

A Yiddish Character Repertoire for Internationalized Domain Names

Cary Karp
Swedish Museum of Natural History

1. Introduction

The registries for the .MUSEUM generic top-level domain and the .SE national top-level domain both have their administrative and operational headquarters in Stockholm, Sweden. The two registries are therefore collaborating on the implementation of Internationalized Domain Names (“IDNs”) derived from all languages with official status in that country. (An IDN is a domain name containing at least one character that is not a basic Latin letter “a-z”, a digit “0-9”, or a hyphen “-”, as defined in <<http://www.faqs.org/rfcs/rfc3490.htm>>. Readers unfamiliar with the basic concepts of domain naming will find a brief review in Appendix 4, below.)

In addition to Swedish, which is the de facto national language, there are five legally recognized minority languages: Finnish, Meänkeli, Romani, Sami, and Yiddish. Strict guidelines issued in 2005 by the national government about names in administrative systems specify a Latin character repertoire that provides a comprehensive basis for the IDN representation of all but one of these languages, in all the varieties used in Sweden. The exception is Yiddish, which is written with an alphabet based on Hebrew script.

The expanded Latin character repertoire in those guidelines (starting on page 5 in the Swedish text at <<http://about.museum/idn/riktlinjer0505.pdf>>), supports Latin-script IDNs without need for explicit reference to any targeted language. However, descriptions of language-based subsets of this repertoire are useful for informational purposes. One such statement can be seen in the *IANA Repository of TLD IDN Practices* where the characters needed specifically for the Swedish language are described and tabulated for the benefit of anyone interested in a listing provided by a registry that conducts its primary business in Swedish <<http://www.iana.org/assignments/idn/se-swedish.html>>.

There is no similar source for IDN-oriented information about Yiddish. Since no national government affords that language stronger recognition than it has in Sweden, it seems appropriate for the .SE and .MUSEUM registries to produce a Yiddish character table suitable for publication in the IANA repository. The following text was prepared in consultation with the national advocacy organization, *The Society for Yiddish and Yiddish Culture in Sweden*, and reviews the factors considered when establishing the technical and policy bases for the two registries' IDN support for Yiddish.

2. Yiddish orthography

There is significant variation in traditional Yiddish orthography and differing approaches are encountered in contemporary publication. All variants use the twenty-two letters of the basic Hebrew alphabet (five of which take different graphic forms when in the final position of a word), plus a number of diacritical marks, called “points”, that may be combined with one or more of the base letters. A given approach to Yiddish orthography can be characterized by the

number and disposition of the points that it uses, but a specific pointed letter will have the same meaning wherever it appears.

Authors of Yiddish text freely use the orthographic conventions of their personal preference without the differences causing any difficulty. Readers recognize common orthographic nuance with ease, or can identify unfamiliar detail after only brief reflection. There is no equivalent latitude when Yiddish letters are used in domain names. These are unique identifiers presented as mnemonics, without any literary context in which to assess variation. If a given letter can appear in a number of different ways, the risk for both inadvertent and deliberate confusion can become uncomfortably high. (This concern is not restricted to Yiddish, where its scope is less daunting than it is with Latin script. The number of diacritical marks that appear with Latin letters is far larger than the number of points used with Yiddish letters.)

The contention that Yiddish can comfortably sustain orthographic variation in general literary contexts (to which domain naming cannot be reckoned) has, nonetheless, been a subject of regular debate. Action toward a uniform Yiddish orthography was initiated in Eastern Europe during the early 20th century, targeted on standardizing school curricula. During the 1930s, the *Yidisher visnshaftlekher institut* (YIVO), began developing this into a written and spoken modern Standard Yiddish. The resulting character repertoire is codified in, *The Standardized Yiddish Orthography: Rules of Yiddish Spelling*, 6th ed., YIVO Institute for Jewish Research, New York, 1999, ISBN 0-914512-25-0 (in Yiddish with introductory material in English), and is commonly taken as normatively descriptive of the modern Standard Yiddish alphabet in contexts where that notion is deemed relevant. The spelling rules together with the prescribed character repertoire, are often abbreviated as “the SYO” and are referred to in that manner below. To whatever extent orthographic uniformity is taken to be a concern with contemporary Yiddish, it must be noted that the appearance of this language (or any other) in domain identifiers can involve constraints not encountered in common usage.

The SYO has been applied in several bilingual Yiddish dictionaries produced since its establishment. One example of particular relevance to the present initiative is, Lennart Kerbel & Peter David, *Jiddisch Svensk Jiddisch Ordbok*, Megilla-Förlaget, Stockholm, 2005, ISBN 91-89340-26-4. Given the international viability of the SYO repertoire and the wide availability of lexicographic and general documentation about it, it has been taken as the baseline reference for the determination of characters that can safely be permitted in IDNs derived from Yiddish.

3. IDN constraints

There is one detail in the IDN protocol that prevents the strict use of the SYO as an IDN character repertoire: it does not permit the final character in a label to be pointed. Since pointed letters frequently appear at the end of Yiddish words, both in the SYO and traditional orthographies, the use of words and names as IDN labels may require some degree of orthographic compromise.

The most thorough means for avoiding such difficulty would be to restrict the available repertoire from the outset, and forgo the use of pointing altogether. This alternative is not supported by the SYO, which lacks the unpointed form

of one of the requisite twenty-two Hebrew letters. However, the addition of that single base character (U+05E4 HEBREW LETTER PEY, using notation that is explained in the following section) is all that is needed to provide an IDN repertoire that does give the holders of Yiddish IDNs full flexibility to point them at individual discretion.

This anticipates not just a variety of orthographic preferences being reflected in IDNs, but also permits sequences of Hebrew letters to appear as IDN labels with no implicit link between them and any Yiddish vocabulary. This suggests that it might further be worth considering a shift entirely away from the present language-based approach, to a script-based character repertoire, satisfactory not just for Yiddish, but also for other languages written with Hebrew script. The most obvious of these is Hebrew itself. The issues attaching to a transition from the language-based tabulation provided in this document, to one based entirely on script, are discussed in Appendix 3.

4. Character Table

The value in the first column of the table gives the position of a character in the *Unicode Character Code Chart* <<http://www.unicode.org/charts/>>, with “U+” prefixed to its numerically assigned “code point” (in hexadecimal form). Two code points appearing in succession as “U+nnnn U+mddd” indicate combining characters that form a single displayed character, and the unprefixed “nnnn..ddd” indicates a continuous range of code points. The second column illustrates the corresponding characters. (Their correct display requires a font in which all are included and the use of software that renders it properly.) The third column provides the Unicode names for the characters. The fourth column lists the romanized Yiddish names for the characters as given by YIVO for an Anglophone audience. They are often spelled differently when appearing in Swedish or other non-English text.

Code Point	Symbol	Unicode Name	Yiddish Name	Note
U+05D0	א	HEBREW LETTER ALEF	<i>shtumer alef</i>	
U+05D0 U+05B7	אֿ	HEBREW LETTER ALEF with HEBREW POINT PATAH	<i>pasekh alef</i>	
U+05D0 U+05B8	אׁ	HEBREW LETTER ALEF with HEBREW POINT QAMATS	<i>komets alef</i>	
U+05D1	ב	HEBREW LETTER BET	<i>beys</i>	5.1
U+05D1 U+05BF	בֿ	HEBREW LETTER BET with HEBREW POINT RAFE	<i>veys</i>	5.2
U+05D2	ג	HEBREW LETTER GIMEL	<i>giml</i>	
U+05D3	ד	HEBREW LETTER DALET	<i>daled</i>	
U+05D4	ה	HEBREW LETTER HE	<i>hey</i>	
U+05D5	ו	HEBREW LETTER VAV	<i>vov</i>	
U+05D5 U+05BC	וּ	HEBREW LETTER VAV with HEBREW POINT DAGESH OR MAPIQ	<i>melupm vov</i>	5.3
U+05D6	ז	HEBREW LETTER ZAYIN	<i>zayen</i>	

U+05D7	ח	HEBREW LETTER HET	<i>khes</i>	
U+05D8	ט	HEBREW LETTER TET	<i>tes</i>	
U+05D9	י	HEBREW LETTER YOD	<i>yud</i>	
U+05D9 U+05B4	י	HEBREW LETTER YOD with HEBREW POINT HIRIQ	<i>khirik yud</i>	5.4
U+05DA	ך	HEBREW LETTER FINAL KAF	<i>langer khof</i>	5.5
U+05DB	כ	HEBREW LETTER KAF	<i>khof</i>	
U+05DB U+05BC	כ	HEBREW LETTER KAF with HEBREW POINT DAGESH OR MAPIQ	<i>kof</i>	5.2
U+05DC	ל	HEBREW LETTER LAMED	<i>lamed</i>	
U+05DD	ם	HEBREW LETTER FINAL MEM	<i>shlos mem</i>	5.5
U+05DE	מ	HEBREW LETTER MEM	<i>mem</i>	
U+05DF	ן	HEBREW LETTER FINAL NUN	<i>langer nun</i>	5.5
U+05E0	נ	HEBREW LETTER NUN	<i>nun</i>	
U+05E1	ס	HEBREW LETTER SAMEKH	<i>samekh</i>	
U+05E2	ע	HEBREW LETTER AYIN	<i>ayen</i>	
U+05E3	ף	HEBREW LETTER FINAL PE	<i>langer fey</i>	5.5
U+05E4	פ	HEBREW LETTER PE	<i>pey</i>	5.6
U+05E4 U+05BC	פ	HEBREW LETTER PE with HEBREW POINT DAGESH OR MAPIQ	<i>pey</i>	
U+05E4 U+05BF	פ	HEBREW LETTER PE with HEBREW POINT RAFE	<i>fey</i>	
U+05E5	ץ	HEBREW LETTER FINAL TSADI	<i>langer tsadek</i>	5.5
U+05E6	צ	HEBREW LETTER TSADI	<i>tsadek</i>	
U+05E7	ק	HEBREW LETTER QOF	<i>kuf</i>	
U+05E8	ר	HEBREW LETTER RESH	<i>reysh</i>	
U+05E9	ש	HEBREW LETTER SHIN	<i>shin</i>	
U+05E9 U+05C2	ש	HEBREW LETTER SHIN with HEBREW POINT SIN DOT	<i>sin</i>	5.2
U+05EA	ת	HEBREW LETTER TAV	<i>sof</i>	
U+05EA U+05BC	ת	HEBREW LETTER TAV with HEBREW POINT DAGESH OR MAPIQ	<i>tof</i>	5.2
U+05F2 U+05B7	יי	HEBREW LIGATURE YIDDISH DOUBLE YOD with HEBREW POINT PATAH	<i>pasekh tsvey yudn</i>	5.7

A label containing a character at any of the code points specified above may not contain any other characters, except for those in the following auxiliary table which may, however, not appear in the first or last positions of a label:

U+002D	-	HYPHEN-MINUS
0030..0039	0 - 9	DIGIT ZERO - DIGIT NINE

5. Discussion of the character repertoire

Both the SYO and traditional orthographies impose contextual constraints on the appearance and placement of several characters in Yiddish words, described in detail below. However, since there is no expectation that an IDN label will be a word (and many are deliberately not), there is no basis for determining the extent to which these word-based restrictions should, or even can, be applied here. With the exception of combining points, which may only attach to the characters they are explicitly associated with in the table, any permissible character may appear at any point in a string. The name holder is responsible for the orthographic rigor of a proper Yiddish word or name when used as an IDN label, including the positioning of final form characters.

As noted in Section 3 above, there is one overriding technical constraint imposed by the IDN protocol on the use of combining marks. This applies to all scripts written right to left, and prohibits any combining mark being placed on the final character in a label. The consequence of this for Yiddish IDNs is that labels requiring pointed characters in the final position are not possible, disallowing for example, the YIVO acronym — יײִװֿ .

The obvious alternatives are either to craft labels so as not to require final pointing, or to accept the compromise use of incongruous unpointed label-final characters. Enabling the latter option requires one modification to the SYO repertoire, which includes the entire Yiddish alphabet in unpointed form with the single exception of the *pey*. The SYO invariably points this with a *dagesh*, but since it is not reasonable to prohibit the *pey* at the end of a label (which would be of precisely the same consequence as barring a Latin “p” from that position) its unpointed form (U+05E4) has been included in the table.

There is no corresponding problem with the *fey*, which shares the same base character, but is pointed with a *rafe* when in word-initial and medial positions. Unlike the *pey*, however, the *fey* has a separate unpointed form when it is word final. This obviates the risk of confusing an unpointed final *pey* with a final *fey*, and otherwise treats both both letters similarly: pointed when in initial and medial positions, and unpointed in final position — פֿ, פֿ, פּ, פּ. When the revised protocol enters into effect and a pointed *pey* can appear anywhere in a label, a reversion to full SYO compliance will be also possible. Policies and procedures can then be devised for the optional conversion of labels registered in the interim.

Difficulty of another sort with pointing may be experienced in display environments that are not configured for the correct rendering of Yiddish characters. In such situations, application programs and fonts may be encountered that do not properly align points with their base characters (not just in domain names, but also in running text). Again, this is not a problem specific to Yiddish. The same concern pertains to the use of composite characters in many other scripts.

The unpointed *pey* will remain in the permitted repertoire, in any case, since it is necessary for the registration of Yiddish IDNs in fully unpointed form. This is a useful option for the holder of a Yiddish IDN who wishes to be certain that it is minimally subject to risk of incorrect display, will not confuse a user unfamiliar with pointing, or otherwise regards pointing as inappropriate in this

context. The additional character is also needed to support labels with varying degrees of pointing intermediate to full SYO detail, and for the application of other orthographic rules.

5.1 The *beys* is often written with a *dagesh* — ם. This is not permitted here because it would result in increased potential both for user confusion and display instability. However, since it is the only commonly encountered non-SYO form absent from the present repertoire, its addition will be considered if there is a clear indication of interest, and recognition of the concomitant difficulties, from prospective name holders.

5.2 This letter only appears in words of Semitic origin (Aramaic and Hebrew) and is not normally pointed in its original orthographic context. The point is therefore also frequently omitted in Yiddish text and, since it cannot appear in the final position of a label in any case, it is recommended that it not be used in IDNs. This consideration applies particularly to the *rafe*, as explained in Appendix 3.

5.3 The *melupm vov* is used for the unambiguous indication of a vocalic *vov* in a sequence of *vovn* and/or *yudn* with vocalic and consonantal components that might be read incorrectly. In cases where pointing is deliberately being avoided but where the intention is for the label to be read as a proper word, a *shtumer alef* can indicate the boundary between consonants and vowels, for example, as וואונדער instead of וונדער, and פראוון rather than פרוון. (This use of *alef* was standard practice prior to the YIVO reform, which sought to terminate it, but is nonetheless frequently encountered in contemporary writing.)

5.4 The *khirik yud* indicates a vocalic *yud* where it might otherwise be read as consonantal. The comments in the preceding subsection also apply to it.

5.5 The characters referencing this subsection are only used in word-final position. Transposed into IDNs, they would be used in the final position in a label, or at the end of a sequence of letters preceding a DIGIT or HYPHEN, or possibly in a manner equivalent to CamelCaps to indicate concatenation.

5.6 As noted above, the unpointed *pey* does not appear in the SYO but is included here because *pey* cannot reasonably be barred from appearing at the end of a label, and the IDN protocol does not permit its pointing when in that position. The additional character is also needed to enable the registration of entirely unpointed labels.

5.7 The pointed *pasekh tsvey yudn* — ם — requires the use of a single-character ligature (U+05F2, pointed with U+05B7) without the possibility of alternate representation by separately typing each of the two *yudn*. The latter mode of keyboard entry of *tsvey yudn* is well-established in user practice, and it is assumed that those who do so recognize the obligatory use of the ligature for the pointed form. It can, in any case, be expected that someone using an improvised alternative to the pointed ligature (such as the three character sequence *yud-pasekh-yud*), and therefore not getting the expected response, would subsequently try the ligated form.

5.8 Potential need is recognized for the punctuation marks *geresh* — ‘ — (U+05F3) and *gershayim* — ” — (U+05F4). These are currently excluded because they are not regularly indicated on Yiddish keyboards (or any others marked with Hebrew script), and an apostrophe (U+0027) and quotation mark (U+0022) are commonly substituted for them. This alternative is, however, not permissible in a domain name. Although the correct Unicode characters can be included in an IDN label, some erudition is required on the part of the user to ensure that a keyboarded transcription of that label is correct.

The *geresh* and the *gershayim* are used in Yiddish to indicate abbreviation and contraction, and cannot be seen as essential elements of an IDN repertoire. These marks do, however, have additional functions in Hebrew and other languages written with Hebrew script, and may be deemed necessary in the repertoires for those languages. If so, they would appear in the general Hebrew script table discussed in Appendix 3, and could then be added to the Yiddish language table here (assuming ongoing need for the parallel maintenance of both language-based and script-based tables).

6.0 Additional registry policies

In addition to the policies implicit in the preceding section, the following further details are in force:

The traditional Yiddish character repertoire includes three digraphs — װ ן ם. These are not listed separately in the character table and are available for inclusion in IDNs as simple sequences of the component characters. However, all three digraphs also appear in the Unicode chart as precomposed ligatures (U+05F0, U+05F1, U+05F2). The ligated and two-character forms are semantically identical and often display indistinguishably. Two IDN labels differing solely in the way the digraphs are represented therefore need to be treated as fully equivalent to each other. This precludes making both forms available for separate registration.

The .MUSEUM and .SE registries support the full SYO repertoire but restrict the use of ligatures to the single case of the *pasekh tsvey yudn*. (The SYO explicitly states that the digraphs are not separate letters of the Yiddish alphabet). It is understood that this may cause some initial confusion for users accustomed to the keyboard entry of the ligature forms of all the digraphs.

There would be no intrinsic difficulty in implementing an alternative procedure that equates every occurrence of a Yiddish ligature with the equivalent two-character digraph, and automatically generates two IDNs that are registered as a single “bundle” (or blocks the one from autonomous delegation). The inverse situation is, however, not as clear cut. It is possible, for example, for two consecutive *vovn* to be separated by a syllable boundary, thus not being correctly representable by a ligature. This is compounded by the availability of non-lexical labels to which the SYO rules are inapplicable. It is not realistically possible to rewrite a sequence, say, of five unpointed *vovn* using ligatures. It would likely be possible to devise a heuristic that could determine if a label adheres to the SYO, and then identify any digraphs it contains and generate the alternate ligated representations. This would, however, only result in heightened confusion with labels containing two-character digraphs that cannot be unambiguous bundled with ligated equivalents.

The same basic consideration applies to IDNs that differ solely by the appearance of pointed letters. Here again, the situation with the Yiddish alphabet does not differ from that with any other script using ligated and diacritically-marked characters. Language-specific considerations will, for example, determine whether or not there is an equivalence between an “ae” digraph and an “æ” ligature, or if either is equivalent to an unlauded “ä”, or if so, can acceptably also be indicated with an unmarked “a”. Even where lexicographic rules might be contemplated for dealing with this, their automated implementation would easily be stymied by differences between the representation of proper names and dictionary words: “encyclopaedia” and “encyclöpædia” could be taken as identical, but “mueller” and “müller” cannot, and “öresund” and “øresund” might be argued either way.

Corresponding difficulties are encountered when differentiating a pointed *alef* from the unmarked base character, and with the similar distinctions between pointed and unpointed *yudn* and *vovn*. There are, for example, situations where a *pasekh alef* and an unpointed *alef* can reasonably be seen as variants of the same character, with a corresponding potential equivalence between the *kometts alef* and the unpointed *alef*. The two pointed forms are, however, completely distinct from each other and it is not possible to associate the unpointed character preferentially with either. Nor is there any algorithmic way to determine if an unpointed *alef* is deliberately being used as an alternative to a character that the SYO requires to be pointed.

The holder of an IDN containing pointed Yiddish characters who also wishes to hold the unpointed equivalent of the same name needs to register each separately. This is analogous to the prospective holder of the Latin label “lättöl” being free to register “lattol”, without either imposing any restriction on the availability of the other, or on any further variants using the more than twenty diacritically marked forms of the base “a” in the Unicode chart, or the similar number of marked forms of “o”.

The number of diacritical permutations of a Yiddish label will, however, typically be significantly smaller than that of a Latin label. In cases where there is an objective basis for determining their equivalence, a registry may therefore reserve all forms of the same Yiddish label for the holder of the first registered variation, or simply bundle them outright. (Registration requests in .SE are processed automatically and every marked variant of a label otherwise consisting of the same sequence of base characters is treated as unique. In .MUSEUM, all requested names are vetted, and what are deemed to be variants of the same label are bundled or blocked on a case-to-case basis.)

Finally, it should be noted that the pointed Yiddish letters in the character table all exist in precomposed form at separate positions in the Unicode chart. Unlike the ligatures, these are remapped to the combining forms by the IDN protocol. Both the one- and two-code point forms can therefore be used as input in a request for DNS resolution, although only the latter form can actually be registered. It must also be noted, however, that the impending revision to the IDN protocol is expected to eliminate all such remapping.

Appendix 1 Contextual rules

The restrictions on combining characters in the main table are restated here as algorithmically enforceable rules. The order of appearance refers to the positions of two immediately adjacent code points in a Unicode string as it is submitted to the ToASCII conversion operation specified in the IDNA protocol.

Rule	Code Point	may only appear	Code Point
1	U+05B4 (HIRIQ)	following	U+05D9 (YOD)
2	U+05B7 (PATAH)	following or following	U+05D0 (ALEF) U+05F2 (DOUBLE YOD) <i>note the final rule below</i>
3	U+05B8 (QAMATS)	following	U+05D0 (ALEF)
4	U+05BC (DAGESH)	following or following or following or following	U+05D5 (VAV) U+05DB (KAF) U+05E4 (PE) U+05EA (TAV)
5	U+05BF (RAFE)	following or following	U+05D1 (BET) U+05E4 (PE)
6	U+05C2 (SIN DOT)	following	U+05E9 (SHIN)
7	U+05F2 (DOUBLE YOD)	if followed by	U+05B7 (PATAH)

Appendix 2 Test labels

The rules in Appendix 1 may be tested with the following labels:

Rule	accepts	rejects
1	אָאָאָ xn-cdb9cb8a	אָאָאָ xn-cdb9cb6h
2	אָאָאָ xn-fdb3cab אָאָאָ xn-fdb3cb8k	אָאָאָ xn-fdb3cb8a
3	אָאָאָ xn-gdb1cab	אָאָאָ xn-gdb1cb2f
4	אָאָאָ xn-kdb3bbs אָאָאָ xn-kdb3bb6b אָאָאָ xn-kdb3bb2f אָאָאָ xn-kdb3bb6h	אָאָאָ xn-kdb3bb8g
5	אָאָאָ xn-ndb7abc אָאָאָ xn-ndb7ab2f	אָאָאָ xn-ndb7ab6b
6	אָאָאָ xn-qdb1ab2h	אָאָאָ xn-qdb1ab8c
7	אָאָאָ xn-fdb3cb8k	אָאָאָ xn-cdb9cb8k

Appendix 3 Script table

A3.1 Basic principles

This document is focused on the development of an IDN character repertoire and policies for a single language. Many details of the Yiddish writing system are specific to it but one of its key components, the Hebrew script, will figure with equal prominence in the tabulation of an IDN repertoire for any other language with which it is written. Domain names do not, however, always represent dictionary words, and nothing intrinsic to a label indicates the language, if any, it is intended to represent.

Consideration of the orthographic detail of a language is an obvious and necessary initial step in preparing for its use in IDNs. Further attention must, however, be given to the way its script is used for writing other languages that are, or are likely to be, similarly reflected in IDNs. Without such action, the way a language-specific detail is treated by one registry could prove to be at odds with the way the same detail is handled by another registry supporting some other language also written the same script, with possibly significant confusion within the broader Internet user community as a consequence. At least one such issue attaches to the present effort with Yiddish, as described in the following section.

Fundamental differences between writing systems give rise to many situations where a given element of a script is used in differing manners from language to language, with potential for confusing anyone without a detailed understanding of the variation in orthographic practices. This must be accepted in IDNs precisely as it is in other contexts where written language appears. Nonetheless, it will benefit the user community if efforts are made to keep the risk of inadvertent difficulty to an absolute minimum. The prototypic contribution to the development of script-based policies serving multiple language communities resulted in the *Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese, and Korean* <<http://www.ietf.org/rfc/rfc3743.txt>>. Similar initiatives are currently being conducted by language communities sharing other scripts, for example, the *Arabic Script IDN Working Group (ASIWG)* <<http://lists.irnic.ir/mailman/listinfo/idna-arabicscript>>.

A3.2 Specimen issue

The Hebrew script provides the basis for the writing systems of a number of languages. Text in three of them — Hebrew, Judeo-Spanish (also called Ladino or Djudezmo), and Yiddish — can easily be located on the Internet, and their writing systems are documented in readily available sources. As a rough indicator of their suitability for illustrating specimen issues that would figure in any effort toward developing an IDN repertoire based on the Hebrew script and sharable across the language boundaries, it can be noted that separate versions of the Wikipedia are maintained in each of these three languages (with no others currently using Hebrew script, but without in any way discounting the possibility of their future appearance).

Hebrew, Judeo-Spanish, and Yiddish are all written with the same basic twenty-two letter alphabet, plus the five separate final forms. They all also represent

phonetic detail that this alphabet does not directly indicate with various systems of pointing, differing in the extent and contexts in which they are applied. There is also some degree of internal variation in the orthographic and typographic traditions of each of the languages.

Issues relating to pointing are unlikely to be a concern for Hebrew, which can dispense with pointing entirely in a context such as IDN. However, Hebrew uses the punctuation mark *geresh* to indicate one specific form of the phonetic variation that Yiddish indicates with pointing (“hard” vs. “soft” pronunciation of consonants, as described below), and uses the punctuation mark *gershayim* in the penultimate position in a sequence of letters to indicate that it is an acronym.

Greater difficulty is presented by Judeo-Spanish, which is commonly printed with two different typefaces, and indicates the same phonetic contrast by the use of pointing in the one of them, and the *geresh* in the other. Other forms of pointing are, however, less common. The points included in the current Yiddish repertoire should prove adequate for Judeo-Spanish, and they are not, the addition of further points would be easily accomplished.

The notation of the distinction between pointed consonants that share the same base character can be illustrated with the four character pairs in the Yiddish table, *beys/veys*, *kof/khof*, *pey/fey*, and *tof/sof*. Each pair shares a single base character that has both a hard and a soft pronunciation, indicated with a *dagesh* and a *rafe*, respectively. The SYO does not apply these marks consistently across all four character pairs, and text that otherwise conforms to it therefore frequently omits the *rafe* from the *fey*, in harmonization with its unpointed final form, and makes the contrastive distinction from a *pey* solely with a *dagesh* in the latter — פּ פ. The similar avoidance of the *rafe* and preferential use of the *dagesh*, is a common alternative for the distinction between *veys* and *beys* — בּ ב (see note 5.1).

Hebrew makes a corresponding hard/soft distinction, but between other base letters such as *gimel*, *zayin*, and *tsadi*. and with the use of another orthographic device; the soft pronunciation is indicated by following the base letter with a *geresh* — זײַ זײ. Judeo-Spanish also makes this distinction for yet another set of characters, and indicates it with a *geresh* in text set with square Hebrew letters (used for all the exemplification thus far). This typeface is commonly used for banner text and headings. Running text, however, is normally set with a semi-cursive typeface called Rashi — א ב ג ד (the first four letters of the alphabet). When it is used, the soft pronunciation is normally indicated by a point above the base letter — אױ אױ — but this is a *varika* (U+FB1E HEBREW POINT JUDEO-SPANISH VARIKA — ױ), not a *rafe*.

There is an annotation in the *Unicode Character Code Chart* stating that the *varika* is “glyph variant” of the *rafe*. This may be a reasonable assertion, but commonly encountered Rashi fonts treat the two marks separately, displaying the horizontal bar for U+05BF and not replacing it with the curved mark, which is only displayed when U+FB1E is explicitly indicated. (The two examples of pointed Rashi letters in the preceding paragraph are, however, presented with a locally-modified version of a widely-used Rashi font, that does make the glyph substitution.)

The situation with Judeo-Spanish is further complicated by it being the one of the three languages with the greatest degree of orthographic variation. Although the typographic practice just described is by far the most common, the *geresh* is sometimes seen in text set in Rashi, and the *rafe* in text set in square Hebrew. A romanized orthography has also been widely adopted by the language community, and both scripts appear in contemporary texts. The trend has been toward increasing romanization, but it remains to be seen if developments such as IDN will bolster efforts to maintain the traditional Hebraic orthography. In any case, the Judeo-Spanish version of the Wikipedia is maintained in both scripts in parallel <<http://lad.wikipedia.org/>>. For the present discussion, it will suffice to note that its Hebrew-script facet uses the *geresh* to indicate the soft pronunciation of a consonant, rather than the *varika* or *rafe*, as illustrated by the native representation of name of the language — גוֹדְיָאוֹ-אִסְפַּאֲנִיּוֹל — and is thus free from any concern with glyph substitution. (See also <<http://www.omniglot.com/writing/ladino.htm>>, where the *geresh* is applied to Rashi letters, as is also done in, Marie-Christine Varol, *Manual of Judeo-Spanish*, University of Maryland Press and L'Asiathèque, 2008, ISBN 978-1-934309-19-3 and ISBN 978-2-915255-75-1.)

The problem with the incorrect display of combining marks mentioned in the next to the last paragraph of Section 5 above, is certain to be observed in the typical working environment when presented with the full range of Judeo-Spanish letters that can be pointed with a *varika* or *rafe*, and may even cause difficulty for the specialist user. The use of the *geresh* in domain names is therefore clearly the more advisable alternative from the perspective of consistent display.

It remains necessary to recognize that a non-obvious key combination is normally required for the entry of a *geresh*, and that the common expedient of substituting an apostrophe is not a viable alternative. The extent to which the utility of including the *geresh* in the IDN repertoire outweighs the keyboard issue has yet to be determined. The issues under consideration, together with those attaching to the *gershayim* and other relevant details, are discussed at <<http://unicode.org/mail-arch/unicode-ml/y2005-m02/0168.html>>, and through the chain of **Reply:** links initiated there.

Since there is no alternative to the *geresh* and *gershayim* in Hebrew text, the sole consideration is whether their availability for IDNs is worth the trade-off on the keyboard side. This assessment needs to be made prior to any conclusive consideration of their appearance in a shared script-based repertoire. A clearly applicable source of guidance on this would be the policies adopted for Hebrew IDNs in the Israeli national TLD, .IL. Such support has, however, not yet been deployed in that domain. Pending the decision either to include the *geresh* in, or to exclude it from, the IDN repertoire deemed necessary there for the Hebrew language, two alternatives remain for a cross-language script-based repertoire.

If the *geresh* is available, the obvious choice would be to use it in both Hebrew and Judeo-Spanish IDNs to indicate the soft pronunciation of the consonants that can be marked in that manner. The simultaneous availability of the *rafe* for the same purpose in Yiddish would provide a second alternative for Judeo-Spanish. A contextual rule preventing the simultaneous appearance of a *geresh* and a *rafe* in the same label would therefore be highly advisable. A

policy would also be needed for dealing with the equivalence between the *rafe* and the *varika*, for example, by the automatic bundling of both forms of any label that is requested using either. It would also be necessary to differentiate between situations where characters marked with the *rafe* and the *geresh* are to be seen as variants of each other, and where they are not.

If the *geresh* is not available, the use of the *rafe* as it appears in the current Yiddish language repertoire can be extended to the additional consonants needed for Judeo-Spanish, which to a reasonable initial approximation are those already in the Yiddish table, plus those that would be followed by a *geresh* in Hebrew if that convention had been supported. However, due to the display instability that the extended use of the *rafe* would entail, it remains the potentially more problematic of the two alternatives.

The *geresh* and *gershayim* are also used in the notation of Hebrew numbers <<http://smontagu.org/writings/HebrewNumbers.html>>. If these are supported as IDN labels, the contextual rules presented at the end of this appendix will require appropriate modification. The two marks are also used to indicate abbreviation and contraction but, as the punctuated indication of such constructs is not supported in conventional labels, particular justification may be needed for including it here. If it is supported, there would be little basis for imposing any contextual rules on either the *geresh* or *gershayim*.

On the basis of all that has been considered in this appendix, the Yiddish language table in the main body of the document could be reframed as a Hebrew script table, as follows. It is being presented to seed further discussion of the coordinated preparation of a character repertoire that is adequate for the representation of IDNs derived from multiple languages written using the Hebrew script, and is not a statement of intent for near-term modification of IDN policies in either .SE or .MUSEUM.

A3.3 Script table

Code Point	Symbol	Unicode Name	Note
U+05B4	.	HEBREW POINT HIRIQ	rule applies
U+05B7	-	HEBREW POINT PATAH	rule applies
U+05B8	ֿ	HEBREW POINT QAMATS	rule applies
U+05BC	װ	HEBREW POINT DAGESH OR MAPIQ	rule applies
U+05BF	ױ	HEBREW POINT RAFE	rule applies
U+05C2	ײ	HEBREW POINT SIN DOT	rule applies
U+05D0	א	HEBREW LETTER ALEF	
U+05D1	ב	HEBREW LETTER BET	
U+05D2	ג	HEBREW LETTER GIMEL	
U+05D3	ד	HEBREW LETTER DALET	
U+05D4	ה	HEBREW LETTER HE	
U+05D5	ו	HEBREW LETTER VAV	

U+05D6	ז	HEBREW LETTER ZAYIN	
U+05D7	ח	HEBREW LETTER HET	
U+05D8	ט	HEBREW LETTER TET	
U+05D9	י	HEBREW LETTER YOD	
U+05DA	ך	HEBREW LETTER FINAL KAF	
U+05DB	כ	HEBREW LETTER KAF	
U+05DC	ל	HEBREW LETTER LAMED	
U+05DD	ם	HEBREW LETTER FINAL MEM	
U+05DE	מ	HEBREW LETTER MEM	
U+05DF	ן	HEBREW LETTER FINAL NUN	
U+05E0	נ	HEBREW LETTER NUN	
U+05E1	ס	HEBREW LETTER SAMEKH	
U+05E2	ע	HEBREW LETTER AYIN	
U+05E3	ף	HEBREW LETTER FINAL PE	
U+05E4	פ	HEBREW LETTER PE	
U+05E5	ץ	HEBREW LETTER FINAL TSADI	
U+05E6	צ	HEBREW LETTER TSADI	
U+05E7	ק	HEBREW LETTER QOF	
U+05E8	ר	HEBREW LETTER RESH	
U+05E9	ש	HEBREW LETTER SHIN	
U+05EA	ת	HEBREW LETTER TAV	
U+05F2	ײ	HEBREW LIGATURE YIDDISH DOUBLE YOD	rule applies
U+05F3	׳	HEBREW PUNCTUATION GERESH	rule applies
U+05F4	״	HEBREW PUNCTUATION GERSHAYIM	rule applies

Notes:

The comment “rule applies” refers to the contextual restrictions imposed on the use of the indicated code point as stated in Appendix 1, plus the following additional constraints:

A. The U+05F3 GERESH is restricted to the position immediately following the character that it modifies. A full list of permissible such code points needs to be added to the rule table before this is implemented, with further extension if Hebrew numbers are also supported.

B. The U+05F4 GERSHAYIM may only appear in the penultimate position in a label.

C. The U+05BF RAPE and the U+05F3 GERESH may not both appear in the same label.

Appendix 4 Basic concepts

A domain name is the sequence of characters to the right of the @-sign in an e-mail address (username@**example.test**), or between the second and third slashes in a Web resource identifier ([http://**example.test**/filename](http://example.test/filename)). It consists of “labels” separated by “dots”, with each label designating a level in the Domain Name System. In the second-level domain **example.test** (with both the labels and the dot being pronounced — “example dot test”) the top-level domain is **test**, (also commonly read with the preceding dot — “dot test”) and the second-level domain is **example**. This can be extended on successively lower levels as **fourthlevel.thirdlevel.secondlevel.toplevel**.

In e-mail addressing and Web resource identification, together with numerous other applications, the characters available for inclusion in a domain name are restricted to the twenty-six letters of the basic Latin alphabet “a-z”, the ten digits “0-9”, and the hyphen “-”. An Internationalized Domain Name appears to contain other characters, but this is done by encoding each IDN label with a sequence of characters taken from the restricted repertoire. For example, the Yiddish translation of **example.test** is **ביישפיל.טעסט**. It is normally displayed in this way as an IDN, but it is actually stored in the Domain Name System in its encrypted form, which is **xn--fdbk5d8ap9b8a8d.xn--deba0ad**.

Software that understands this scheme displays the additional characters as a user expects to see them, transparently encoding and decoding them as required (sometimes needing explicit configuration before displaying the unencoded forms of characters that do not otherwise appear in the locale to which the software is set). There is a live test site at <http://ביישפיל.טעסט/>. Along with ten further resources labeled with non-Latin equivalents to **example.test**, it is available to enable users to assess the IDN-compliance of their individual working environments. An English language gateway to the evaluation facility is located at [<http://idn.icann.org/>](http://idn.icann.org/).

Version 4.5
rev 17 January 2009

Author’s address:

Cary Karp
Swedish Museum of Natural History
Frescativägen 40
SE-11418 Stockholm