, but no security concerns, the code point is included in the MSR. The Generation Panel will be best situated to review these particular code points and to propose a disposition for them in the proposed LGR. In general, it is expected that Generation Panels will propose to include only a subset of code points that are in scope for their respective scripts.

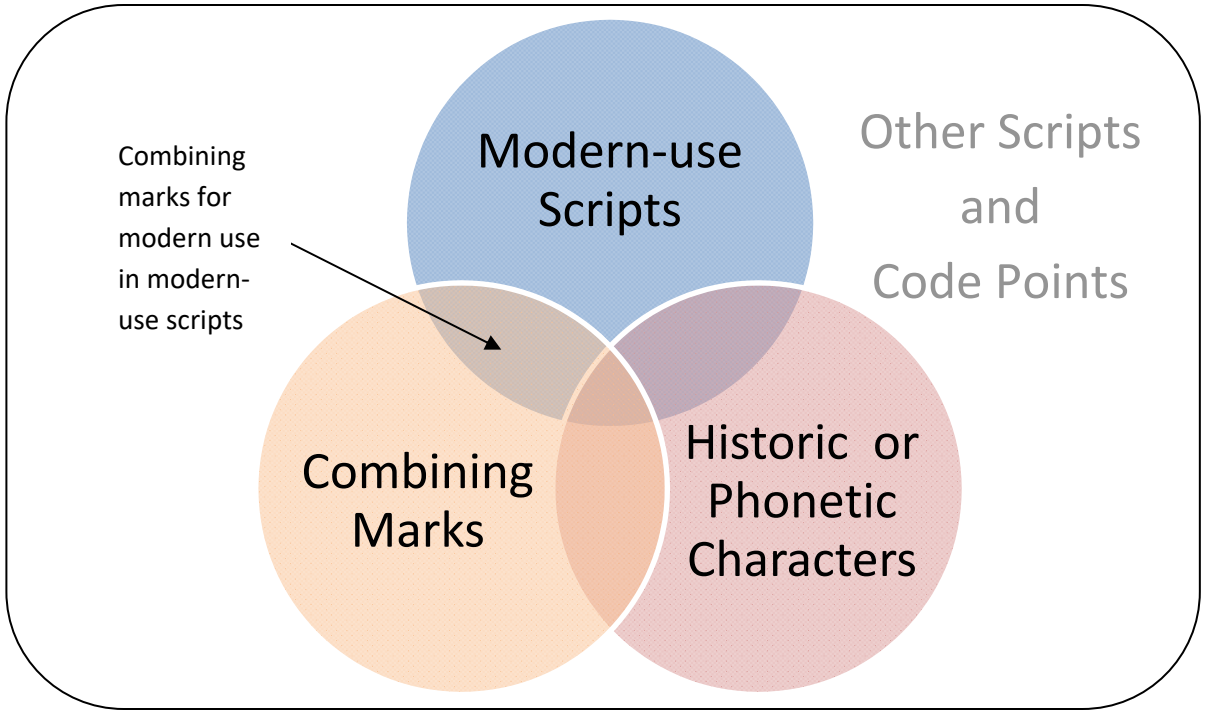
The Integration Panel is tasked to evaluate its actions in light of the Principles laid out in the [Procedure]. The methodology followed by the Integration Panel ensures that the Stability, Inclusion and Conservatism Principles may be fully applied to the final result (RZ-LGR). It recognizes that the MSR is merely an interim step in the development on the LGR, and that any code points included in it are not automatically added to the RZ-LGR; the MSR is only one of several constraints on the final LGR.

The expectation is that the Generation Panel will give these code points the benefit of very careful review and that they will be accompanied by a detailed rationale, should they be included in the LGR proposal. In turn the IP will apply these Principles when reviewing LGR proposals for integration.

2.3 Process of Deciding the MSR

The methodology followed by the Integration Panel started with a determination of scripts that are in "common widespread everyday use", based on the classification suggested in [UTS39] and refined further based on the relevance of each script to IDN labels applied for as part of the gTLD [NEWGTLD] and ccTLD [IDNFT] processes. These are called "modern-use" scripts in the following. Their selection is described in more detail in Section 3, below.

The set of code points for these scripts was taken from Unicode 11.00 [Unicode110], the latest published version of that standard for which IDNA 2008 tables [IDNAREG] are available at the time of defining MSR-5. These tables define a subset of Unicode code points that are deemed PVALID by applying the methodology of [RFC5892] to the repertoire of Unicode 11.0.0. This PVALID subset of modern-use scripts is indicated by the blue circle in the diagram.



The Integration Panel then created a list of code points to further exclude from the MSR, based on their status as unambiguously encoded for specialized purposes, such as characters used historically or in phonetic or other notations, but also based on some other problematic status values (not shown separately in the diagram).

Combining marks may be present in any script. Many are solely intended for historic or other specialized use (such as phonetic transcriptions), and are therefore excluded. The remaining ones were investigated for other problematic issues and whether they are used in everyday writing. Only some of those in the intersection between combining marks and modern use (indicated by the arrow) are thus included in the MSR.

Just because a script is in modern use does not mean that all its code points are in use for everyday modern writing. There are many code points encoded for historic or phonetic use. Only those which also have a modern use in everyday writing are retained in the MSR. They are part of the set shown by the intersection of the respective circles.

The MSR thus corresponds to a subset of code points identified as both belonging to modern-use scripts and as being in actual active modern use.

The following sections describe this process and discuss the resulting MSR in more detail.

3 Scripts

The root zone is to cater to significant modern use. The first step in deciding the Maximal Starting Repertoire is to consider (based on the Inclusion Principle) the scripts that should be supported.

3.1 Comprehensiveness and Staging

Ideally, the MSR would be comprehensive, that is, include all scripts eligible for the root zone from its first version. With respect to the *Stability Principle* and the *Least Astonishment Principle* a fully comprehensive MSR would guarantee that all issues relating to the possible interaction among all scripts can be fully investigated in the development of the RZ-LGR. From a practical perspective, the IP decided that doing so would be prohibitive because of the additional time needed to investigate certain scripts, and perhaps unnecessary for two main reasons. First, not all scripts are related closely enough so that they affect each other for the purpose of LGR development. Second, it is not realistic to expect that Generation Panels will be formed and complete their work for all eligible scripts within the same time frame.

Consequently, the Integration Panel accepts the pragmatic reality that both the MSR and the RZ-LGR will be rolled out in stages. Key to the success of such an approach is to ensure that all related scripts are always considered together, whether in defining the MSR or in creating the LGR.

For each stage of the work, the corresponding version of the MSR is immutable and represents a fixed set of code points. Generally, a given version of the LGR will be based on the most recently available MSR, but there is no lock-step process. For example, multiple versions of the LGR may be released without necessarily re-issuing the MSR, or vice versa.

The first version of the MSR [MSR-1] deferred some scripts, to balance timeliness with comprehensiveness (see Section **Error! Reference source not found.**). MSR-2 included the bulk of these deferred scripts. MSR-3 added six code points to two existing scripts. MSR-4 added fifteen code points to two existing scripts. MSR-5 adds 4 code points to 3 scripts. MSR-5 is the foundation for any subsequent LGR. Future versions of the MSR will cater to additional repertoires that become eligible for the root zone, whether due to additions in the Unicode repertoire for scripts already included in the MSR, or the adoption of existing or newly encoded scripts for modern use. (See also Section 3.9).

All future versions of the MSR and all versions of the RZ-LGR must retain full backwards compatibility, so that they preserve the output of any label registration against the old LGR, when applied to an updated LGR or an LGR resulting from a later version of the MSR. For the MSR, any part of the repertoire not already used as basis for any script LGR is not required to be retained in future versions. Likewise, for the RZ-LGR, any repertoire that has not yet been used for label registration is not required to be retained in future versions. Nevertheless, the IP anticipates that succeeding versions of the MSR or the RZ-LGR will be strict supersets of their respective predecessors.

It is expected that registrations that predate the initial release of an LGR covering the respective script will be allowed to remain, even if in conflict, but without becoming a binding precedent for the LGR itself.

3.2 What Defines a Related Script?

Historical derivation is one element that the IP evaluated carefully. For example, all the alphabetic scripts of Europe and the Middle East are derived ultimately from the Phoenician alphabet. But this fact by itself is not relevant in designing label generation rules; rather there must be sufficient common features among scripts for there to be confusion among code points, or the structural properties of the way code points of the script combine must pose common problems to a digital presentation system. (Hebrew and Arabic share bidirectional features, so to the extent that the LGR is sensitive to bidirectional issues, the Integration Panel would need to make sure any solution fits both of these scripts. Likewise, the Brahmi scripts of India, Bangladesh, Nepal and Sri Lanka follow comparable rules of consonant-vowel combination in rendering.)

In addition to questions of script relation, the Integration Panel also considered the **encoding model**, because it is the digital representation of a script that is of interest. Thai and Lao are both encoded on a model based on the TIS (Thai Industrial Standard). Khmer, for example, would not be. It differs in encoding model from the other two. Similarly, the Ethiopic script is not considered related to the logically-similar Indian scripts, since Unicode has adopted a completely different encoding model for it.

The existence of a large **set of confusables**[4] between two scripts is something to which the Integration Panel paid close attention. Such confusability is a good indication that the Integration Panel will want to consider two scripts at the same time: they are not "separable" and must be both-or-neither in the MSR and the LGR.

3.3 Separable Scripts

Separable scripts are not mutually related. As a consequence, they can be considered independently, both for inclusion in the MSR, but also for purposes of integration. LGRs for separable scripts can be reviewed independently by the Integration Panel without the risk from possible interactions. The same is not true for related scripts.

3.4 Deferred Scripts

MSR-1 deferred a number of scripts to a later version of the MSR, despite being considered eligible in principle for the Root Zone by the IP. This was done primarily in the interest of managing the work and to allow a timely release of an initial MSR (see Section3.8.2). All previously deferred scripts are included in the MSR starting with MSR-2. There are currently no scripts identified as “deferred”.

3.5 Historical and Obsolete Scripts

With the first release of the MSR, the Integration Panel — confirming the position in the [Procedure] — took the decision to simplify the issues rather radically by designating all **historical and obsolete** scripts ineligible from ever being in the root. In other words, certain scripts, by virtue of their restricted sphere

[4] These can take the form of outright *homoglyphs*, discussed below, or resemble the type of close and systematic visual relation between, for example, certain Devanagari and Gurmukhi code points, e.g.:
Devanagari 0935 व (VA), 0909 उ (U), 092F य (YA), 093F ि (VOWEL SIGN I), 0940 ी (VOWEL SIGN II)
vs
Gurmukhi 0A15 ਕ (KA), 0A24 ਤ (TA), 0A27 ਧ (DHA), 0A3F ਿ (VOWEL SIGN I), 0A40 ੀ (VOWEL SIGN II)

of use, are removed from consideration: they are not just "deferred" but "ineligible" by definition. This has the consequence that the decision to remove them is effectively **permanent**, and is not subject to later change by adding these scripts in the ordinary course of updating the MSR. Allowing their status to change in this way would give rise to precisely the risks addressed in the [Procedure]. The unforeseen and rather unlikely event of a major change in some community leading to a full revival of the affected script might offer the sole possible exception (see also Section 3.10).

On the other hand, there are a number of scripts that, while they are currently considered ineligible for the root based on limited use, cannot be ruled out in permanence. For some, there is active promotion of their use in the community and consequently, their status may change in the future. The IP will continue to monitor developments for these scripts (see Section 3.9).

3.6 Selecting Scripts and Code Points for the MSR

The [Procedure] gives the general prescription for developing the MSR:

> *"The maximal set of code points for the root zone is itself a subset of Unicode created by the Integration Panel via an application of IDNA2008 and the principles in IABCP. Further, it does not include code points defined as restricted for identifiers, as specified in Table 1 of UTS#39. [Section B.3.4.1]"*

The starting point for MSR-5 was the intersection between the latest published version of the Unicode Standards, Unicode 11.0, and the [IDNA 2008] PVALID set derived [IANAREG] from the prescriptions of [RFC5892].[5] Therefore, all code points newly added to Unicode 12.0 through Unicode 13.0 are not included in this intersection for MSR-5, but deferred for future consideration.

In further analyzing the set resulting from this intersection, the additional data tables provided by the Unicode Consortium for Unicode Technical Standard #46 [UTS46] and Unicode Technical Standard #39 [UTS39] for Unicode11.0 were helpful inputs, as well as the tables in Unicode Standard Annex #31 [UAX31]. MSR-5 provides full coverage of eligible modern-use scripts (see discussion below).

3.7 Scripts Appropriate for Use in Identifiers

Section B.5.3.2 of the [Procedure] states:

> *"In section 3.1, Unicode Technical Standard#39 "Unicode Security Mechanisms" [UTS39] includes a mechanism for evaluating Assigned Code Points to determine whether they are appropriate for use in identifiers. This determination is based in part on whether a code point is part of a script not used for writing a living language, or a script that is of limited use, or otherwise not yet widely used, as defined in UAX#31 "Unicode Identifier and Pattern Syntax", Tables 4 through 7 [UAX31]. "*

[5] At the time of writing, the latest available set of tables in the IANA registry [IANAREG]is based on Unicode 11.0.

and gives additional guidance to the Integration Panel on using its judgment in fine-tuning this determination. In [UAX31] the Unicode Consortium provides a classification of scripts, which provided the starting point.[6]

3.8 Modern Use Scripts

Table 5 in [UAX31] lists those scripts that are identified as being recommended for support in identifiers because they are "in widespread modern customary use, or ... regional scripts in modern customary use by large communities."

The IP has identified Bopomofo as a modern use script that is limited to educational use and therefore not eligible for the root zone. There are also two collections of code points that are used with multiple scripts. The reasons for making these categorizations are discussed below.

The MSR now covers all of the scripts in this table except Bopomofo, as indicated.

UAX#31, Table 5. Recommended Scripts[7]

Script ID	Description	Remarks
Zyyy	Common	*Shared across scripts*
Zinh	Inherited	*Shared across scripts*
Arab	Arabic	MSR-1, extended in MSR-5
Armn	Armenian	Added in MSR-2
Beng	Bengali[8]	MSR-1
Bopo	Bopomofo	*Educational use only*
Cyrl	Cyrillic	MSR-1
Deva	Devanagari	MSR-1, extended in MSR-5
Ethi	Ethiopic	Added in MSR-2
Geor	Georgian	MSR-1
Grek	Greek	MSR-1
Gujr	Gujarati	MSR-1
Guru	Gurmukhi	MSR-1
Hani	Han	MSR-1, extended in MSR-3
Hang	Hangul	MSR-1
Hebr	Hebrew	MSR-1
Hira	Hiragana	MSR-1
Knda	Kannada	MSR-1
Kana	Katakana	MSR-1
Khmr	Khmer	Added in MSR-2
Laoo	Lao	MSR-1

[6] The Unicode Consortium has revised its classification of scripts somewhat in recent editions of UAX#31. However, the set of “Recommended” scripts is unaffected.

[7] The left column in the original tables from [UAX#31] indicates the script membership using a format suitable for regular expressions. The reproduction here presents the corresponding Script ID value directly. Table 5 has been augmented by a remarks column not found in the original.

[8] This script is commonly known as Bangla. To some users it is known as Assamese.

Latn	Latin	MSR-1, extended in MSR-3through MSR-5
Mlym	Malayalam	MSR-1
Mymr	Myanmar	Added in MSR-2, extended in MSR-4
Orya	Oriya	MSR-1
Sinh	Sinhala	MSR-1
Taml	Tamil	MSR-1
Telu	Telugu	MSR-1
Thaa	Thaana	Added in MSR-2
Thai	Thai	MSR-1
Tibt	Tibetan	Added in MSR-2

The following subsections discuss the disposition of these scripts in terms of the MSR.

3.8.1 Common and Inherited

These two script categories are for code points that are shared among scripts or are among the combining marks that have not been given a specific script assignment in the Unicode Standard (whether or not they are actually used with more than one script). Some code points from these two script categories will need to be included in the LGR, but they require careful scrutiny, in certain cases by more than one of the GPs.

3.8.2 Scripts included in MSR-1

The set of scripts included in MSR-1 was based on the set of scripts for which applications had been received as part of the IDN ccTLD Fast Track Process [IDNFT] and New gTLD Program [NEWGTLD], as well as some additional scripts that are strongly related to one or more of these scripts. These scripts are identified as “MSR-1” in the table above.

3.8.3 Scripts added in MSR-2

In contrast, some scripts are not strongly related, but instead separable. Separable scripts may be considered in isolation for purpose of creating the MSR because the LGRs proposed for them are not expected to interact with LGRs for other scripts. Some of these separable scripts were known to be eligible in principle for the Root Zone, but were deferred at the time of creating MSR-1. They are included in all versions of the MSR from MSR-2, thus completing the coverage of eligible modern-use scripts.

These additional scripts are:

- Armenian,
- Ethiopic,
- Khmer,
- Myanmar,
- Thaana and
- Tibetan.

3.8.4 Scripts added in MSR-3 through MSR-5

None.

3.8.5 Modern Scripts Ineligible for the Root Zone

Based on its usage, the Integration Panel decided that Bopomofo is not of interest for IDN TLDs and it is not included in the MSR. The Integration Panel deems it unlikely that it will be part of any future MSR releases.

3.9 Scripts for Possible Future MSRs

The scripts listed in [UAX31], Table 7, "Limited Use Scripts" (as reproduced below) are not candidates for the MSR at this point. However, the Integration Panel adopts a neutral attitude with regard to reviewing them for inclusion in a future update of the MSR. These scripts are not closely related to any of the modern use scripts already included in the MSR.

UAX#31 - Table 7. Limited Use Scripts[9]

Script ID	Description
Adlm	Adlam
Bali	Balinese
Bamu	Bamum
Batk	Batak
Cakm	Chakma
Cans	Canadian Aboriginal
Cham	Cham
Cher	Cherokee
Java	Javanese
Kali	Kayah Li
Lana	Tai Tham
Lepc	Lepcha
Limb	Limbu
Lisu	Lisu
Mand	Mandaic
Plrd	Miao
Mong	Mongolian
Mtei	Meetei Mayek
Newa	Newa
Nkoo	Nko
Olck	Ol Chiki
Osge	Osage
Rohg	Hanifi Rohingya
Saur	Saurashtra
Sund	Sundanese
Sylo	Syloti Nagri
Syrc	Syriac

[9] This document deliberately reflects the status as of Unicode 11.0.0.

Tale	Tai Le
Talu	New Tai Lue
Tavt	Tai Viet
Tfng	Tifinagh
Vaii	Vai
Wcho	Wancho
Yiii	Yi

3.10 Scripts Identified in UAX#31 as Not Suitable for identifiers

[UAX31], Table 4, ”Excluded Scripts” lists all scripts in Unicode 11.0 not recommended for identifies nor in limited use. These are a mixed bag of ancient, recently obsolete, educational, and apparently declining scripts, including any scripts with encoding models that would make them otherwise unsuitable.

In agreement with suggested treatment of these scripts in the [Procedure], the Integration Panel confirms that none of these scripts should be eligible for IDN TLDs now or in the foreseeable future. It is arguable that, in response to substantial and sustained shifts in script use by the relevant communities, the status of one or the other of these scripts might at some later time need to be changed to Limited Use. Barring that eventuality, these scripts are permanently excluded from the root zone.

UAX#31 Table 4. Excluded Scripts

Script ID	Description
Aghb	Caucasian Albanian
Ahom	Ahom
Armi	Imperial Aramaic
Avst	Avestan
Bass	Bassa Vah
Bhks	Bhaiksuki
Brah	Brahmi
Bugi	Buginese
Buhd	Buhid
Cari	Carian
Chrs	Chorasmian
Copt	Coptic
Cprt	Cypriot
Diak	Dives Akuru
Dogr	Dogra
Dsrt	Deseret
Dupl	Duployan
Egyp	Egyptian Hieroglyphs
Elba	Elbasan
Elym	Elymaic
Glag	Glagolitic
Gong	Gunjala Gondi

Gonm	Masaram Gondi
Goth	Gothic
Gran	Grantha
Hano	Hanunoo
Hatr	Hatran
Hluw	Anatolian Hieroglyphs
Hmng	Pahawh Hmong
Hung	Old Hungarian
Ital	Old Italic
Khar	Kharoshthi
Khoj	Khojki
Kits	Khitan Small Script
Kthi	Kaithi
Lina	Linear A
Linb	Linear B
Lyci	Lycian
Lydi	Lydian
Maka	Makasar
Mahj	Mahajani
Mani	Manichaean
Marc	Marchen
Medf	Medefaidrin
Mend	Mende Kikakui
Mero	Meroitic Hieroglyphs
Merc	Meroitic Cursive
Modi	Modi
Mong	Mongolian
Mroo	Mro
Mult	Multani
Narb	Old North Arabian
Nbat	Nabataean
Nand	Nandinagari
Nshu	Nushu
Ogam	Ogham
Orkh	Old Turkic
Osma	Osmanya
Palm	Palmyrene
Pauc	Pau Cin Hau
Perm	Old Permic
Phag	Phags-pa
Phlp	Psalter Pahlavi
Phli	Inscriptional Pahlavi
Phnx	Phoenician
Prti	Inscriptional Parthian
Rjng	Rejang
Runr	Runic
Samr	Samaritan
Sarb	Old South Arabian

Sgnw	SignWriting
Shrd	Sharada
Shaw	Shavian
Sidd	Siddham
Sind	Khudawadi
Sogd	Sogdian
Sogo	Old Sogdian
Sora	Sora Sompeng
Soyo	Soyombo
Tagb	Tagbanwa
Tglg	Tagalog
Tang	Tangut
Takr	Takri
Tirh	Tirhuta
Ugar	Ugaritic
Wara	Warang Citi
Xpeo	Old Persian
Xsux	Cuneiform
Yezi	Yezidi
Zanb	Zanabazar Square

4 Exclusions of Individual Code Points or Ranges

4.1 Historic and Phonetic Extensions to Modern Scripts

As the universal character set, Unicode caters not only to modern, everyday use, but also to the scholarly use of scripts, including code points for historic and other obsolete letters as well as extensions for phonetic use. The Integration Panel feels that for a subset of these, the special nature of these code points is well-established enough to warrant their a-priori exclusion from the MSR. This reflects in almost all cases the explicit and documented decision by the Unicode Consortium to encode them for these specialized purposes. It thus helps the Generation Panels focus on modern use.

In deciding on exclusion based on historic, phonetic, or other limited use, the Integration Panel followed the principle discussed above of allowing code points to remain in the MSR any time their status could not be definitely confirmed. In other words, if there is a possibility that some code point also occurs in everyday modern use, perhaps for a significant minority language, it was not filtered out from the MSR, but left to evaluation by the Generation Panels. The Integration Panel requires that any justification for the inclusion of such a code point in the LGR would meet the highest standards.[10]In the following sections, we discuss specific code points excluded from the MSR on these grounds.

[10] For MSR-3, three previously excluded phonetic characters from the Latin script were added to the MSR after further review indicated their possible use in general-use orthographies.

4.2 Code Points That Pose Special Risks

There are a number of code points that pose a special risk to the DNS and implementations, whether due to confusability with ASCII punctuation, instability of encoding, or other reasons. Such code points must be excluded from the LGR, and where these issues can be discovered ahead of time, are best excluded already from the MSR.

4.3 Code Points with Strong Justification to Exclude

There are a number of code points for which there are strong reasons to exclude them from the Maximal Starting Repertoire a priori.

A code point assigned to a character which has any of the following characteristics is excluded from the MSR. The character is:

- archaic, historic, symbolic, and has little chance to gain use in modern context,
- PVALID as unintended consequence of the IDNA2008 algorithm[11],
- highly confusable with an existing and common punctuation character[12]
- exclusively used for phonetic, liturgical or other specialized purposes,

Historic usage in writing systems of India primarily includes special characters for Sanskrit and Vedic. The Integration Panel felt that code points only used in that context should be excluded as should code points assigned to characters for liturgical use, such as Hebrew cantillation marks or Arabic Koranic annotations.

In the Latin and Cyrillic scripts there are dozens of characters that have been encoded to support one or more systems of phonetic notation. Where it could be established unambiguously that a code point was encoded solely for that purpose, the Integration Panel decided to exclude it from the MSR.

4.4 Code Points That May or May Not be Excludable from the Root Zone LGR

The following factors indicate that a code point should not be included in the RZ-LGR unless a very careful analysis determines otherwise. In order to facilitate such an analysis, the Integration Panel has chosen to leave these code points in the MSR but to require very strong supporting justification from the Generation Panels in order for any of the code points to appear in the LGR.

This applies to any code point assigned to a character that shows any of the following characteristics. The character is:

- currently obsolete, but with some probability of future or near term re-use in modern context

[11] A good example is 101FD PHAISTOS DISC SIGN COMBINING OBLIQUE STROKE, a code point that is part of an undeciphered script and that is PVALID only because it is formally a combining mark (thus formally of script Zinh).

[12] Most code points that fall within Unicode General Category 'P', the union of {Pc, Pd, Pe, Pf, Pi, Po, Ps} will be excluded, but the procedure notes that General Category does not perfectly align with the distinction desired by the Letter Principle. For example, the Tibetan *intersyllabic tsheg* (0F0B) has General Category of Po, but when it occurs medially, it is an obligatory mark of word-internal punctuation.

- used in a minority writing system, but whether it is required in the context of LGR needs more study
- primarily intended for historic use (such as Sanskrit), but whether or not it also has modern use needs more analysis
- primarily intended for a specialized use such as liturgical or phonetic, but more analysis needs to be done on whether it has generalized use that needs to be supported
- one that presents a compatibility issue with respect to IDNA2003

The Integration Panel relied on expert advice and public sources, such as the Unicode Standard and the Document Register of the Unicode Technical Committee, in establishing the intent behind a character assignment. Where a probability of dual use in modern writing could not be excluded, the code points were generally retained in the MSR.

For the Latin and Cyrillic scripts, in particular, the Integration Panel has opted to include those code points in the MSR that, while ostensibly encoded to support one or more systems of phonetic notation, might represent cases of actual or potential dual-use in the context of modern writing systems. Doing so enables the Generation Panels to review whether the code point is in fact required for the RZ-LGR, and if so to document the facts confirming that assessment.

4.5 Non-spacing Combining Marks

Many non-spacing combining marks (sometimes called *diacritics*) for writing systems that make heavy use of pre-composed forms have been excluded from the MSR. This concerns mostly Latin, Greek and Cyrillic orthographies. While many have well-established specialized purposes, such as for Polytonic Greek (a historic orthography) or phonetic transcriptions, it has not been possible for other combining marks to exclude the possibility that they see some use in modern orthographies (especially in African writing systems). Whenever that is the case, the proper venue for further evaluation of these code points would be the relevant Generation Panel; in such case, the Integration Panel has retained the code points in the MSR.

The Integration panel excluded code points assigned to combining marks that are *exclusively* intended for:

- specialized use for medieval and other transcriptions,
- phonetic use, or that are
- otherwise inappropriate for the DNS root zone.

Many combining marks are used in sequences that can be normalized to precomposed characters (combinations of base character and diacritic assigned a single code point). These tend to be the most common, most widely used, and most productively used combining marks; therefore, they can be expected to occur also in novel combinations that would require an explicit combining mark even in Normalization Form C (NFC). Because such combinations do not require a listing in the Unicode Standard before being usable, it is not possible to rule out, or limit, their applicability to living orthographies

without further evaluation by the Generation Panels. These code points for combining marks have been included in the MSR.

The Integration Panel expects that a Generation Panel can provide evidence of at least one combination not corresponding to a precomposed character for each combining mark it intends to include in the LGR (alternatively, document that the mark is an independent letter, for example a vowel sign, and ordinarily required).

In addition, any direct evidence that a mark may be needed for everyday writing despite it never being used in a precomposed character has been taken as a reason to include the mark in the MSR; one example is the *vertical bar below* for Yoruba (U+0329). The aim was to cover the productive marks, as well as any marks attested for orthographic use, so that the MSR does not accidentally exclude any actual code point (or combining sequence) needed for some orthography eligible to be supported in the root zone.

The actual set of combining marks allowable in the LGR will be smaller than the set included in the MSR, because it will be limited to those marks that are actually required for at least one combining sequence not expressible in NFC. In addition, where the number of such attested sequences is known and limited, GPs are encouraged to enumerate the sequences where feasible, rather than adding the "bare" combining mark to the repertoire. This would serve to prevent such marks from combining with every other allowed code point in the GP's repertoire.

At the same time, GPs are encouraged to consider variant relations based on NFD. That is, if two combining marks are variants of each other, then they should be made blocking variants of each other even in precomposed characters. For example, based on interchangeable and inconsistent use a strong argument can be made that the *comma below* and the *cedilla* are combining marks that are not merely similar in shape, but rather are frequently (if inadvertently) treated as "variants" of each other[13]; if treated as variants, S with cedilla below and S with comma below would then block each other. (Another diacritic that may be substituted or effectively indistinguishable from either comma below or cedilla is *dot below*[14]).

Because all the labels allowed by IDNA2008 are precomposed using Normalization Form C, all precomposed characters containing either of two combining marks that are variants of each other should also be blocked variants of each other (to the extent that the base characters agree as well). Because of the fact that the LGR will be stated in NFC, this would require more entries in the table, but the Conservatism Principle would argue against allowing these to remain unrestricted, therefore in favor of explicitly listing the cases as blocked variants.

[13] The letters S or T with comma below were not encoded separately from S and T with cedilla in the early character encodings or versions of Unicode before version 3.0. According to [Latvian], comma below remains an acceptable rendering for Latvian letters with cedilla.

[14] In some writing systems, users may substitute dot below for cedilla below, if only as a common fallback. (See [Marshallese]).

The MSR includes U+0307 COMBINING DOT ABOVE. If this code point is included in an LGR, special care must be taken to prevent it from being used in conjunction with “soft-dotted” or “dot-less” code points, like “I” and “j” or “dotless-I” and “dotless-j”, as the these combinations would be indistinguishable from the ordinary characters “I” and “j”.[15]

5 Discussion of Particular Code Points

The following sections give detailed rationale for excluding certain code points (and in a few cases for not excluding them). The listings of code points in each subsection are not exhaustive unless so indicated; for full listing see the annotated code table file.

In the MSR non-CJK code tables, excluded code points are annotated as follows:

Main annotation	Additional detail, as appropriate
Obsolete	historic, archaic
Limited or declining use	educational, threatened, nearly extinct
Symbol	characters classified as letters that are symbolic in nature
Numeric	characters used in numerical context
Punctuation	characters classified as letters that look like punctuation
CONTEXTJ	context - join controls
CONTEXTO	context - others
Unstable	encoding model changed
Deprecated	no longer in use, alternate code preferred
Technical use	phonetic, poetry
Religious use	annotation, cantillation
Homoglyph	digraph of x y
Deferred repertoire	12.0 and 13.0 repertoire not yet included in IANA tables

In making the determinations described in this section, the Integration Panel has relied on a number of sources, including expertise of both panel members and external advisors. Written sources consulted include the RFCs listed in the References, the descriptions of script use in [Unicode 6.3] and later versions, as well as individual documents submitted during the character encoding process [UTC].

5.1 Digits and Hyphen

All digits and Hyphen are excluded from the MSR. This follows from the Letter Principle in the [Procedure], which states that only code points exclusively used in writing words are to be included in root zone labels. There are the occasional orthographies that use digits and punctuation as part of words. Where these code points do not occur *exclusively* inside words, they are prohibited by the Letter Principle. This is in contrast to code points that are considered letters or ideographs but may also be

[15] In this context, GPs should note that some fonts may omit the dot on the letter ‘i’, making it indistinguishable from dotless-i. While not common, the existence of such fonts implies that users of languages not using a dotless-I will readily accept it as substitution for the dotted ‘I’.

used to express numbers, such as ideograph U+4E00 which also signifies “one” or use of Latin letters that may be used as roman numerals.

[TBD]The MSR identifies and documents all code points excluded solely due to the letter principle. For some scripts, there exist native digit sets that have fallen out of use in favor of ASCII digits. Such digits would also be excludable as “historic” or “obsolete” and therefore would not be considered as being excluded solely due to the letter principle.

5.2 CONTEXT O Code Points

All remaining code points requiring a CONTEXTO rule in IDNA2008 are excluded from the MSR.

- U+00B7 MIDDLE DOT
- U+0375 GREEK LOWER NUMERAL SIGN (keraia)
- U+05F3 HEBREW PUNCTUATION GERESH
- U+05F4 HEBREW PUNCTUATION GERSHAYIM
- U+30FB KATAKANA MIDDLE DOT

5.3 CONTEXT J Code Points

All code points requiring a CONTEXTJ rule in IDNA2008 are excluded from the MSR.

- U+200C ZERO WIDTH NON-JOINER
- U+200D ZERO WIDTH JOINER

5.4 Code Points Restricted for Identifiers

Code Points listed in Table 1 of [UTS39] as restricted from use in identifiers are the final set of code points explicitly mandated for exclusion in the [Procedure]. This set overlaps some of the sets already described. Several subsets of these are unambiguously identified.

A very small number of code points in the Unicode Standard are formally deprecated. This means that, their use is in all instances strongly discouraged, often in favor of an alternate code point or code point sequence. For example:

- U+0673 ARABIC LETTER ALEF WITH WAVY HAMZA BELOW
- U+17A3 KHMER INDEPENDENT VOWEL QAQU+17A4 KMHER INDEPENDENT VOWEL QAA

The Unicode property XID_CONTINUE is false for some characters that are PVALID and not already listed above:

- U+06FD ARABIC SIGN SINDHI AMPERSAND
- U+06FE ARABIC SIGN SINDHI POSTPOSITION MEN
- U+2E2F VERTICAL TILDE

The other subsets of code points listed in Table 1 of [UTS39] as restricted for use in identifiers either never apply to code points that are PVALID in IDNA 2008, or are subject to judgment calls and their disposition for the purpose of the MSR is covered in the following discussion.

5.5 Compatibility with IDNA2003

In IDNA2003, case folding is applied which creates compatibility issues between IDNA2008 and IDNA2003 for several code points. This arguably makes the affected code points candidates for summary exclusion from the MSR on grounds of Longevity (§2.1).

Case folding removes the following two code points from the final string and replaces them with other code points; however, both code points are reasonably frequent in their respective orthographies.

- U+00DF LATIN SMALL LETTER SHARP S
- U+03C2 GREEK SMALL LETTER FINAL SIGMA

In Greek, the final sigma (ς) would be the form normally chosen whenever sigma (σ) ends a word. However, in cases of labels based on compound words, the final sigma may also occur in mid-label, and in cases of abbreviations, the regular sigma may also occur at the end. The Greek Issues Report [GreekVIP] contains a discussion on the final sigma but suggests that it should be allowed in the LGR.

In German, while an "ss" is often an acceptable fall-back for the *sharp s* (ß)[16], German orthography has changed such that there is now a distinct difference between “ß” and “ss” in pronunciation of the preceding vowel. Second level domains exist that have moved to supporting *sharp s* without restrictions and that treat both *sharp s* and “ss” as unrelated for purposes of delegation. The Latin Issues Report [LatinVIP] does not address this code point.

The Integration Panel admits both of these code points to the MSR, with the purpose of allowing the respective Generation Panels to perform an in-depth review and to propose a way to handle them in the LGR that best balances community requirements and DNS stability, usability and security.

For the final sigma, it would be possible to define a "when" rule in the LGR that would disallow this code point except in the final position in a label. The function of such a rule would be to limit delegations, but it would not account for the exceptional placements of sigma and final sigma. It would also be possible to define sigma and final sigma to be (blocked) variants of each other. The Integration Panel encourages the Greek Generation Panel to study the issue.

Likewise, the Integration Panel expects the Latin Generation Panel to carefully review the feasibility and risks of supporting the *sharp s* in the LGR and, if it should consider the inclusion of this code point in the LGR, to investigate the case for or against making it a blocked variant of "ss".

Another code point with IDNA 2003 compatibility issues is assigned to the character

- 0131 LATIN SMALL LETTER DOTLESS I

[16] Because both uppercase to the same sequence “SS”, they also resolved the same under IDNA2003. The issue is further complicated by the recent adoption of the uppercase letter “ß” as a formally permitted alternative to “SS”. Any variant defined for sequence “ss” would *overlap* with variants defined for “s”. See discussion in RFC 8228.

This character has case mappings that are locale sensitive and thus were an issue for IDNA 2003. In IDNA 2008 there is no case mapping. The Integration Panel expects the Latin Generation Panel to investigate the need to address any compatibility issues related to this code point, and if found, suggest means to mitigate them.

5.6 Code Points for Which the Encoding or Usage May be Unstable

5.6.1 Unified Ideograph-4CA4

This code point represents an incorrect unification of two ideographs with different radicals:

To remedy this error, a future version of Unicode will likely introduce a new, disunified code point.

This renders the encoding of U+4CA4 unstable. As a result, it cannot be included in the MSR. This may not be a transient issue, because legacy implementations can be expected to exist for a considerable time, making the use of U+4AC4 (and its proposed disunified counter-part) too risky for use in the root.

5.6.2 Candrabindu

Unicode 7.0 adds 3 code points named *candrabindu* to Indic scripts:

- U+0C00 TELUGU SIGN COMBINING CANDRABINDU ABOVE
- U+0C81 KANNADA SIGN CANDRABINDU
- U+0D01 MALAYALAM SIGN CANDRABINDU

Some characters in Indic script have a history of disunification: after a period where the Devanagari block code point was assumed to be shared among scripts, separate script-specific code points were added for some scripts. Such disunification would render the encoding unstable.

Based on information received [MSRGupta], this is not the case for these additions of *candrabindu* code points; there is no evidence of shared script use of existing *candrabindu* code points, and such use would not be supported in existing rendering engines. As a consequence, the additions do not prevent the inclusion of the existing code points

- U+0901 DEVANAGARI SIGN CANDRABINDU
- U+0981 BENGALI SIGN CANDRABINDU
- U+0A81 GUJARATI SIGN CANDRABINDU
- U+0B01 ORIYA SIGN CANDRABINDU

in the MSR.

A fully consistent treatment of *candrabindu* characters across the scripts, unfortunately, will not be possible in the first phase of the RZ-LGR work, because the additions remain outside the Unicode 6.3 cutoff for MSR-1, MSR-2, MSR-3 and MSR-4. The Integration Panel requested the affected Generation Panel to consider such newly added code points while performing its analysis to reduce the risk of any later compatibility issues.

All indications are that the *candrabindu* characters are not equally needed for everyday modern usage in their respective scripts and languages. Based on the information received, the Integration Panel did not add these to MSR-5 when the base version of Unicode was moved to Unicode 11.0. The Integration panel expects that any future inclusion in an updated LGR be accompanied by a detailed justification.

5.7 Confusability and Homoglyphs

The Integration Panel, for the purpose of creating the MSR, generally did not consider confusability between code points that otherwise qualify for inclusion. The expectation is that each Generation Panel will apply particular scrutiny in such cases and will propose whether such cases should be handled by defining blocked variants, by not including code points in the LGR or by relying on standard processes outside the LGR to address the issue.

Homoglyphs are characters which are of essentially identical appearance by design, instead of merely similar appearance. In many cases, homoglyphs arise because Unicode assigned a duplicate code point to the “same” character, based on different use, or to avoid having to give a single character membership in multiple scripts. Often the reason that characters may be homoglyphs is because of historical derivation from the same source or because of having been adopted (borrowed) from another script. Where confusability is based on homoglyphs, the Integration Panel makes a distinction between homoglyphs of PVALID code points and homoglyphs of code points that are not PVALID in IDNA 2008.

The expectation of the Integration Panel is that homoglyphs of single PVALID code points will be addressed in each Generation Panel's LGR proposal, and further, that the LGR will exclude homoglyphs from the repertoire or define them as blocked variants, unless the Generation Panel can provide an acceptable justification for a different treatment. In contrast, the Integration Panel has excluded homoglyphs of code points that are not PVALID because it considers such homoglyphs ineligible for the LGR. This applies in particular to homoglyphs of ASCII Punctuation.

5.7.1 Cross-script Homoglyphs

There are a number of homoglyphs of code points that cross scripts. These occur, for example, between Latin and Cyrillic, or Latin and Greek, or Cyrillic and Greek. There are no obstacles to defining blocked variants across script boundaries in the Integrated RZ-LGR. The Integration Panel does not expect that cross-script homoglyphs would ever become allocatable variants, because that would imply mixed-script repertoires.

Because simple glyph shapes like this give effectively no hint of script identity, the IP encourages the Generation Panels to consider cross-script variants in such cases even for otherwise unrelated scripts.[17]

Examples involving a “circle glyph” include

- U+006F LATIN SMALL LETTER O
- U+03BF GREEK SMALL LETTER OMICRON
- U+043E CYRILLIC SMALL LETTER O
- U+0585 ARMENIAN SMALL LETTER OH
- U+0B20 ORIYA LETTER TTHA
- U+0D20 MALAYALAM LETTER TTHA
- U+101D MYANMAR LETTER WA

Among related scripts, there may be pairs of code points that are identical or nearly identical despite having more complex shapes. Where these can be used to form a label that is a homograph of a label in another script, they should be investigated for variant status.

5.7.2 Script-internal Homoglyphs

The Arabic script has extensive sets of in-script homoglyphs, depending on position in the word. For example, the following two code points have identical glyphs if in initial form.

- U+0643 ARABIC LETTER KAF
- U+06A9 ARABIC LETTER KEHEH

These two characters have identical appearance if at the end of a word.

- U+0647 ARABIC LETTER HEH
- U+06D5 ARABIC LETTER AE

Additional detail can be found in [Arabic VIP]. The Integration Panel has not addressed these positional homoglyphs in the MSR and expects the Generation Panel to conduct an in-depth analysis of the issues presented by these homoglyphs and further expects that any proposal to include them in the RZ-LGR will be accompanied by a suitable proposal on how to mitigate their effect.

5.7.3 Digraphs

Digraphs are a subset of script-internal homoglyphs, where a code point codes for an entity that can otherwise be rendered by a sequence of two code points. (In contrast to a combining sequence, the term digraph usually refers to a sequence of two base letters).

There are digraphs used in Yiddish writing systems using the Hebrew script that are homoglyphs of sequences of ordinary Hebrew characters and therefore indistinguishable no matter the font used. See the discussion in Section 5.16.

[17] Note that there already is a delegated ASCII-TLD “.ooo”.

Digraph code points also exist in the Latin script. However, all instances identified so far are for limited use and therefore excluded from the MSR on that basis already.

The Tibetan script contains digraphs of vowel signs. While vowel signs are combining marks in Unicode, the code points in question are homoglyphs of a sequence of other vowel signs. They do not have a decomposition in Unicode, therefore both the sequence and the digraph version are independently PVALID.

- U+0F7B TIBETAN VOWEL SIGN EE
- U+0F7D TIBETAN VOWEL SING OO

The former is a digraph of <U+0F7A, U+0F7A>, that is, a sequence of two instances of Tibetan vowel sign E and the latter a digraph of <U+0F7C, U+0F7C>, that is, a sequence of two instances of Tibetan vowel sign O. Thus, the digraph code points match the result double application of these vowels signs.

5.7.4 Script-internal Near Homoglyphs (ASCII Lookalikes)

The Integration Panel is concerned with the potential risk associated with code points that are nearly indistinguishable from their ASCII counter parts. For example, in any font not employing a glyph with a "handle" for the letter "a", the code point

- U+0251 ɑ LATIN SMALL LETTER ALPHA

is (for all practical purposes) indistinguishable from U+0061 a LATIN SMALL LETTER A unless presented side-by-side. (The character U+0251 is used by Fe'fe'e in Cameroon and the African Reference Alphabet). The Integration Panel believes that it would be premature to eliminate this and similar characters from the MSR. Instead, it is anticipated that the Latin Generation Panel, the Integration Panel, and expert advisors will engage in a dialogue aimed at a thorough analysis of this issue, to come to a resolution whether this and other code points presenting the same issues should be included in the LGR, and if so, whether they should become a blocked variants of their respective ASCII counterparts.

Allowing such characters in the RZ-LGR without making them blocked variants would mean relying on other parts of the registration process, such as a string similarity review that would need to be performed on all proposed TLDs. The Generation Panel and Integration Panel may conceivably come to the view that this would in fact be the most appropriate place in the process to address this issue.

5.7.5 Homoglyphs of Punctuation

In general, where code points are homoglyphs or near homoglyphs of code points that are not PVALID, usually punctuation characters, the Integration Panel has not included such code points in the MSR.

In particular, the following code points are highly confusable with or outright homoglyphs of code points, such as common punctuation characters like apostrophe or exclamation mark, that are not PVALID in IDNA2008 or excluded for other reasons:

- U+01C0..U+01C3 LATIN LETTER DENTAL CLICK..LATIN LETTER RETROFLEX CLICK

- U+02B9..U+02C1 MODIFIER LETTER PRIME..MODIFIER LETTER REVERSED GLOTTAL STOP
- U+02C6..U+02D1 MODIFIER LETTER CIRCUMFLEX ACCENT..MODIFIER LETTER HALF TRIANGULAR COLON
- U+02EC MODIFIER LETTER VOICING
- U+02EE MODIFIER LETTER DOUBLE APOSTROPHE
- U+A78C LATIN SMALL LETTER SALTILLO

Note that many of these characters are themselves PVALID only because of their status as "letters" by virtue of having been re-encoded by Unicode with code points classified as "modifier letters". The set of modifier letters includes these code points as clones of punctuation marks for use when writing systems employ such marks as part of words. The Integration Panel considers them an unacceptable risk for the root zone and has not included them in the MSR.

This is in keeping with the “Letter Principle”, called out in the [Procedure].

The Integration Panel recognizes that several of these code points, in particular the following six, are widely used and prominently occur in their respective writing systems. Nevertheless, the Integration Panel concludes that security concerns outweigh an interest in more naturally mnemonic TLDs, and has removed the code points from the MSR.

- U+01C0 | LATIN LETTER DENTAL CLICK
- U+01C1 ‖ LATIN LETTER LATERAL CLICK
- U+01C2 ǂ LATIN LETTER ALVEOLAR CLICK
- U+01C3 ! LATIN LETTER RETROFLEX CLICK
- U+02BB ʻ MODIFIER LETTER TURNED COMMA
- U+A78C ' LATIN SMALL LETTER SALTILLO

In particular, U+01C0 and U+01C1 are indistinguishable from the punctuation marks U+007C and U+2016 in certain user interface fonts. U+01C2 has a more distant resemblance to a line-drawing symbol U+256A; it is included here for consistency. U+01C3 is always indistinguishable from an exclamation mark.

Code point 1E37 ḷ LATIN SMALL LETTER L WITH DOT BELOW also resembles an exclamation mark (!) but appears sufficiently differentiated not to require peremptory exclusion. Nevertheless, GPs planning to include it are expected to consider any security risks introduced as well as a propose how to mitigate these as far as possible.

5.7.6 Dual Representation

When code points result in a dual representation, for example as result of spelling reform substituting one form for another, conservatism and the avoidance of risk demand particular caution. One example is U+0D4C ൗ·ൗ MALAYALAM VOWEL SIGN AU.

The Unicode Standard and other sources on Malayalam mention this vowel sign not only as “archaic”, but as superseded by another code point. A typical explanation of Vowel Sign Au, reads something like this:

“Malayalam has two forms for Vowel Sign Au. The ancient form has vowel markers on both sides of the consonant, as in കൊൗ ‘kau’. This is encoded in Unicode as ൊൗ U+0D4C MALAYALAM VOWEL SIGN AU. However, this is an archaic and obsolete method. The modern form has only one vowel marker, to the right side of the consonant, as in കൗ ‘kau’. This modern form is encoded as U+0D57 ൗ MALAYALAM AU LENGTH MARK.”

If the code point 0D4C were permitted in the Root Zone, it would represent a dual encoding that, at the minimum, would need to be investigated for a variant relation with U+0D57. It would be unlikely that any dual representation (particularly if it includes an archaic form) would be acceptable for the root. It is best not to introduce such a risk into the RZ-LGR to begin with. Therefore, the conservative response must be to disallow such a code point into the MSR, unless solid evidence is available that it is unavoidable for some reason.

5.8 IDNA 2008 Gaps and Side effects

The rules determining the PVALID status in IDNA2008 are based on a series of Unicode properties, so that IDNA2008 PVALID status can be easily updated as Unicode adds new code points and assigns properties to them. However, the rules admit some rather inappropriate characters because of accidents of character classification in The Unicode Standard.

The following code points, while formally classified as “letters”, really encode symbols, number forms or punctuation and are thus excluded from the MSR:

- U+093E DEVANAGARI SIGN AVAGRAHA[18]
- U+09BD BENGALI SIGN AVAGRAHA
- U+0ABD GUJARATI SIGN AVAGRAHA
- U+0B3D ORIYA SIGN AVAGRAHA
- U+0C3D TELUGU SIGN AVAGRAHA
- U+0CBD KANNADA SIGN AVAGRAHA
- U+0D3D MALAYALAM SIGN AVAGRAHA
- U+0F18 TIBETAN ASTROLOGICAL SIGN -KHYUD PA
- U+0F19 TIBETAN ASTROLOGICAL SIGN SDONG TSHUGS
- U+0F35 TIBETAN MARK NGAS BZUNG NYA ZLA
- U+0F37 TIBETAN MARK NGAS BZUNG SGOR RTAGS
- U+0F3E TIBETAN SIGN YAR TSHES
- U+0F3F TIBETAN SIGN MAR TSHES
- U+0FC6 TIBETAN SYMBOL PADMA GDAN
- U+214E TURNED SMALL F

[18] The Avagraha (U+093D), coming after a word-final vowel, and marking the elision of a word-initial short a in the next word can be compared to the use of an apostrophe in English. It is mostly used in Sanskrit texts.

https://en.wikipedia.org/wiki/Avagraha shows that *Avagraha* and its analogues in other Neo-Brahmi scripts have multiple uses in modern Indic languages, although they are all species of punctuation rather than letters, motivating their exclusion from the MSR.

- U+2184 LATIN SMALL LETTER REVERSED C
- U+2E2F VERTICAL TILDE
- U+3006 IDEOGRAPHIC CLOSING MARK
- U+302A..U+302D [4] IDEOGRAPHIC LEVEL TONE MARK..IDEOGRAPHIC ENTERING TONE MARK
- U+303C MASU MARK
- U+A9CF JAVANESE PANGRANGKEP
- U+A717..U+A71A [4] MODIFIER LETTER DOT VERTICAL BAR..MODIFIER LETTER LOWER RIGHT CORNER ANGLE
- U+A71B..U+A71F [5] MODIFIER LETTER RAISED UP ARROW..MODIFIER LETTER LOW INVERTED EXCLAMATION MARK

Affected code points are further annotated in the PDF file that lists the code tables.

IDNA2008 admits all combining marks, including the following that are highly specialized, or need particular conventions for correct usage, or both. Based on this analysis, the Integration Panel has not included them in the MSR.

- U+FE20..U+FE23 [4] COMBINING LIGATURE LEFT HALF..COMBINING DOUBLE TILDE RIGHT HALF
- U+FE24..U+FE26 [3] COMBINING MACRON LEFT HALF..COMBINING CONJOINING MACRON
- U+101FD PHAISTOS DISC SIGN COMBINING OBLIQUE STROKE

Additional affected code points are indicated in the PDF file that lists the code tables.

5.9 IDNA 2008 Exceptionally PVALID Code Points

The following code points are symbols that are exceptionally defined as PVALID in IDNA 2008, but because of the Letter Principle they are considered inappropriate for the Root Zone.

- U+06FD ARABIC SIGN SINDHI AMPERSAND
- U+06FE ARABIC SIGN SINDHI POSPOSITION MEN

A third code point, U+0F0B TIBETAN MARK INTERSYLLABIC TSHEG is exceptionally treated as PVALID in IDNA 2008 based on the fact that it is an obligatory mark that occurs between *all* syllables in every multi-syllabic word in Tibetan. While Unicode classifies this character as a punctuation mark (General_Category=Po), and also excludes it from the set of characters recommended for identifiers (XID_Continue=false) it can be argued that, excluding it would likely render supporting the Tibetan script as such pointless for the root. Unlike 002D HYPHEN, for example, the mark appears not directly confusable with other punctuation marks, and while it occurs inside words, it never occurs between them. It also occurs in every single word of more than one syllable.

The Integration Panel therefore resolved to include this code point in the MSR, to enable review and affirmation by a Tibetan Generation Panel. Because it would be inadvisable to allow the delegation of labels that differ only by the addition of an extraneous *tsheg*, the Tibetan Generation Panel would be advised to consider defining a rule that restricts this code point to single instances occurring medially.

5.10 Code Points Exclusively Used for Religious or Liturgical Purposes

Code points that are used exclusively for religious or liturgical purposes have been excluded from the MSR. Examples include two of three alphabets used for Georgian that are ecclesiastical in nature and restricted to liturgical use. The Arabic script has many code points exclusively used for annotating the Koran, and not used for everyday writing outside this context. The Hebrew script has code points for cantillation marks, a liturgical use. Isolated code points in other scripts were excluded on the same basis, such as:

- U+0950 DEVANAGARI OM
- U+0BD0 TAMIL OM
- U+0A74 GURMUKHI EK ONKAR
- U+0AD0 GUJARATI OM
- U+0F00 TIBETAN SYLLABLE OM

Additional code points affected are indicated in the PDF file that lists the code tables.

5.11 Threatened or Declining Languages or Orthographies

For the Latin and Cyrillic script in particular, but not limited to these scripts, there are many orthographies for languages that, while still in use, may be in decline to the point that they have fallen out of use for everyday writing, not uncommonly as result of another (majority) language being used for commercial and administrative purposes.

In some cases, such as for several orthographies in Cyrillic, the language community may have shifted to writing in a different script in recent times. In such cases, the orthography itself is obsolete even though the language community may be active and vigorous.

Where the Integration Panel was able to establish to its satisfaction that a given code point was assigned a character solely for use in a disused orthography, or for a language in serious decline, the code point has been removed from the MSR. These exclusions are fundamentally equivalent to the exclusion of disused or limited use scripts.

Affected code points are indicated in the PDF file that lists the code tables.

In making this determination, the classification of languages on the EGIDS (Expanded Graded Intergenerational Disruption Scale) documented in [SIL-Ethnologue] was used to derive a proxy measure of the *effective demand* for the corresponding writing systems. The EGIDS is based on a concept of *established vitality* which is a more useful consideration than mere population size. It does not correlate perfectly with script usage, not least because some writing systems are not stable or standardized, while the languages themselves may be. Also, as noted for Cyrillic above, a particular orthography may have fallen out of use because of other factors, or it may be limited to the preservation of cultural heritage rather than used actively in everyday affairs. Therefore, the Integration Panel considered a minimal

EGIDS score a necessary rather than a sufficient condition to assume that an orthography is in modern wide-spread use.[19]

For the MSR the IP used the cut-off between EGIDS level 4 and level 5:

4: *Educational*

> Language in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.

5: *Developing*

> Language in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.

It should be evident that the EGIDS level merely captures a snapshot of a potentially dynamic situation related to language use. Languages may gain, or lose, vitality over time. The same is true for the related writing systems, which ultimately are the object of support by LGRs. Writing systems are further impacted by literacy rates, orthographic change, or switch in preferred script. In making its determination, nevertheless, the IP must rely on present facts, not fallible long-term predictions about language or writing system usage trends.

For MSR-4, the following 12 code points were included based on reports by the Myanmar Generation Panel that the orthographies that require them are seeing more widespread use.

- U+1028 MYANMAR LETTER MON E
- U+1033 MYANMAR VOWEL SIGN MON II
- U+1034 MYANMAR VOWEL SIGN MON O
- U+105A MYANMAR LETTER MON NGA
- U+105B MYANMAR LETTER MON JHA
- U+105C MYANMAR LETTER MON BBA
- U+105D MYANMAR LETTER MON BBE
- U+105E MYANMAR CONSONANT SIGN MON MEDIAL NA
- U+105F MYANMAR CONSONANT SIGN MON MEDIAL MA
- U+1060 MYANMAR CONSONANT SIGN MON MEDIAL LA
- U+108F MYANMAR SIGN RUMAI PALAUNG TONE-5
- U+AA7B MYANMAR SIGN PAO KAREN TONE

5.12 Historical, Obsolete, or Deprecated Code Points

Scripts encoded for historical purposes or for obsolete orthographies are out of scope and therefore excluded from the MSR. Likewise, extensions to eligible scripts encoded exclusively for historical or obsolete orthographies of those scripts have been excluded from the MSR. In some cases, a determination could not be made with certainty, and the affected code points were retained in the

[19] For MSR-4, the Myanmar Generation Panel adduced evidence that 12 previously excluded code points should be considered in widespread use. These have now been added.

MSR, with the expectation that the GP considering that language would not include them in an LGR proposal without strong affirmative evidence of significant everyday modern use.

Affected code points are indicated in the PDF file that lists the code tables.

5.13 Technical Use

Many code points have been added to Unicode for specialized purposes, such as transliteration, phonetic transcription, discussion of poetry and other such technical use. Despite the fact that such code points may be correctly classified as letters in a technical sense of that term, they have been excluded from the MSR wherever their status could be determined unambiguously by the Integration Panel.

Affected code points are indicated in the PDF file that lists the non-CJK code tables.

For MSR-3, three previously excluded code points for the Latin script were found on review to appear to be part of modern orthographies that are in active and widespread use.

- U+0268 LATIN SMALL LETTER I WITH STROKE
- U+0272 LATIN SMALL LETTER N WITH LEFT HOOK
- U+1E3D LATIN SMALL LETTER L WITH CIRCUMFLEX BELOW

For MSR-4, an additional three previously excluded code points for the Latin script were found on review to appear to be part of modern orthographies that are in active and widespread use.

- U+1E13 LATIN SMALL LETTER D WITH CIRCUMFLEX BELOW
- U+1E4B LATIN SMALL LETTER N WITH CIRCUMFLEX BELOW
- U+1E71 LATIN SMALL LETTER T WITH CIRCUMFLEX BELOW

These code points are now included in the MSR.

5.14 Han Ideographs

There is a general difficulty in making a hard cutoff for the purpose of delineating "everyday use" Han Ideographs from historical, local or special purpose ideographs. Over the years there have been several attempts at defining a minimal, but sufficient set of characters for modern use. One such effort has been the set of International Ideographs Core [IICORE]; this set accounts for modern, everyday use of Han ideographs in writing the Chinese, Japanese and Korean languages (CJK).

In creating the MSR, the Integration Panel reviewed existing IDN tables for CJK domains and compared them to various subsets, including IICORE, defined in the Unicode Consortium's Unihan database [UAX38]. From this analysis, it appears that the union of certain IDN tables ([JP] and [ZH]) plus the IICORE is most likely to produce a starting set that satisfies the requirement of being larger than the expected final LGR, while at the same time not being overly inclusive.

For MSR-3, the Integration Panel included in its review the set of code points specified in the Hong Kong Supplementary Character Set, 2016 [HKSCS] but not included in the .Asia set. This lead to the addition of

two code points: U+3A5C (㩜) and U+58B5 (壵), also considered by CDNC for addition. A further extension by a single code point, U+20B9F (叱), was motivated by the desire for full coverage of the most recent version of Japanese set of Jōyō kanji [Joyo]. MSR-3 therefore extended the number of source sets for Han ideographs to five.

Chinese Characters (Han ideographs) for the MSR are listed in a separate PDF file "MSR 5 Annotated Han Tables 2021-04-06". This file uses a different convention for excluded code points, by showing them without highlighting instead of pink. The additional annotations follow the latest version of the Unicode Standard and are provided for ease of reference only.

5.14.1 Special Code Points

Three code points require special consideration in context of Han Ideographs.

1) The code point U+3005 々 IDEOGRAPHIC ITERATION MARK is in essence a symbolic notation. It is not a CJK Unified Ideograph although it may sometimes be used as a simplified form of U+4EDD 仝. Any generation panels working on that character should determine whether U+3005 should be made a character variant of U+4EDD.
2) The code point U+3006 〆 IDEOGRAPHIC CLOSING MARK is also in essence a symbolic notation. It is not a CJK Unified Ideograph although it may sometimes be used as an abbreviation of U+7DE0 締, or as a substitute for U+9589 閉, or a character variant of U+4E44 乂. Any generation panels working on that character should determine whether U+3006 should be made a character variant of U+4E44, or any other considerations towards its other related characters.
3) The code point U+9DC0 鷀 is a CJK Unified Ideograph and part of [HKSCS]. It is also part of a complex correlation between 3 code points: U+9DBF 鶯, U+9DC0 鷀, and U+9E5A 鸙. Ideally, U+9DC0 should have been the traditional variant of U+9E5A, but U+9DBF was created earlier and ended up being the commonly accepted variant. It is however important that Generation Panels evaluate these 3 code points together, even if eventually U+9DC0 is not added to any LGR.

5.15 Korean Jamo and Hangul

Modern Korean is written with 11,172 Hangul syllables, which, in turn are combinations of elements, called Jamo. The MSR contains all modern Hangul but none of the Jamo characters, which are only needed for non-modern Hangul.

The modern Hangul syllables for the MSR are listed in a separate PDF file "MSR 5 Annotated Hangul Tables 2021-04-06".

5.16 Hebrew

There are IDN tables that support extensions for Yiddish, which include points (combining marks used to indicate vowels) and digraphs. Digraphs are essentially homoglyphs for a sequence of two characters,

except that, when a combining mark is applied to them, the positioning applies to the digraph as a unit. Points are highly confusable with each other, a concern that a Hebrew GP would need to address if it were to include them in the LGR.

Existing IDN tables restrict both points and digraphs to fixed combinations (sequences of code points). This technique is reasonably simple and results in a limited number of possible sequences with the result that the allowed code points and sequences are again rather distinct; it is thus recommended to the attention of the Generation Panel.

5.17 Whole Block Exclusions

Many Unicode blocks contain one or more code points that are PVALID in IDNA 2008, but all of these code points are excluded from the MSR. Such blocks were often expressly encoded in Unicode for phonetic or historic extensions to the relevant script, and therefore do not contain any code points in widespread everyday common use. The following table lists the code point ranges and names for these blocks, as well as the rationale for exclusion of all the IDNA 2008 PVALID code points in them. Being entirely excluded from the MSR, these blocks are not documented in the PDF files.

- 02B0-02FF Spacing Modifier Letters; technical use / punctuation used as letter
- 1380-139F Ethiopic Supplement; limited use (Sebatbeit)
- 1CD0-1CFF Vedic Extensions; obsolete (historic)
- 1D00-1D7F Phonetic Extensions; technical use (phonetics)
- 1D80-1DBF Phonetic Extensions Supplement; technical use (phonetics)
- 1F00-1FFF Greek Extended; obsolete (polytonic)
- 2000-206F General Punctuation; CONTEXTJ (Join Controls)
- 2100-214F Letterlike Symbols; obsolete
- 2150-218F Number Forms; obsolete (ancient number)
- 2D00-2D2F Georgian Supplement; religious use (ecclesiastical alphabet)
- 2DE0-2DFF Cyrillic Extended-A; obsolete
- 2E00-2E7F Supplemental Punctuation; punctuation used as letter
- 31F0-31FF Katakana Phonetic Extensions; obsolete (historic)
- A640-A69F Cyrillic Extended-B; obsolete
- A700-A71F Modifier Tone Letters; punctuation used as letter
- A8E0-A8FF Devanagari Extended; technical use
- A980-A9DF Javanese; technical use
- A9E0-A9FF Myanmar Extended-B; limited use (Tai Ling) or 7.0 additions (Shan)
- AB30-AB6F Latin Extended-E; obsolete
- FE20-FE2F Combining Half Marks; technical use
- FE70-FEFF Arabic Presentation Forms B; technical use (glyph fragment)
- 101D0-101FF Phaistos Disc; obsolete (undeciphered)
- 102E0-102FF Coptic Epact Numbers; limited use
- 11300-1137F Grantha; limited use
- 16FE0-16FFF Ideographic Symbols and Punctuation; obsolete

- 1B000-1B00F Kana Supplement; obsolete
- 1B130-1B16F Small Kana Extension; obsolete
- 2A700-2B734 CJK Unified Ideographs Extension C; not in modern subset
- 2B740-2B81D CJK Unified Ideographs Extension D; not in modern subset
- 2B820-2CEA1 CJK Unified Ideographs Extension E; not in modern subset
- 2CEB0-2EBE0 CJK Unified Ideographs Extension F; not in modern subset

Any blocks that exclusively contain code points from any scripts not included in the MSR are not listed in the code tables and are also not listed here. The exception in the list above are a few blocks containing IDNA 2008 PVALID code points formally given the Unicode script property value “Common”, which would technically have made them eligible to be used in context of one or more of the included scripts.

6 Default Whole Label Evaluation (WLE) Rules

The purpose of WLE rules for the RZ-LGR is to allow automatic exclusion of labels that present particular challenges in display and processing, such as a label leading off with a combining mark, because that mark would tend to combine visually with the character in front of it.

While there may be other conditions that render a "random" label problematic in some of the complex scripts, the Integration Panel sees this as remit of the Generation Panels for such scripts and has included only a single rule intended to make labels invalid that lead off with a combining mark.

For example, rules forbidding incidence of initial or multiple dependent vowels in Brahmi scripts may be considered by the appropriate Generation Panel, and, if it they are found in agreement with the principles, might be approved for the LGR.

Note, because of the prohibition of script mixing and restricted repertoire, the Bidi Rule of [RFC5893] is automatically satisfied for all possible labels. The same applies to existing rules about digits and hyphen (which are not present in the root zone).

In accordance with [RFC7940] the default rules also contain explicit action statements that assign dispositions to variant labels based on the dispositions provided for variant code points, such as causing a variant label to be blocked if it contains any blocked variant code points.

7 Generation Panels’ Use of the MSR

As stated in the [Procedure], the Integration Panel is" *tasked with establishing the* ***maximal set of code points*** *(see Section B.5.3.2 of the Procedure) and* ***default whole label variant evaluation rules*** *(see Section B.5.5. of the Procedure) for the root zone,* ***which serve as starting point for the generation panels***" (emphasis added). These constitute the MSR. The MSR and the [Procedure] are used by the GPs as the starting point for their work.

This section gives additional guidance and direction for the GPs when evaluating the MSR. It assumes that the reader is familiar with the [Procedure]. Guidelines for the full set of Generation Panel tasks can be found in [Guidelines].

7.1 Repertoire

As stated in the [Procedure], "*The generation panel's starting point is a subset of the maximal set of code points for the root zone. From that maximal set, the generation panel picks the set of Unicode characters used in the writing systems in question.*"

The MSR is the fixed outer limit of the code point repertoire potentially available for the RZ-LGR. Following the Inclusion Principle, the Generation Panels are expected to build their proposed repertoire "from the ground up" — positively affirming each and every code point in their LGR proposals. Code points that are not part of the MSR must not be included as part of the repertoire in an LGR proposal.

As stated in the [Procedure], LGR proposals for the root zone will be created on a per-script basis, with normally no script mixing, and in particular no mixture of Latin (ASCII) letters with other scripts. Therefore, the repertoire for any LGR proposal from a given GP is expected to be the intersection of the MSR and the set of code points associated with the script in question. There are some exceptions based on the shared use of, for example, the Han script; including the way the Japanese writing system uses a mix of Hiragana, Katakana and Han code points while being treated as the script "Jpan", based on the script code defined in the ISO 15924 registry. Finally, some code points in the MSR formally have INHERITED (Zinh) as their Unicode script property value; these code points will normally be eligible to be part of the repertoire of any script for which their use is required.

For convenience of the Generation Panels, the XML file containing the normative definition of the MSR identifies the script of each code point. A small number of code points may be used with multiple scripts.

As required by the Inclusion Principle in the [Procedure], the Integration Panel expects the Generation Panels to justify the inclusion of every single code point in their proposed repertoire. While the Integration Panel may accept a summary justification for the core alphabet(s) in a script, the less common characters and sequences will have to be documented individually.

Adherence to these guidelines has the effect that the Inclusion Principle and Conservatism Principle from the [Procedure] may be fully applied to the LGR; nevertheless, even though the MSR (being an interim step) will include code points that, after further review by the Generation Panel, or after final review by the Integration Panel, are found to not satisfy these principles and therefore will not be part of the final, integrated LGR.

Some code points included in the MSR have ambiguous status or are potentially problematic for the root zone, but were included in the MSR expressly for the purpose of allowing the proper Generation Panel to research them. These include, but are not limited to the code points mentioned as problematic or ambiguous in Section 5. Generation Panels are advised that while inclusion of any code point into the LGR requires an affirmative decision under the Inclusion Principle, any potentially problematic code

points are expected to meet particularly high standards of justification before they would be acceptable to the Integration Panel for inclusion in the integrated RZ-LGR. Generation Panels that intend to submit such code points in their LGR proposals are encouraged to discuss this choice with the Integration Panel before submission.

7.2 Variants

In addition to deciding on a repertoire, the Generation Panels must decide whether any variant relationships between code points exist, and if so, must specify them. For purposes of the RZ-LGR, each code point variant must have exactly one disposition value; from these the disposition of any variant label containing them is calculated. How variants are specified in the XML format [RFC7940] is beyond the scope of this document.[20]

For each variant, the Generation Panel must make a determination about whether the presence of one variant character in a label will block another label that has the other variant code point (blocked variant), or whether the second label could be allocated later to the same applicant (allocatable). Note that assigning a disposition of "allocatable" does not mean that the second label will actually be delegated in the root zone, only that such a delegation may happen; as indicated in the [Procedure], ICANN is currently in the process of determining how "allocatable" labels will be handled.

In contrast, the effect of blocked variants is completely predictable. Because that effect prevents delegations, it can be argued that blocked variants tend to make the DNS more, not less robust - and are thus in many cases the more conservative alternative, even compared to not defining a variant relation at all. On the other hand, allocatable variants (to the degree they are delegated) do impact the DNS and its users and the conservative choice is to minimize the number of delegated variant labels. Generation Panels should consider how the Conservatism principle applies and how this affects the decision to define variant code points as allocatable, and how best to minimize their number. See also [RFC8228].

Generation Panels considering defining variants should carefully review all sections of the [Procedure] that concern variants. Appendixes E and F of the [Procedure] give useful, non-normative examples of how variants might be assigned. Finally, in Section A.3.3, the [Procedure] states:

> "It may be argued that the LGR process should be set up to minimize the number of variants defined. The benefits of a strictly minimal variant set apply only to those variants for which the returned disposition would be "allocatable". From the Conservatism Principle (if not others), it follows that the number of allocatable variants should be minimized. But the LGR process is also a way to identify all those variants that should be unambiguously blocked from allocation. Instead of minimizing the set of blocked variants, it would appear possible to simplify the evaluation of new candidate labels by maximizing the generation of such labels, thus removing them from the set that

[20] The discussion In [RFC8228] is hereby brought to the attention of Generation Panels. That RFC describes additional steps in defining variants that may be needed to ensure that variant sets are well-behaved when applied to code point sequences or in conjunction with context rules on code points.

must be subject to case-by-case analysis. In other words, the output of this procedure should aim to maximize the number of blocked variants, and to minimize the number of allocatable variants."

7.3 Restrictions on Combining Sequences

Some combining marks are used properly only in a very small number code point sequences for a particular script. A GP for such a script needs to evaluate the utility of each combining mark. Limiting the acceptable combinations of a combining mark to a small subset of characters is likely to be justified by the Conservatism Principle. Such limitations need also to be considered in light of the Simplicity and Predictability principles:

Simplicity: Overly complex rules are to be avoided, in favor of rules easily understood by users with only some background.

Predictability: People with reasonable knowledge of the topic should, by and large, reach the same conclusions about which code points should be included.

If a combining mark can be used sensibly with only a few characters, the Generation Panel may decide to add only the allowed combinations to the LGR, which would limit the use of the combining mark. On the other hand, if a combining mark is used with a wide variety of characters, the Generation Panel may decide to add the combining mark by itself to the repertoire but then needs to provide proper justification for allowing arbitrary combinations.

Complex rules that would allow a combining mark based on complicated context (other than fixed sequences) would likely run afoul of the Simplicity Principle; although something like a requirement for well-formed syllables might be appropriate for some scripts[21] in light of the adverse effects of such ill-formed syllables. Nevertheless, the least complex formulation of such rules should be aimed at, even at the loss of some linguistic fidelity. Any intention along those lines should be discussed with the Integration Panel ahead of time.

7.4 Whole Label Evaluation Rules

All LGR proposals by Generation Panels must include the default WLE rules from the MSR. They may include additional WLE rules (expressed in the XML representation) as long as they satisfy the principles in the [Procedure] and are appropriate for the root zone. If the same label can be formed using different scripts' LGRs, all WLE rules affecting it must lead to the same result. Generation Panels are advised to discuss any tentative WLE rules with the Integration Panel before submitting them as part of an LGR proposal.

7.5 Coordination between GPs

To allow the Integration Panel to create an integrated LGR for the root zone requires that proposed LGRs for related scripts are available so they can be reviewed together. Attempts to integrate each proposal in isolation would create unacceptable risks of incompatibilities and risks violating the *Stability*

[21] These scripts include the neo-Brahmi scripts of India, Sinhalese, Tamil, Khmer, Lao, Thai, Myanmar and Tibetan.

Principle and the *Least Astonishment Principle*. This has some straightforward consequences for the work of GPs covering related scripts. As stated in the [Procedure]:

> *"Panels for **related or structurally similar scripts** are encouraged to communicate or cooperate in the interest of arriving at a more consistent treatment of repertoires and variants for the root zone."* (Emphasis added).

Ideally, GPs for related scripts would be active at a similar phase of development and coordinate their efforts, so as to resolve any issues arising out of the relationship between the scripts in question. To facilitate the procedure-mandated dialogue between the panels, GPs are encouraged to keep the IP advised of their plans for and progress of such coordination.

Each Generation Panel still submits a separate LGR per script. Even in cases of significant overlap (as between Chinese and Japanese use of the Han script) the coordinated repertoires may differ. (For example, the Chinese LGR would not be expected to include Japanese only ideographs in its repertoire).

Where there is an overlap between the repertoires, any variant mappings specified in the integrated RZ-LGR must be the same. However, whether a particular variant mapping (by its assignment of type) results in a blocked label or and allocated one may be different for each script LGR. Therefore, the types assigned to variant mappings are specific to each script's LGR.

There is a provision that allows the proposed LGR to contain variant mappings to code points that are not in the repertoire, to facilitate the specification of shared variants in a symmetric and transitive manner, even where the repertoires differ. However, in some instances, an LGR may simply rely on having the variants imposed during integration.

Where a requirement for coordination between GPs may exist, a GP may submit a preliminary, not yet coordinated LGR if they would like ICANN to perform a review of the contents to determine issues that need to be addressed in such coordination.

8 Summary of Changes

1. MSR-1 added 32,790 code points for 22 scripts and one default WLE rule.
2. MSR-2 contains these changes from MSR-1:
   - Addition of 700 code points for six previously deferred scripts
   - No additions to the repertoire of scripts covered in MSR-1
   - No changes to the base Unicode Version (6.3)
   - No changes in default WLE rules
3. MSR-3 contains these changes from MSR-2:
   - Addition of 3 code points to the Hani script
   - Addition of 3 code points to the Latn script
   - A small adjustment in the presentation of non-cjk repertoire (highlighting)
   - Rationale for continued exclusion of Avagraha code points
   - Additional editorial clarifications and corrections

- No changes to the base Unicode Version (6.3)
- No changes in default WLE rules

4. MSR-4 contains these changes from MSR-3:
   - Addition of 3 code points to the Latn script
   - Addition of 12 code points to the Mymr script
   - Additional editorial clarifications and corrections
   - No changes to the base Unicode Version (6.3)
   - No changes in default WLE rules
5. MSR-5 contains these changes from MSR-4
   - Addition of 2 code point to the Arabic script
   - Addition of 1 code point to the Devanagari script
   - Addition of 1 code point to the Latin script
   - Change of Unicode base version from 6.3 to 11.0

8.1 Code points by script

Script tag[22]	MSR-1	MSR-2	MSR-3	MSR-4	MSR-5
Arab	239	239	239	239	241
Armn	-	38	38	38	38
Beng	64	64	64	64	64
Cyrl	93	93	93	93	93
Deva	91	91	91	91	92
Ethi	-	364	364	364	364
Geor	37	37	37	37	37
Grek	36	36	36	36	36
Gujr	66	66	66	66	66
Guru	61	61	61	61	61
Hang	11172	11172	11172	11172	11172
Hani	19852	19852	19855	19855	19855
Hebr	46	46	46	46	46
Hira	89	89	89	89	89
Kana	92	92	92	92	92
Khmr	-	78	78	78	78
Knda	68	68	68	68	68
Laoo	53	53	53	53	53
Latn	305	305	308	311	312
Mlym	73	73	73	73	73
Mymr	-	90	90	102	102
Orya	66	66	66	66	66
Sinh	79	79	79	79	79
Taml	49	49	49	49	49
Telu	67	67	67	67	67
Thaa	-	50	50	50	50
Thai	71	71	71	71	71
Tibt	-	80	80	80	80
Zinh	21	21	21	21	21
Total	32790	33490	33496	33511	33515

9 Contributors

MSR-3 was developed by the Integration Panel, with expert input from external advisors and community members, as well as support by ICANN staff members. As the repertoire is cumulative, contributions to earlier versions of the MSR are listed.

1. Integration Panel Members

Marc Blanchet
Asmus Freytag
Nicholas Ostler

[22] Code points with multiple tags are listed only once.

Michel Suignard
Wil Tan

2. Advisors

Michael Everson (MSR-1)
Paul Hoffman (MSR-1)
Thomas Milo (MSR-1, Arabic)

3. Community Members

The integration panel gratefully acknowledges the information provided by the following members of the community:
Andrew Cunningham (MSR-1, Harari)
Chea Sak Huor (MSR-2, Khmer)
Christopher Fynn (MSR-2, Tibetan)
Thura Hlaing (MSR-2, Myanmar)
Andreas Wetter(MSR-2, Ethiopic)

4. ICANN Staff

Sarmad Hussain
Pitinan Kooarmornpatana

10 Advisor Reports

In accordance with the LGR procedure, the Integration Panel relied on the contribution of advisors to the development and review of MSR-1. None of the advisors elected to submit a separate report. No advisors were consulted during the development of subsequent versions.

11 References

[ArabicVIP] Hussain, S, *et al.* "Internationalized Domain Names Variant Issues Project Arabic Case Study Team Issues Report". (Marina del Rey, California: ICANN, October 2011). http://archive.icann.org/en/topics/new-gtlds/arabic-vip-issues-report-07oct11-en.pdf.

[CALL-FOR-PANELS] "Call for Generation Panels to Develop Root Zone Label Generation Rules" http://www.icann.org/en/news/announcements/announcement-11jul13-en.htm.

[ChineseVIP] Lee, X. *et al.*, "Report on Chinese Variants in Internationalized Top-Level Domains". (Marina del Rey, California: ICANN, October 2011). http://archive.icann.org/en/topics/new-gtlds/chinese-vip-issues-report-03oct11-en.pdf.

[CyrillicVIP] Sozonov, A. *et al.*, "IDN Variant TLDs – Cyrillic Script Issues". (Marina del Rey, California: ICANN, October 2011). http://archive.icann.org/en/topics/new-gtlds/cyrillic-vip-issues-report-06oct11-en.pdf.

[DevanagariVIP] Govind, *et al.,* "Devanāgarī VIP Team Issues Report". (Marina del Rey, California: ICANN, October 2011). http://archive.icann.org/en/topics/new-gtlds/devanagari-vip-issues-report-03oct11-en.pdf.

[GreekVIP] Segredakis, V., et al., "Study of the issues present in the registration of IDN TLDs in GREEK characters". (Marina del Rey, California: ICANN, October 2011). http://archive.icann.org/en/topics/new-gtlds/greek-vip-issues-report-07oct11-en.pdf.

[Guidelines]Internet Corporation for Assigned Names and Numbers, "Guidelines for Developing Script-Specific Label Generation Rules for Integration into the Root Zone LGR". (Los Angeles, California: ICANN, December 2014) https://community.icann.org/download/attachments/43989034/Guidelines-for-LGR-2014-12-02.pdf.

[HKSCS] Office of the Government Chief Information Officer & Official Languages Division, "Hong Kong Supplementary Character Set, 2016" (Hong Kong: Civil Service Bureau, The Government of the Hong Kong Special Administrative Region, May 2017) Available as https://www.ogcio.gov.hk/en/business/tech_promotion/ccli/terms/doc/e_hkscs_2016.pdf from: https://www.ogcio.gov.hk/en/business/tech_promotion/ccli/terms/terms_38a_hkscs2016.htm, Visited: 2017-12-01.

[IABCP] Sullivan, A., *et al.,* "Principles for Unicode Code Point Inclusion in Labels in the DNS". Internet Architecture Board (IAB) = RFC 6912 http://tools.ietf.org/html/rfc6912.

[IDNAREG] IANA Registry: "IDNA Parameters". For Unicode 11.0 available at: https://www.iana.org/assignments/idna-tables-11.0.0/idna-tables-11.0.0.xhtml Visited 2021-03-16

[IDNFT] ICANN, IDN ccTLD Fast Track Process, http://www.icann.org/en/resources/idn/fast-track. Visited 2014-02-18.

[IICORE] *International Ideographs Core (IICORE),* http://www.ogcio.gov.hk/en/business/tech_promotion/ccli/iso_10646/iicore.htm. Visited 2014-01-07

[ISO15924] *Codes for the representation of names of scripts*, ISO 15924:2004. Available from http://www.unicode.org/iso15924/. Visited 2018-01-05.

[Joyo] List of 2136 jōyō kanji (常用漢字), issued in 2010 by the Japanese Ministry of Education, as listed in: https://en.wikipedia.org/wiki/List_of_jōyō_kanji, Visited 2018-02-05.

[JP] IDN Tables for the .jp domain (Japanese) dated 2005-08-30 deposited by Japan Registry Services Co., Ltd.
http://www.iana.org/domains/idn-tables/tables/jp_ja-jp_1.2.html

[LatinVIP] Frakes, J, *et al.,* "Considerations in the use of the Latin script in variant internationalized top-level domains: Final report of the ICANN VIP Study Group for the Latin script". (Marina del Rey, California: ICANN, October 2011). http://archive.icann.org/en/topics/new-gtlds/latin-vip-issues-report-07oct11-en.pdf.

[Latvian] Wikipedia. "Latvian Orthography", https://en.wikipedia.org/wiki/Latvian_orthography, Visited 2018-02-28.

[NEWGTLD] ICANN, New Generic Top Level Domains, http://newgtlds.icann.org, Visited 2014-02-18.

[Marshallese] Wikipedia, "Marshallese Language", https://en.wikipedia.org/wiki/Marshallese_language#Display_issues, Visited 2018-02-28.

[MSR-1] Internet Corporation for Assigned Names and Numbers, "Maximal Starting Repertoire – MSR-1", (Los Angeles, California: ICANN, June, 2014). Available from https://www.icann.org/resources/pages/msr-2015-06-21-en.

[MSR-2] Internet Corporation for Assigned Names and Numbers, "Maximal Starting Repertoire – MSR-2", (Los Angeles, California: ICANN, June 2014). Available from https://www.icann.org/resources/pages/msr-2015-06-21-en.

[MSR-3] Internet Corporation for Assigned Names and Numbers, "Maximal Starting Repertoire – MSR-3", (Los Angeles, California: ICANN, March 2018). Available from https://www.icann.org/resources/pages/msr-2015-06-21-en.

[MSR-4] Internet Corporation for Assigned Names and Numbers, "Maximal Starting Repertoire – MSR-4", (Los Angeles, California: ICANN, January 2019). Available from https://www.icann.org/resources/pages/msr-2015-06-21-en.

[MSRGupta] Nehu Gupta et al., "Comments on Maximal Starting Repertoire – MSR-1 Overview and Rationale", http://forum.icann.org/lists/comments-msr-03mar14/pdfgYaBIQ8s9G.pdf

[Procedure] Internet Corporation for Assigned Names and Numbers, "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels." (Los Angeles, California: ICANN, March, 2013) http://www.icann.org/en/resources/idn/variant-tlds/draft-lgr-procedure-20mar13-en.pdf

[RFC1035] Mockapetris, P., "Domain names - implementation and specification", RFC 1035, November 1987.

[RFC3743] Konishi, K., Huang, K., Qian, H., and Y. Ko, "Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese, and Korean", RFC 3743, April 2004.

[RFC4290] Klensin, J., "Suggested Practices for Registration of Internationalized Domain Names (IDN)", RFC 4290, December 2005.

[RFC5646] Phillips, A. and M. Davis, Eds., “Tags for Identifying Languages”, RFC 5646, BCP 47, September 2009.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, August 2010.

[RFC5891] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", RFC 5891, August 2010.

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, August 2010.

[RFC5893] Alvestrand, H., Ed., and C. Karp, "Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)", RFC 5893, August 2010.

[RFC5894] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale", RFC 5894, August 2010.

[RFC5895] Resnick, P. and P. Hoffman, "Mapping Characters for Internationalized Domain Names in Applications (IDNA) 2008", RFC 5895, September 2010.

[RFC6912] Sullivan, A., *et al.,* “Principles for Unicode Code Point Inclusion in Labels in the DNS”, RFC 6912, April 2013. = IABCP

[RFC7940] Davies, K. and A. Freytag, "Representing Label Generation Rulesets using XML", RFC 7940, August 2016, https://tools.ietf.org/html/rfc7940. Visited 2017-12-01.

[RCF8228] A. Freytag, “Guidance on Designing Label Generation Rulesets (LGRs) Supporting Variant Labels, August 2017, https://tools.ietf.org/html/rfc8228. Visited 2017-12-01.

[SIL-Ethnologue] Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2014. *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. Online version available as http://www.ethnologue.com.

[UAX24] UAX #24: *Unicode Script Property.* An integral part of The Unicode Standard. Most recent version available from http://www.unicode.org/reports/tr24/. Version 11.0.0 available as https://www.unicode.org/reports/tr24/tr24-28.html.

[UAX29] UAX #29: *Unicode Text Segmentation.* An integral part of The Unicode Standard. Most recent version available from http://www.unicode.org/reports/tr29/. Version 11.0.0 available as http://www.unicode.org/reports/tr29/tr29-33.html**Error! Hyperlink reference not valid.**.

[UAX31] UAX #31: *Unicode Identifier and Pattern Syntax.* An integral part of The Unicode Standard. Most recent version available from http://www.unicode.org/reports/tr31/. Version 6.3 available as http://www.unicode.org/reports/tr31/tr31-29.html.

[Unicode63] The Unicode Consortium. The Unicode Standard, Version 6.3.0, defined by: "The Unicode Standard, Version 6.3.0", (Mountain View, CA: The Unicode Consortium, 2013. ISBN 978-1-936213-08-5). http://www.unicode.org/versions/Unicode6.3.0/.

[Unicode110] The Unicode Consortium. The Unicode Standard, Version 11.0.0, defined by: "The Unicode Standard, Version 11.1.0", (Mountain View, CA: The Unicode Consortium, 2018. ISBN 978-1-936213-19-1). http://www.unicode.org/versions/Unicode11.0.0/.

[UAX38] Unicode Standard Annex #38, "Unicode Han Database (Unihan)" edited by John H. Jenkins 井作恆, Richard Cook 曲理查 and Ken Lunde 小林劍, an integral part of The Unicode Standard. Most recent version available from http://www.unicode.org/reports/tr38/
Version 11.0.0 available as http://www.unicode.org/reports/tr38/tr38-25.html.

[UTC] Unicode Technical Committee. The document register for the UTC can be found at http://www.unicode.org/L2/L-curdoc.htm.

[UTS39] UTS#39: Unicode Security Mechanisms. Available from http://www.unicode.org/reports/tr39/.

[UTS46] UTS#46: Unicode IDNA Compatibility Processing. Available from http://www.unicode.org/reports/tr46/.

[ZH] DotAsia Organisation, ".ASIA ZH IDN Language Table", 2011-05-04, http://www.iana.org/domains/idn-tables/tables/asia_zh_1.1.txt.