

# 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” — names containing at least one character that is not a basic Latin letter “a-z”, a digit “0-9”, or a hyphen “-” <<http://www.faqs.org/rfcs/rfc3490.htm>>) derived from all languages with official status in that country. 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 provided 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 <[http://www.verva.se/web/t/Page\\_\\_\\_1100.aspx](http://www.verva.se/web/t/Page___1100.aspx)>. The exception is Yiddish, which is written with an alphabet based on Hebrew script.

The expanded character repertoire in those guidelines supports Latin-script IDNs without need for explicit reference to any of the targeted languages. However, descriptions of language-based subsets of the repertoire may be useful for informational purposes. One such statement can be seen in the *IANA Repository of TLD IDN Practices* where the characters specifically needed for the Swedish language are documented 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 reviews the factors considered when establishing the technical and policy bases for the two registries' IDN support for Yiddish, and the document for the IANA Repository will be extracted from it.

## 2. Yiddish orthography

There is significant variation in traditional Yiddish orthography and differing approaches are encountered in contemporary publication. However, all variants use the twenty-two characters of the basic Hebrew script (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 characters. A given approach to Yiddish orthography can be characterized by the number and disposition of the points that it uses, but a specific pointed character will have the same meaning wherever it appears.

Authors of Yiddish text can freely use the orthographic conventions of their personal preference without the differences causing any difficulty. Most readers recognize common orthographic nuance with ease or can identify unfamiliar detail after only brief reflection. There is no equivalent latitude when Yiddish characters 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. Both the orthographic rules and the prescribed character repertoire are 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 used 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 general international viability of the SYO repertoire and the wide of 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. The Yiddish Alphabet**

The table in the following section presents the full SYO repertoire with the single further inclusion an unpointed *pey*. The reasons for this are discussed together with other considerations pertaining to the use of specific characters, in sections 5 and 6 below. In brief summary, it can be noted that the unpointed *pey* enables the use of the all letters in the Yiddish alphabet without any pointing, both where the name holder explicitly prefers this alternative, or wishes to register the unpointed and pointed forms of a name in parallel. This also supports IDNs in other languages commonly written with unpointed Hebrew script, most obviously Hebrew itself.

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 the numerically assigned “code point” (in hexadecimal form). Two code points appearing in succession as “U+nnnn U+mmmm” indicate combining characters that form a single displayed character, and the unprefixed “nnnn..mmm” 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.

#### 4. Character Table

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	אָ	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>	
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	וֿ	HEBREW LETTER VAV with HEBREW POINT DAGESH OR MAPIQ	<i>melupm vov</i>	5.2
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	יֿ	HEBREW LETTER YOD with HEBREW POINT HIRIQ	<i>khirik yud</i>	5.3
U+05DA	ך	HEBREW LETTER FINAL KAF	<i>langer khof</i>	5.4
U+05DB	כ	HEBREW LETTER KAF	<i>khof</i>	
U+05DB U+05BC	כֿ	HEBREW LETTER KAF with HEBREW POINT DAGESH OR MAPIQ	<i>kof</i>	
U+05DC	ל	HEBREW LETTER LAMED	<i>lamed</i>	

U+05DD	ם	HEBREW LETTER FINAL MEM	<i>shlos mem</i>	5.4
U+05DE	מ	HEBREW LETTER MEM	<i>mem</i>	
U+05DF	ן	HEBREW LETTER FINAL NUN	<i>langer nun</i>	5.4
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>	5.4
U+05E4	פ	HEBREW LETTER PE	<i>pey</i>	5.5
U+05E4 U+05BC	פּ	HEBREW LETTER PE with HEBREW POINT DAGESH OR MAPIQ	<i>pey</i>	
U+05E4 U+05BF	פֿ	HEBREW LETTER PE with HEBREW POINT RAFE	<i>fey</i>	
U+05E5	ץ	HEBREW LETTER FINAL TSADI	<i>langer tsadek</i>	5.4
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	שׁ	HEBREW LETTER SHIN with HEBREW POINT SIN DOT	<i>sin</i>	
U+05EA U+05BC	תּ	HEBREW LETTER TAV with HEBREW POINT DAGESH OR MAPIQ	<i>tof</i>	
U+05EA	ת	HEBREW LETTER TAV	<i>sof</i>	
U+05F2 U+05B7	ײ	HEBREW LIGATURE YIDDISH DOUBLE YOD with HEBREW POINT PATAH	<i>pasekh tsvey yudn</i>	5.6

In addition to the characters and code points specified above, a Yiddish IDN label may include the following characters and code points, but not in the first or last positions:

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

## 5. Discussion of the character repertoire

The SYO imposes 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, 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.

There is one overriding technical constraint imposed by the IDN protocol on the use of combining marks in all scripts written right to left — none may be attached to the final character in a label. (This restriction is expected to be eliminated in a revision of the protocol that should enter into effect in 2009.) The consequence of this for Yiddish IDNs is that labels requiring pointed characters in the final position are not currently possible (disallowing, for example, the YIVO acronym — ץײױװ).

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*, but implies a need for abstaining from pointing any other *pey* in the same label. Since the *pey* includes a *dagesh* in even the most frugally pointed Yiddish orthography, with the unpointed form normally being read as a *fey*, if confusion is to be minimized in a pointed label ending with a *pey*, it may be advisable to indicate any *fey* in the same label with a *rafe*.

Difficulty of another sort with pointing may be experienced in display environments that have not been deliberately configured for the correct rendering of Yiddish characters. In such situations, software applications 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.

As a further effect of allowing the unpointed *pey*, it is possible to register Yiddish IDNs in fully unpointed form. This can be 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 tabulated repertoire also supports labels with varying degrees of pointing intermediate to full SYO detail. Where no orthographic or typographic compromise is acceptable, the Yiddish attribute of a label can be indicated with a sequence of letters that is seen as a specifically Yiddish word, or as clearly derived from one, without requiring pointing to be immediately recognizable as such.

**5.1** The *beys* is often written with a *dagesh* — בּ. 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 and the user community.

**5.2** 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 correct word, a *shtumer alef* can indicate the boundary between consonants and vowels, for example, as וואַנדרער instead of ווינדער, and פּראַווין rather than פּרוין. (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.)

**5.3** 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.

**5.4** The characters referenced to 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.

**5.5** As noted above, the unpointed *pey* does not appear in the SYO but is included here because *pey* cannot be barred from appearing at the end of a label, and the IDN protocol does not permit its pointing when in that position.

**5.6** The pointed *pasekh tsvey yudn* — ײ — 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*. Since the latter mode of keyboard entry of *tsvey yudn* is well-established in user practice, the obligatory use of the pointed form of the ligature must be assumed to be known to users. It can also be expected that someone using an improvised alternative to the pointed ligature (such as the three character sequence *yud-pasekh-yud*) in a resolution request and therefore not getting the intended response, would subsequently try the two-character form.

## **6.0 Registry policies**

The traditional Yiddish character repertoire includes three digraphs — ײ ױ װ. 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”. 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 a digraph and not correctly representable by a ligature. This is compounded beyond utility 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 “encyclopædia” 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 a graphically simplified alternative to what the SYO would require to be pointed.

The holder of an IDN containing pointed Yiddish characters who also wishes to hold the unpointed equivalent of the same name may be required to register each separately (which is .SE policy). This is analogous to the prospective holder of the Latin label “lättöl” being free to register the undecorated correlate “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”. However, since the number of diacritical permutations of a Yiddish label will typically be significantly smaller than that of a Latin label, a registry may reserve all forms of the same Yiddish label for the holder of the first registered variation, or simply bundle them outright.

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 likely to eliminate all such remapping, leaving it to application software to provide that functionality.

## Appendix 1

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

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

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

Version 3.5  
20 August 2008

Author's address:

Cary Karp  
Swedish Museum of Natural History  
Frescativägen 40  
SE-10405 Stockholm