Integration Panel:

Maximal Starting Repertoire — MSR-1 Overview and Rationale

REVISION – June 6, 2014

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]. This document gives the 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]

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

are repeated in this document, but the [Procedure] as a whole is the formal framework on which the MSR is based.

2 Maximal Starting Repertoire (MSR-1)

2.1 Files

MSR-1 is provided as collection of files. The current document 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 additional guidance to Generation Panels on using the MSR as basis for their LGR proposals to the Integration Panel.

The normative definition of MSR-1 is provided as an XML file MSR-1-Repertoire+WLE-Rules-20140606. The MSR is expressed using a standard format defined in "Representing Label Generation Rulesets in XML" [XML-LGR]. This format provides for 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 used for the MSR supports the specification of variants and their disposition; however, these features are unused in the MSR. The Generation Panels are expected to use the XML format when submitting their script-specific LGR proposals to the Integration Panel (including variant information where applicable).

A non-normative PDF file MSR-1-Annotated-non-CJK-Tables-20140606 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 without highlighting (that is, white cells) are for code points that are not PVALID in IDNA 2008 [RFC5892][IDNAREG], or otherwise excluded in a generic fashion from the root zone (digits, hyphen). Code cells with yellow highlighting are part of the MSR. Code cells with pink highlighting are ***excluded*** from the MSR, but shown for reference for any block of code points that also contains part of the repertoire of the MSR.

The tabular listing of Unicode character names contains additional information about certain code points; for excluded code points listed, it includes a shorthand notation for the principal rationale leading to the exclusion of the code point.

For CJK Unified ideographs the file repeats the source information published in the Unicode Standard.

Because of size, the tables showing repertoires for Han ideographs MSR-1-Annotated-Han-Tables-20140606 and Hangul syllables MSR-1-Annotated-Hangul-Tables-20140606 are broken off into separate PDF files. For these files, no highlighting (white cells) represents code points excluded from the MSR.

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 defined in [IABCP]. While these principles apply to the overall process of defining the integrated LGR for the root zone, 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.
  - 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 status of any individual code point, 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 the Generation Panels will propose to include only a subset of code points that are in scope for their respective scripts.

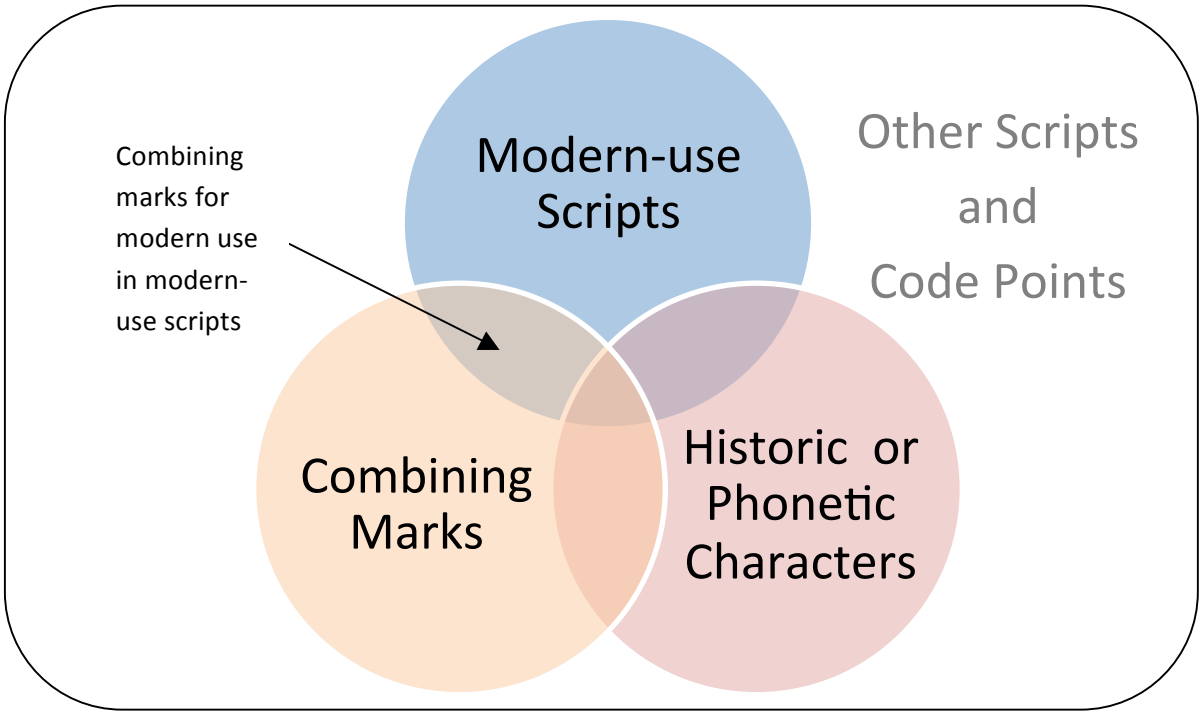
The Integration Panel is tasked to evaluate its actions in light of the Principles laid out in the [Procedure]. The methodology followed by the Integration Panel ensures that the Stability, Inclusion and Conservatism Principles may be fully applied to the final result (LGR) but recognizes that the MSR is merely an interim step in the development on the LGR, and that any code points included in it, do not automatically get added to the 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 use those 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 modern-use scripts, based on the classification suggested in [UTS39] and refined further based on the relevance of each script to IDN label applied for as part of the gTLD [NEWGTLD] and ccTLD [IDNFT] processes.

The set of code points for these scripts was taken from Unicode 6.3, the latest published version of that standard at the time of defining MSR-1. This set was then intersected with the set of code points that are deemed PVALID by applying the methodology of [RFC5892] to the repertoire of Unicode 6.3. This is indicated by the blue circle in the diagram.



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

Combining marks may be present in any script. Many are solely intended for historic or other specialized use, and are therefore excluded. The remaining 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 belonging to Modern-use scripts.

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 in 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 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 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. All work done on the first version of the LGR (LGR-1) will be based on MSR-1. If it becomes necessary to stage the release of the LGR, for example because not all Generation Panels are able to submit proposals at the same time, subsequent versions of the LGR may be released without necessarily re-issuing the MSR.

MSR-1 defers some part of the possible repertoire, so as to balance timeliness with comprehensiveness (see Section 3.5.3). At some future point in time, another version, MSR-2 will be developed that includes these deferred repertoires. MSR-2 would be the foundation for any subsequent LGR.

All future versions of the MSR and all versions of the 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. Repertoire that has not been used for label registration is not required to be retained in future versions.

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, 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 scripts 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**[2] 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.

A number of separable scripts have been deferred to a later version of the MSR, primarily in the interest of managing the work and allowing a timely release of an initial MSR (see Section 3.5.3). The intention is to include these scripts in future releases of the MSR. However, not all scripts that are excluded from MSR-1 can be considered “deferred” in this way.

With release of MSR-1, the Integration Panel is confirming the position in the [Procedure], and has taken the decisions 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 of use, are removed from consideration: they are not just "separable" by definition but "separated". 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. Doing so 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.

[2] These can take the form of 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 □ (VA), 0909 □ (U), 092F □ (YA), 093F □ (VOWEL SIGN I), 0940 □ (VOWEL SIGN II)
vs
Gurmukhi 0A15 □ (KA), 0A24 □ (TA), 0A27 □ (DHA), 0A3F □ (VOWEL SIGN I), 0A40 □ (VOWEL SIGN II)

3.3 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 the MSR was the intersection between the latest published version of the Unicode Standards, Unicode 6.3, and the [IDNA 2008] PVALID set derived [IANAREG] from the prescriptions of [RFC5892]. In further analyzing the set, the additional data tables provided by the Unicode Consortium for Unicode Technical Standard #46 [UTS46] and Unicode Technical Standard #39 [UTS39] for Unicode 6.3 were helpful inputs, as well as the tables in Unicode Standard Annex #31 [UAX31]. In addition, the IP resolved to defer a number of eligible scripts to a future version of the MSR (see discussion below).

3.4 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]. "*

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.

3.5 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 chosen to categorize the scripts in this list based on whether or not there has been a TLD application that used the script, and further on whether a script is part of a of a known group for official country languages, together with at least one scripts with a TLD application. The latter occurs for the official languages of India.

The IP further notes whether or not the script is separable from other scripts, that is, whether it may be considered in isolation without adversely affecting the remainder of the MSR or the LGR. There are also two collections of code points that are used with multiple scripts. The reasons for making these categorizations are discussed below.

UAX#31, Table 5. Recommended Scripts[3]

Script ID	Description	Remarks
Zyyy	Common	Shared across scripts
Zinh	Inherited	Shared across scripts
Arab	Arabic	Applied TLD
Armn	Armenian	Separable
Beng	Bengali[4]	Applied TLD
Bopo	Bopomofo	Separable, educational use only
Cyrl	Cyrillic	Applied TLD
Deva	Devanagari	Applied TLD
Ethi	Ethiopic	Separable
Geor	Georgian	Applied TLD
Grek	Greek	Applied TLD
Gujr	Gujarati	Applied TLD
Guru	Gurmukhi	Applied TLD
Hani	Han	Applied TLD
Hang	Hangul	Applied TLD
Hebr	Hebrew	Applied TLD
Hira	Hiragana	Applied TLD
Knda	Kannada	Required to represent official language of India
Kana	Katakana	Applied TLD
Khmr	Khmer	Separable
Laoo	Lao	Related to Thai, not separable
Latn	Latin	Applied TLD
Mlym	Malayalam	Required to represent official language of India
Mymr	Myanmar	Separable
Orya	Oriya	Required to represent official language of India
Sinh	Sinhala	Applied TLD
Taml	Tamil	Applied TLD
Telu	Telugu	Applied TLD
Thaa	Thaana	Separable
Thai	Thai	Applied TLD
Tibt	Tibetan	Separable

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

3.5.1 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

[3] 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.

[4] This script is known as Assamese to many users

(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.

3.5.2 Scripts with Pending TLD Applications and Closely Related Scripts

As part of the IDN ccTLD Fast Track Process [IDNFT] and New gTLD Program [NEWGTLD], ICANN has received IDN TLD applications for 18 scripts. These are identified as "Applied TLD" in the remarks column and make up the majority of the scripts in this table. There are additional scripts that are strongly related to one or more of these scripts. This includes a number of scripts that belong in a group of scripts for official country languages with applied for scripts. Under the consideration of Comprehensiveness, it is very desirable to be able to review them at the same time, hence the decision to also include them in the first release of the MSR. This leads to a list of 19 scripts that are included in MSR-1, all but one of which were already called out by ICANN in the call for Generation Panels [CALL-FOR-PANELS][5].

3.5.3 Modern Scripts Deferred for a Later Update of the MSR

In contrast, some scripts are not strongly related, but instead separable. Separable scripts do not have to be included in the initial review of the Unicode repertoire because the LGRs proposed for them are not expected to interact with LGRs for other scripts. Some of them have issues that are expected to require additional time for the Integration Panel to resolve. Therefore, they have been postponed to a future version of the MSR.

3.5.4 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.

3.6 Scripts for Possible Future MSRs

The scripts listed in UAX#31 Table 6, ","Aspirational Use Scripts" and Table 7, "Limited Use Scripts" are not candidates for the MSR at this point, but 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 included in MSR.

UAX#31 - Table 6. Aspirational Use Scripts

Script ID	Description
Cans	Canadian_Aboriginal
Plrd	Miao
Mong	Mongolian
Tfng	Tifinagh
Yiii	Yi

[5] The [CALL-FOR-PANELS] uses "Japanese" as a collective term that includes Hiragana and Katakana.

UAX#31 - Table 7. Limited Use Scripts

Script ID	Description
Bali	Balinese
Bamu	Bamum
Batk	Batak
Cakm	Chakma
Cham	Cham
Cher	Cherokee
Java	Javanese
Kali	Kayah_Li
Lana	Tai Tham
Lepc	Lepcha
Limb	Limbu
Lisu	Lisu
Mand	Mandaic
Mtei	Meetei Mayek[6]
Nkoo	Nko
Olck	Ol Chiki6
Saur	Saurashtra
Sund	Sundanese
Sylo	Syloti_Nagri
Syrc	Syriac
Tale	Tai Le
Talu	New Tai Lue
Tavt	Tai Viet
Vaii	Vai

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

[UAX31], Table 4, is intended mainly as control to make sure that all scripts in Unicode 6.3 are covered in this document. These are a mixed bag of ancient, recently obsolete, educational, and apparently declining scripts. 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. Candidate Scripts for Exclusion from Identifiers

Script ID	Description
Armi	Imperial Aramaic
Avst	Avestan

[6] The script is considered "resurgent" by some, which would normally imply a grouping under "aspirational". For the purposes of the Root Zone LGR the difference in classification is not significant.

Brah	**Brahmi**
Bugi	**Buginese**
Buhd	**Buhid**
Cari	**Carian**
Copt	**Coptic**
Cprt	**Cypriot**
Dsrt	**Deseret**
Egyp	**Egyptian Hieroglyphs**
Glag	**Glagolitic**
Goth	**Gothic**
Hano	**Hanunoo**
Ital	**Old Italic**
Khar	**Kharoshthi**
Kthi	**Kaithi**
Linb	**Linear B**
Lyci	**Lycian**
Lydi	**Lydian**
Mero	**Meroitic Hieroglyphs**
Merc	**Meroitic Cursive**
Ogam	**Ogham**
Orkh	**Old Turkic**
Osma	**Osmanya**
Phag	**Phags Pa**
Phli	**Inscriptional Pahlavi**
Phnx	**Phoenician**
Prti	**Inscriptional Parthian**
Rjng	**Rejang**
Runr	**Runic**
Samr	**Samaritan**
Sarb	**Old South Arabian**
Shrd	**Sharada**
Shaw	**Shavian**
Sora	**Sora Sompeng**
Tagb	**Tagbanwa**
Tglg	**Tagalog**
Takr	**Takri**
Ugar	**Ugaritic**
Xpeo	**Old Persian**
Xsux	**Cuneiform**

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 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 or phonetic 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.

In the following sections we discuss specific code points excluded from the MSR on these grounds.

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[7],
- highly confusable with an existing and common punctuation character (i.e. with one that falls within Unicode General Category 'P', the union of {Pc, Pd, Pe, Pf, Pi, Po, Ps}),
- 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 characters like 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 notations. Where it could be established unambiguously that a character was encoded solely for that purpose, the Integration Panel decided to exclude it from the MSR.

[7] 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 combing mark.

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 Root Zone 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,
- 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 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.

In the Latin and Cyrillic scripts, in particular, there are dozens of characters that have been encoded to support one or more systems of phonetic notations. Cases of actual or potential dual-use with modern writing systems do exist, and the Integration Panel has opted in those cases to include the code points in the MSR to enable Generation Panel review of and confirmation the code point is in fact required for the Root Zone LGR.

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 for medieval and other transcriptions,
- phonetic use, or

- otherwise inappropriate for the DNS root zone.

Combining marks used in precomposed characters (combinations of base character and diacritic assigned a single code point) 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.

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 (sequence) needed for some orthography worth supporting in the root zone.

The actual set of combining marks allowable in the LGR will be smaller, 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, rather than adding the "bare" combining mark to the repertoire, where feasible. 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 are not merely similar in shape, but rather are frequently (if inadvertently) treated as "variants" of each other; under the suggested treatment S with cedilla below and S with comma below would then block each other.

Because all the labels allowed by IDNA2008 are precomposed using Normalization Form C, all precomposed characters containing either of 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.

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 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] 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.

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

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

The following characters have compatibility issues with IDNA2003 which makes them candidates for summary exclusion from the MSR on grounds of Longevity (§2.1):

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

In IDNA2003, case folding is applied which removes them from the final string and replaces them with other characters; however both characters are reasonably frequent in their respective orthographies. In Greek, the final sigma (ς) would be the form chosen whenever sigma (σ) ends a label. 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), German orthography has changed such that there is now a distinct difference 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. 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

This character has case mappings that are locale sensitive and thus were an issue for IDNA 2003. In IDNA 2008 there is no 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 Ideogracph-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 (not yet published as of the publication of the present document) will add 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 pending 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 pending 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 root zone LGR work, because the pending additions remain outside the Unicode 6.3 cutoff for MSR-1. The Integration Panel suggests that the Generation Panel include the pending code points in its analysis to reduce 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. Therefore, the Integration Panel expects that their inclusion in an 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 character membership in multiple scripts. 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 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.

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 Root Zone LGR. The Integration Panel does not expect that cross-script homoglyphs would ever become allocatable variants, because that would imply mixed-script repertoires.

Examples include

- U+006F LATIN SMALL LETTER O
- U+03BF GREEK SMALL LETTER OMICRON
- U+043E CYRILLIC SMALL LETTER O

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 ARABIE 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 Root Zone LGR will be accompanied by a suitable proposal on how to mitigate their effect.

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. See the discussion in Section 5.15.

Digraphs code points also exist in the Latin script. However, all instances identified are for limited use and therefore already excluded from the MSR.

5.7.3 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 (practically) indistinguishable from U+0061 a LATIN SMALL LETTER A. (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 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.

5.7.4 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

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. 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

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+06FD ARABIC SIGN SINDHI AMPERSAND
- U+06FE ARABIC SIGN SINDHI POSPOSITION MEN
- U+214E TURNED SMALL F
- 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

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 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

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

5.10 Threatened or Declining Languages or Orthographies

For the Latin and Cyrillic script in particular 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, usually 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. 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.

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. 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 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.

5.11 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 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.

5.12 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 code tables.

5.13 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 a superset of certain IDN tables 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.

Chinese Characters (Han ideographs) for the MSR are listed in a separate PDF file MSR-1-Annotated-Han-Tables-20140606. This file uses a different convention for excluded code points, by showing them without highlighting instead of pink. The additional annotations follow the Unicode Standard and are provided for ease of reference only.

5.13.1 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 〆. 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. It is also part of a complex correlation between 3 code points: U+9DBF 鶿, U+9DC0 鷀, and U+9E5A 鸕. 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 generations panels evaluate these 3 code points together, even if eventually U+9DC0 is not added to any LGR.

5.14 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 for the MSR are listed in a separate PDF file MSR-1-Annotated-Hangul-Tables-20140606.

5.15 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.

5.16 Whole Block Exclusions

The following 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. Many of these blocks were expressly encoded in Unicode for phonetic or historic extensions to the relevant script, and thus excluded from the MSR. The following table lists these code point rand and name of 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
- 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
- A720-A7FF Latin Extended-D; technical use (phonetic)/obsolete/punctuation
- A8E0-A8FF Devanagari Extended; technical use
- A980-A9CD Javanese; technical use
- FE20-FE2F Combining Half Marks; technical use
- FE70-FEFF Arabic Presentation Forms B; technical use (glyph fragment)

- 101D0-101FF Phaistos Disc; obsolete (undeciphered)
- 1B000-1B00F Kana Supplement; obsolete
- 2A700-2B734 CJK Unified Ideographs Extension C; not in modern subset
- 2B740-2B81D CJK Unified Ideographs Extension D; not in modern subset

6 Default Whole Label Evaluation (WLE) Rules

The purpose of WLE rules for the Root Zone 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 [XML-LGR] 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 are the MSR. The MSR and the [Procedure] are used by the GPs as starting points 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].

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 Root Zone 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 in an LGR proposal.

As stated in the [Procedure], LGR proposals for the root zone will be created on a per-script basis, with no script mixing. Therefore repertoire for any LGR proposal from a given GP is expected to be a strict subset of the MSR and the code points associated with script in question. There are some exceptions based on the shared use of, for example, the Han script; another example is the way the Japanese writing system will use 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. For convenience of the Generation Panels, the XML file 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 Root Zone 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 Root Zone 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 [XML-LGR] is beyond the scope of this document.

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 (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.

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.

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 Indic Syllables might be appropriate in light of the adverse effects of such ill-formed syllables. 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. 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 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). If there is an overlap between the repertoires, any variant mappings specified must be consistent. However, whether a particular variant results in a blocked label or not may be different for each LGR.

8 Contributors

MSR-1 was developed by the Integration Panel, with expert input from external advisors and support by ICANN staff members.

1. Integration Panel Members

Marc Blanchet
Asmus Freytag
Nicholas Ostler
Michel Suignard
Wil Tan

2. Advisors

Michael Everson
Paul Hoffman
Thomas Milo

3. ICANN Staff

Nicoleta Munteanu
Naela Sarras

9 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.

10 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.

[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.

[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 6.3 available at: http://www.iana.org/assignments/idna-tables-6.3.0/idna-tables-6.3.0.xml. Visited 2013-11-20.

[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 2012-09-21.

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

[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

[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 6.3 available as http://www.unicode.org/reports/tr24/tr24-21.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 6.3 available as http://www.unicode.org/reports/tr29/tr29-23.html.

[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/tr31-19.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/.

[UAX38] Unicode Standard Annex #38, "Unicode Han Database (Unihan)" edited 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 6.3 available as http://www.unicode.org/reports/tr15/tr38-15.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/. Visited 2012-09-21.

[UTS46] UTS#46: Unicode IDNA Compatibility Processing. Available from http://www.unicode.org/reports/tr46/. Visited 2013-11-11.

[XML-LGR] Davies, K. and A. Freytag, "Representing Label Generation Rulesets using XML", http://tools.ietf.org/html/draft-davies-idntables/.Visited 2014-06-06.