**Proposed Update** Unicode® Technical Standard #46

UNICODE IDNA COMPATIBILITY PROCESSING

<table>
<thead>
<tr>
<th>Version</th>
<th>10.0.0 (draft 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editors</td>
<td>Mark Davis (<a href="mailto:markdavis@google.com">markdavis@google.com</a>), Michel Suignard (<a href="mailto:michel@suignard.com">michel@suignard.com</a>)</td>
</tr>
<tr>
<td>Date</td>
<td>2017–02–23</td>
</tr>
<tr>
<td>This Version</td>
<td><a href="http://www.unicode.org/reports/tr46/tr46-18.html">http://www.unicode.org/reports/tr46/tr46-18.html</a></td>
</tr>
<tr>
<td>Previous Version</td>
<td><a href="http://www.unicode.org/reports/tr46/tr46-17.html">http://www.unicode.org/reports/tr46/tr46-17.html</a></td>
</tr>
<tr>
<td>Latest Version</td>
<td><a href="http://www.unicode.org/reports/tr46/">http://www.unicode.org/reports/tr46/</a></td>
</tr>
<tr>
<td>Latest Proposed Update</td>
<td><a href="http://www.unicode.org/reports/tr46/proposed.html">http://www.unicode.org/reports/tr46/proposed.html</a></td>
</tr>
<tr>
<td>Revision</td>
<td>18</td>
</tr>
</tbody>
</table>

**Summary**

Client software, such as browsers and emailers, faces a difficult transition from the version of international domain names approved in 2003 (IDNA2003), to the revision approved in 2010 (IDNA2008). The specification in this document provides a mechanism that minimizes the impact of this transition for client software, allowing client software to access domains that are valid under either system.

The specification provides two main features: One is a comprehensive mapping to support current user expectations for casing and other variants of domain names. Such a mapping is allowed by IDNA2008. The second is a compatibility mechanism that supports the existing domain names that were allowed under IDNA2003. This second feature is intended to improve client behavior during the transitional period.

**Status**

This is a draft document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.

http://www.unicode.org/reports/tr46/tr46-18.html
A Unicode Technical Standard (UTS) is an independent specification. Conformance to the Unicode Standard does not imply conformance to any UTS.

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this document is found in the References. For the latest version of the Unicode Standard, see [Unicode]. For a list of current Unicode Technical Reports, see [Reports]. For more information about versions of the Unicode Standard, see [Versions].

Contents

1 Introduction
   1.1 IDNA2003
   1.2 IDNA2008
   1.3 Transition Considerations
      1.3.1 Mapping
      1.3.2 Deviations
      Table 1. Deviation Characters
2 Unicode IDNA Compatibility Processing
   2.1 Display of Internationalized Domain Names
   2.2 Registries
   2.3 Notation
3 Conformance
   3.1 STD3 Rules
4 Processing
   4.1 Validity Criteria
      4.1.1 UseSTD3ASCIIIRules
      4.1.2 Right-to-Left Scripts
   4.2 ToASCII
   4.3 ToUnicode
   4.4 Preprocessing for IDNA2008
   4.5 Implementation Notes
      Table 2. Examples of Processing
5 IDNA Mapping Table
   Table 2b. Data File Fields
6 Mapping Table Derivation
   Step 1: Define a base mapping
   Step 2: Specify the base valid set
      Table 3. Base Valid Set
   Step 3: Specify the base exclusion set
   Step 4: Specify the deviation set
   Step 5: Specify grandfathered changes
   Step 6: Produce the initial status and mapping values
   Step 7: Produce the final status and mapping values
7 IDNA Comparison
   Table 4. IDNA Comparisons
   7.1 Implications for Implementers
8 Conformance Testing
   8.1 Format
Acknowledgments
References
Modifications
1 Introduction

One of the great strengths of domain names is universality. The URL http://Apple.com goes to Apple’s website from anywhere in the world, using any browser. The email address markdavis@google.com can be used to send email to an editor of this specification from anywhere in the world, using any emailer.

Initially, domain names were restricted to ASCII characters. This was a significant burden on people using other characters. Suppose, for example, that the domain name system had been invented by Greeks, and one could only use Greek characters in URLs. Rather than apple.com, one would have to write something like άττιπε.κομ. An English speaker would not only have to be acquainted with Greek characters, but would also have to pick those Greek letters that would correspond to the desired English letters. One would have to guess at the spelling of particular words, because there are not exact matches between scripts.

Most of the world’s population faced this situation until recently, because their languages use non-ASCII characters. A system was introduced in 2003 for internationalized domain names (IDN). This system is called Internationalizing Domain Names for Applications, or IDNA2003 for short. This mechanism supports IDNs by means of a client software transformation into a format known as Punycode. A revision of IDNA was approved in 2010 (IDNA2008). This revision has a number of incompatibilities with IDNA2003.

The incompatibilities force implementers of client software, such as browsers and emailers, to face difficult choices during the transition period as registries shift from IDNA2003 to IDNA2008. This document specifies a mechanism that minimizes the impact of this transition for client software, allowing client software to access domains that are valid under either system.

The specification provides two main features. The first is a comprehensive mapping to support current user expectations for casing and other variants of domain names. Such a mapping is allowed by IDNA2008. The second feature is a compatibility mechanism that supports the existing domain names that were allowed under IDNA2003. This second feature is intended to improve client behavior during the transitional period. This specification contains both normative and informative material. Only the conformance clauses and the text that they directly or indirectly reference are considered normative.

1.1 IDNA2003

The series of RFCs collectively known as IDNA2003 [IDNA2003] allows domain names to contain non-ASCII Unicode characters, which includes not only the characters needed for Latin-script languages other than English (such as Á, Ó, or ß), but also different scripts, such as Greek, Cyrillic, Tamil, or Korean. An internationalized domain name such as Bücher.de can then be used in an "internationalized" URL, called an IRI, such as http://Bücher.de#titel.

The IDNA mechanism for allowing non-ASCII Unicode characters in domain names involves applying the following steps to each label in the domain name that contains Unicode characters:

1. Transforming (mapping) a Unicode string to remove case and other variant differences.
2. Checking the resulting string for validity, according to certain rules.
3. Transforming the Unicode characters into a DNS-compatible ASCII string using a specialized encoding called Punycode [RFC3492].

For example, typing the IRI http://Bücher.de into the address bar of any modern browser goes to a corresponding site, even though the "ü" is not an ASCII character. This works because the IDN in that IRI resolves to the Punycode string which is actually stored by the DNS for that site. Similarly, when a browser interprets a web page containing a link such as <a href="http://Bücher.de">, the appropriate site is reached. (In this document, phrases such as "a browser interprets" refer to domain names parsed out of IRIs entered in an address bar as well as to those contained in links internal to HTML text.)

In the case of IDN Bücher.de, the Punycode value actually used for the domain names on the wire is xn--bcher-kva.de. The Punycode version is also typically transformed back into Unicode form for display. The resulting display string will be a string which has already been mapped according to the IDNA2003 rules. This example results in a display string for the IRI that has been casefolded to lowercase:


A major limitation of IDNA2003 is its restriction to the repertoire of characters in Unicode 3.2, which means that some modern languages are excluded or not fully supported. Furthermore, within the constraints of IDNA2003, there is no simple way to extend the repertoire. IDNA2003 also does not make it clear to users of registries exactly which string they are registering for a domain name (between Bücher.de and buecher.de, for example).

1.2 IDNA2008

In early 2010, a new version of IDNA was approved. Like IDNA2003, this version consists of a collection of RFCs and is called IDNA2008 [IDNA2008]. IDNA2008 is intended to solve the major problems in IDNA2003. It extends the valid repertoire of characters in domain names, and establishes an automatic process for updating to future versions of the Unicode Standard. Furthermore, it defines the concept of a valid domain name clearly, so that registrants understand exactly what domain name string is being registered.

Processing in IDNA2008 is identical to IDNA2003 for many common domain names. Both IDNA2003 and IDNA2008 transform a Unicode domain name in an IRI (like http://öbb.at) to the Punycode version (like http://xn--bb-eka.at). However, IDNA2008 does not maintain strict backward compatibility with IDNA2003. The main differences are:

- **Additions.** Some IDNs are invalid in IDNA2003, but valid in IDNA2008.
- **Subtractions.** Some IDNs are valid in IDNA2003, but invalid in IDNA2008.
- **Deviations.** Some IDNs are valid in both, but resolve to different destinations.

For more details, see Section 7, IDNA Comparison.

1.3 Transition Considerations

The differences between IDNA2008 and IDNA2003 may cause interoperability and security problems. They affect extremely common characters, such as all uppercase characters, all halfwidth or fullwidth characters (commonly used in Japan, China, and Korea), and certain other characters like the German eszett (U+00DF ß LATIN SMALL LETTER SHARP S) and Greek final sigma (U+03C2 ç GREEK SMALL LETTER FINAL SIGMA).
1.3.1 Mapping

IDNA2003 requires a mapping phase, which maps ÖBB.at to öbb.at, for example. Mapping typically involves mapping uppercase characters to their lowercase pairs, but it also involves other types of mappings between equivalent characters, such as mapping halfwidth katakana characters to normal katakana characters in Japanese. The mapping phase in IDNA2003 was included to match the insensitivity of ASCII domain names. Users are accustomed to having both CNN.com and cnn.com work identically. They expect domain names with accents to have the same casing behavior, so that ÖBB.at is the same as öbb.at. There are variations similar to case differences in other scripts. The IDNA2003 mapping is based on data specified in the Unicode Standard, Version 3.2; this mapping was later formalized as the Unicode property [NFKC_Casefold].

Note that case-folding generates a stable form of a string that erases functional case-differences. It is not the same as lowercasing. In particular, the lowercase Cherokee characters added in Unicode Version 8.0 are case-folded to their uppercase counterparts.

IDNA2008 does not require a mapping phase, but does permit one (called "Local Mapping" or "Custom Mapping"). For more information on the permitted mappings, see the Protocol document of [IDNA2008], Section 4.2, Permitted Character and Label Validation and Section 5.2, Conversion to Unicode.

The UTS #46 specification defines a mapping consistent with the normative requirements of the IDNA2008 protocol, and which is as compatible as possible with IDNA2003. For client software, this provides behavior that is the most consistent with user expectations about the handling of domain names with existing data—namely, that domain names will map consistently both on clients supporting IDNA2003 and on clients supporting IDNA2008 with the UTS #46 mapping.

1.3.2 Deviations

There are a few situations where the use of IDNA2008 without compatibility mapping will result in the resolution of IDNs to different IP addresses from in IDNA2003, unless the registry or registrant takes special action. This affects a very small number of characters, but because these characters are very common in particular languages, a significant number of domain names in those languages are affected. This set of characters is referred to as "Deviations" and is shown in Table 1, Deviation Characters, illustrated in the context of IRIs.

<table>
<thead>
<tr>
<th>Char</th>
<th>Example</th>
<th>IDNA2003 Result</th>
<th>IDNA2008 Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>ØDF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ø3C2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZWJ</td>
<td>href=&quot;http://@.com&quot;</td>
<td>http://@.com → <a href="http://xn-">http://xn-</a></td>
<td>http://@.com → <a href="http://xn-">http://xn-</a></td>
</tr>
<tr>
<td>200D</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For more information on the rationale for the occurrence of these Deviations in IDNA2008, see the [IDN FAQ].

The differences in interpretation of Deviation characters result in potential for security exploits. Consider a scenario involving http://www.sparkasse-gießen.de, a German IRI containing an IDN for "Gießen Savings and Loan".


2. She visits her friend Bob, and checks her bank statement on his browser. His browser supports IDNA2008. Under those rules, http://www.sparkasse-gießen.de is also valid, but converts to a different Punycode domain name in http://xn--sparkasse-gießen-2ib.de. This can lead to a different site with the IP address 101.123.145.167, a spoof site.

Alice ends up at the phishing site, supplies her bank password, and her money is stolen. While the .DE register (DENIC) might have a policy about bundling all of the variants of β together (so that they all have the same owner) it is not required of registries. It is unlikely that all registries will have and enforce such a bundling policy in all such cases.

There are two Deviations of particular concern. IDNA2008 allows the joiner characters (ZWJ and ZWNJ) in labels. By contrast, these are removed by the mapping in IDNA2003. When used in the intended contexts in particular scripts, the joiner characters produce a noticeable change in displayed text. However, when used between any other characters in those scripts, or in any other scripts, they are invisible. For example, when used between the Latin characters "a" and "b" there is no visible different: the sequence "a<ZWJ>b" looks just like "ab".

Because of the visual confusability introduced by the joiner characters, IDNA2008 provides a special category for them called CONTEXTJ, and only permits CONTEXTJ characters in limited contexts: certain sequences of Arabic or Indic characters. However, applications that perform IDNA2008 lookup are not required to check for these contexts, so overall security is dependent on registries having correct implementations. Moreover, the IDNA2008 context restrictions do not catch most cases where distinct domain names have visually confusable appearances because of ZWJ and ZWNJ.

2 Unicode IDNA Compatibility Processing

To satisfy user expectations for mapping, and provide maximal compatibility with IDNA2003, this document specifies a mapping for use with IDNA2008. In addition, to transition more smoothly to IDNA2008, this document provides a Unicode algorithm for a standardized processing that allows conformant implementations to minimize the security and interoperability problems caused by the differences between IDNA2003 and IDNA2008. This Unicode IDNA Compatibility Processing is structured according to
IDNA2003 principles, but extends those principles to Unicode 5.2 and later. It also incorporates the repertoire extensions provided by IDNA2008.

Where the transition processing is not needed, UTS #46 can be used purely as a preprocessing (local mapping) for IDNA2008 by claiming conformance specifically to Conformance Clause C3.

By using this Compatibility Processing, a domain name such as ÖBB.at will be mapped to the valid domain name öbb.at, thus matching user expectation for case behavior in domain names. For transitional use, the Compatibility Processing also allows domain names containing symbols and punctuation that were valid in IDNA2003, such as \.com (which has an associated web page). Such domain names containing symbols will gradually disappear as registries shift to IDNA2008.

Implementations may also restrict or flag (in a UI) domain names that include symbols and punctuation. For more information, see Unicode Technical Report # 36, Unicode Security Considerations [UTR36].

Using the Unicode IDNA Compatibility Processing to transform an IDN into a form suitable for DNS lookup is similar to the tactic of "try IDNA2008 then try IDNA2003". However, this approach avoids a potentially problematic dual lookup. It allows browsers and other clients, such as search engines, to have a single processing step, without the burden of maintaining two different implementations and multiple tables. It accounts for a number of edge cases that would cause problems, and provides a stable definition with predictable results.

The Unicode IDNA Compatibility Processing also provides alternate mappings for the Deviation characters. This facilitates the transition from IDNA2003 to IDNA2008. It is up to the registries to decide how to handle the transition, for example, by either bundling or blocking the Deviation characters that they support. The course of the transition will also depend on how soon the IDNA2003 client software is retired.

The term "registries" includes far more than top-level registries, such as for .de or .com. For example, .blogspot.com has more domain names registered than most top-level registries. There may be different policies in place for a registry and any of its subregistries. Thus millions of registries need to be considered in a transition strategy, not just hundreds.

The retirement of IDNA2003 client software may also take considerable time. IE6 was superseded in October 2006, yet as of April 2010, it still has a usage share of over 20%—higher than all other browsers except IE8! In lookup software, transitions may be fine-grained: for example, it may be possible to transition to IDNA2008 rules regarding Deviations for .blogspot.com at a given point but not for .com, or vice versa. If .de bundles or blocks the Deviation characters, then clients could transition Deviations for .de, but not for (say) .blogspot.de. Moreover, client software with a UI, such as the address bar in a browser, could provide more options for the transition. A full discussion of such transition strategies is outside of the scope of this document.

During the interim, authors of documents, such as HTML documents, can unambiguously refer to the IDNA2008 interpretation of characters by explicitly using the Punycode form of the domain name label.

There are two slightly different compatibility mechanisms for domain names during a transition and afterward. UTS #46 therefore specifies two specific types of processing: Transitional Processing (Conformance Clause C1) and Nontransitional Processing.
(Conformance Clause C2). The only difference between them is the handling of the four Deviation characters.

Summarized briefly, UTS #46 builds upon IDNA2008 in three areas:

- **Mapping.** The UTS #46 mapping is used to maintain maximal compatibility and meet user expectations. It is conformant to IDNA2008, which allows for mapping input.
- **Symbols and Punctuation.** UTS #46 supports processing of symbols and punctuation during the transitional period. The transition will be smooth: as registries move to IDNA2008 the DNS lookups of IDNs with symbols will simply be refused. At that point, in practice, there is full compatibility with IDNA2008.
- **Deviations.** UTS #46 provides two ways of handling these to support a transition. Transitional Processing should only be used immediately before a DNS lookup in the circumstances where the registry does not guarantee a strategy of bundling or blocking. Nontransitional Processing, which is fully compatible with IDNA2008, should be used in all other cases.

For a demonstration of differences between IDNA2003, IDNA2008, and the Unicode IDNA Compatibility Processing, see the [IDN_Demo]. For more detail on the differences, see Section 7, IDNA Comparison. UTS #46 does not change any of the terms defined in IDNA2008, such as A-Label or U-Label.

Neither the Unicode IDNA Compatibility Processing nor IDNA2008 address security problems associated with confusables (the so-called "paypal.com" problem). IDNA2008 disallows certain symbols and punctuation characters that can be used for spoofing, such as spoofs of the slash character ("/"). However, these are an extremely small fraction of the confusable characters used for spoofing. Moreover, confusable characters themselves account for a small proportion of phishing problems: most are cases like "secure-wellsfargo.com". For more information, see [Bortzmeyer] and the [IDN FAQ]. It is strongly recommended that Unicode Technical Report #36, Unicode Security Considerations [UTR36] and Unicode Technical Standard #39, Unicode Security Mechanisms [UTS39] be consulted for information on dealing with confusables, both for client software and registries. In particular, [UTS39] provides information that can be used to drastically reduce the number of confusables when dealing with international domain names, much beyond what IDNA2008 does. See also the [DemoConf].

### 2.1 Display of Internationalized Domain Names

IDNA2003 applications customarily display the processed string to the user. This improves security by reducing the opportunity for visual confusability. Thus, for example, the URL [http://google.com](http://google.com) (with a capital I in place of the L) is revealed as [http://googie.com](http://googie.com). However, for Deviations the distinction between the original and processed form is especially important for users. Thus the Nontransitional Processing should be used for displaying domain names. This is the same as Transitional Processing, except that it excludes the Deviations: ß and ç and the joiners. It is thus fully compatible with IDNA2008 for these Deviation characters.

Except for direct DNS lookup during the transitional period, the Nontransitional Processing should always be used, preserving the Deviation characters in the original string as per IDNA2008. Once the transition for Deviation characters is over, Nontransitional Processing can be used exclusively.

### 2.2 Registries
This specification is primarily targeted at applications doing lookup of IDNs. There is, however, one strong recommendation for registries: do not allow the registration of labels that are invalid according to Nontransitional Processing, and for a transitional period, bundle or block labels containing Deviation characters.

These tactics can be described as follows:

- **Bundling**: If the transitional and non-transitional forms differ, and are both registered, the registrant for each must be the same.

- **Blocking**: If transitional and nontransitional forms differ, allow the registration of only one, and block the others. Registries that do not allow any Deviation characters at all count as **blocking**.

The label that is actually registered and inserted into a registry has always been processed. For example, xn–bcher-kva corresponds to bücher. However, it may be useful for a registry to also ask for "unprocessed" labels, such as Bücher, as part of the registration process, so that they are aware of the registrant's intent. However, such unprocessed labels must be handled carefully:

- Storing the unprocessed label as the sequence of characters that the registrant really wanted to apply for.

- Processing the unprocessed label, and displaying the processed label to the registrant for confirmation.

- Proceeding with the regular registration process using only the processed label.

### 2.3 Notation

Sets of code points are defined using properties and the syntax of Unicode Technical Standard #18, Unicode Regular Expressions [UTS18]. For example, the set of combining marks is represented by the syntax \( \text{\textbackslash d} \{ \text{\textbackslash gc}\} \). Additionally, the "+" indicates the addition of elements to a set, for clarity.

In this document, a *label* is a substring of a domain name. That substring is bounded on both sides by either the start or the end of the string, or any of the following characters, called *label-separators*:

1. U+002E ( . ) FULL STOP
2. U+FF0E ( 。 ) FULLWIDTH FULL STOP
3. U+3002 ( 。 ) IDEOGRAPHIC FULL STOP
4. U+FF61 ( ． ) HALFWIDTH IDEOGRAPHIC FULL STOP

Many people use the terms "domain names" and "host names" interchangeably. This document follows [RFC3490] in use of the term "domain name".

### 3 Conformance

The requirements for conformance on implementations of the Unicode IDNA Compatibility Processing algorithm are stated in the following clauses. An implementation can claim conformance to any or all of these clauses independently.

**C1.** Given a version of Unicode and a **Unicode String**, a conformant implementation of **Transitional Processing** shall replicate the results given by applying the **Transitional Processing** algorithm specified by Section 4, **Processing**.
C2. Given a version of Unicode and a Unicode String, a conformant implementation of Nontransitional Processing shall replicate the results given by applying the Nontransitional Processing algorithm specified by Section 4, Processing.

C3. Given a version of Unicode and a Unicode String, a conformant implementation of Preprocessing for IDNA2008 shall replicate the results specified by Section 4.4, Preprocessing for IDNA2008.

These specifications are logical ones, designed to be straightforward to describe. An actual implementation is free to use different methods as long the result is the same as that specified by the logical algorithm.

Any conformant implementation may also have tighter validity criteria than those imposed by Section 4.1, Validity Criteria. For example, an application could disallow or warn of domain name labels with certain characteristics, such as:

- labels with certain combinations of scripts (Safari)
- labels with characters outside of the user's specified languages (IE)
- labels with certain confusables characters (Firefox)
- labels that are detected by the Google Safe Browsing API [SafeBrowsing]
- labels that do not meet the validity requirements of IDNA2008
- labels produced by toUnicode that would not meet the label validity requirements if toASCII were performed.
- labels containing characters which are not contained in the General Security Profile for Identifiers from Unicode Technical Standard #39, Unicode Security Mechanisms [UTS39]
- labels that do not satisfy Restriction Level 4, Moderately Restrictive from Unicode Technical Standard #39, Unicode Security Mechanisms [UTS39]


3.1 STD3 Rules

IDNA2003 provides for a flag, UseSTD3ASCIIIRules, that allows for implementations to choose whether or not to abide by the rules in [STD3]. These rules exclude ASCII characters outside the set consisting of A-Z, a-z, 0-9, and U+002D ( - ) HYPHEN-MINUS. For example, some browsers also allow characters such as U+005F ( _ ) LOW LINE (underbar) in domain names, and thus use UseSTD3ASCIIIRules=false, plus their own validity checks for the other ASCII characters.

While UseSTD3ASCIIIRules=true is strongly recommended, Section 5, IDNA Mapping Table provides data to allow implementations to support UseSTD3ASCIIIRules=false for compatibility with IDNA2003 implementations where necessary. The mapping table does this: providing the status values and mapping values for both UseSTD3ASCIIIRules=true and UseSTD3ASCIIIRules=false. Implementations that use UseSTD3ASCIIIRules=false will need to apply their own validation to the mapped values as indicated in Section 4.1, Validity Criteria.

4 Processing

The input to Unicode IDNA Compatibility Processing is a prospective domain name string expressed in Unicode, and a choice of Transitional or Nontransitional Processing. The
domain name consists of a sequence of labels with dot separators, such as "Bücher.de". For more information about the composition of a URL, see Section 3.5 of [STD13].

Main Processing Steps

The following steps, performed in order, successively alter the input domain_name string and then output it as a converted Unicode string, plus a flag to indicate whether there was an error. Even if an error occurs, the conversion of the string is performed as much as is possible.

Input

- A prospective domain_name expressed as a sequence of Unicode code points
- A boolean flag: UseSTD3ASCIIIRules
- A boolean flag: CheckHyphens
- A boolean flag: CheckBidi
- A boolean flag: CheckJoiners
- A processing_option with one of two values:
  - Transitional_Processing for transitional handling of Deviation characters, or
  - Nontransitional_Processing otherwise

Processing

1. Map For each code point in the domain_name string, look up the status value in Section 5, IDNA Mapping Table, and take the following actions:
   - disallowed: Leave the code point unchanged in the string, and record that there was an error.
   - ignored: Remove the code point from the string. This is equivalent to mapping the code point to an empty string.
   - mapped: Replace the code point in the string by the value for the mapping in Section 5, IDNA Mapping Table.
   - deviation:
     - For Transitional_Processing, replace the code point in the string by the value for the mapping in Section 5, IDNA Mapping Table .
     - For Nontransitional_Processing, leave the code point unchanged in the string.
   - valid: Leave the code point unchanged in the string.

2. Normalize. Normalize the domain_name string to Unicode Normalization Form C.

3. Break. Break the string into labels at U+002E ( . ) FULL STOP.

4. Convert/Validate. For each label in the domain_name string:
   - If the label starts with "xn--":
     1. Attempt to convert the rest of the label to Unicode according to Punycode [RFC3492]. If that conversion fails, record that there was an error, and continue with the next label. Otherwise replace the original label in the string by the results of the conversion.
     2. Verify that the label meets the validity criteria in Section 4.1, Validity Criteria for Nontransitional Processing. If any of the validity criteria are not satisfied, record that there was an error.
   - If the label does not start with "xn--":
• Verify that the label meets the validity criteria in Section 4.1, Validity Criteria for the input Processing choice (Transitional or Nontransitional). If any of the validity criteria are not satisfied, record that there was an error.

Any input domain_name string that does not record an error has been successfully processed according to this specification. Conversely, if an input domain_name string causes an error, then the processing of the input domain_name string fails. Determining what to do with error input is up to the caller, and not in the scope of this document. The processing is idempotent—reapplying the processing to the output will make no further changes. For examples, see Table 2, Examples of Transitional Processing.

Implementations may make further modifications to the resulting Unicode string when showing it to the user. For example, it is recommended that disallowed characters be replaced by a U+FFFD to made them visible to the user. Similarly, labels that fail processing during steps 4 or 5 may be marked by the insertion of a U+FFFD or other visual device.

With either Transitional or Nontransitional Processing, sources already in Punycode are validated without mapping. In particular, Punycode containing Deviation characters, such as href="xn--fu-hia.de" (for fuß.de) is not remapped. This provides a mechanism allowing explicit use of Deviation characters even during a transition period.

4.1 Validity Criteria

Each of the following criteria must be satisfied for a label:

1. The label must be in Unicode Normalization Form NFC.
2. If CheckHyphens, the label must not contain a U+002D HYPHEN-MINUS character in both the third and fourth positions.
3. If CheckHyphens, the label must neither begin nor end with a U+002D HYPHEN-MINUS character.
4. The label must not contain a U+002E ( . ) FULL STOP.
5. The label must not begin with a combining mark, that is: General_Category=Mark.
6. Each code point in the label must only have certain status values according to Section 5, IDNA Mapping Table:
   1. For Transitional Processing, each value must be valid.
   2. For Nontransitional Processing, each value must be either valid or deviation.
7. If CheckJoiners, the label must satisfy the ContextJ rules from Appendix A, in The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) [IDNA2008].
8. If CheckBidi, the label must satisfy the Bidi rules from Section 2, The Bidi Rule, in Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA) [IDNA2008].

The first 5 criteria are from [IDNA2008]. Criterion #2 in particular is meant to allow for future label extensions beyond just xn--, such as for future versions of IDNA. Some implementations appear to consider such extentions unlikely, and allow labels such as "r3--sn-apo3qvuoxubt-j5pe".

Review Notes:
1. If CheckJoiners, those implementations that support symbols (including emoji) in labels cannot support emoji ZWJ sequences, such as women professions. Should we provide an option for doing the IDNA CheckJoiners test, but allowing emoji zwj sequences for a specified version of Unicode Emoji? For example:

7. If CheckJoiners, the label must satisfy the ContextJ rules from Appendix A, in The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) [IDNA2008], except that if EmojiVersion=0, ZWJ characters are allowed if they are within Emoji ZWJ Sequences specified for Unicode Emoji Version=EmojiVersion.

2. The IDNA ContextJ rules for ZWNJ and ZWJ are coarse tests; they do not exclude as much as they could. One option is to use the rules in UAX #31 (Layout_and_Format_Control_Characters) to have a more fine-grained approach.

Any particular application may have tighter validity criteria, as discussed in Section 3, Conformance. For example, implementations are encouraged to apply the Bidi and ContextJ validity criteria from [IDNA2008]. Thus these conditions are also flagged in the data files for Section 8, Conformance Testing, for ease of testing.

1. **Bidi rules** from Section 2, The Bidi Rule, in Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA) [IDNA2008].

2. **ContextJ rules** from Appendix A, in The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) [IDNA2008].

### 4.1.1 UseSTD3ASCIIRules

If UseSTD3ASCIIRules=false, then the validity tests for ASCII characters are not provided by the table status values, but are implementation-dependent. For example, if an implementation allows the characters [\u0020-~] and also the underbar (_) then it needs to use the table values for UseSTD3ASCIIRules=false, and test for any other ASCII characters as part of its validity criteria. These ASCII characters may have resulted from a mapping: for example, a U+005F ( _ ) LOW LINE (underbar) may have originally been a U+FF3F ( ___ ) FULLWIDTH LOW LINE.

There are a very small number of non-ASCII characters with the data file status disallowed_STD3_valid:

- U+2260 ( ≠ ) NOT EQUAL TO
- U+226E ( ≠ ) NOT LESS-THEAN
- U+226F ( > ) NOT GREATER-THEAN

Those characters are disallowed with UseSTD3ASCIIRules=true because the set of characters in their canonical decompositions are not entirely in the valid set (Step 7 of the Table Derivation). However, they are allowed with UseSTD3ASCIIRules=false, because the base characters of their canonical decompositions, U+003D ( = ) EQUALS SIGN, U+003C ( < ) LESS-THEAN SIGN, and U+003E ( > ) GREATER-THEAN SIGN, are each valid under that option. If an implementation uses UseSTD3ASCIIRules=false but disallows any of these three ASCII characters, then it must also disallow the corresponding precomposed character for its negation.

### 4.1.2 Right-to-Left Scripts
In addition, the label should meet the requirements for right-to-left characters specified in the Right-to-Left Scripts document of [DNA2008], and for the CONTEXTJ requirements in the Protocol document of [DNA2008]. It is strongly recommended that Unicode Technical Report #36, Unicode Security Considerations [UTR36] and Unicode Technical Standard #39, Unicode Security Mechanisms [UTS39] be consulted for information on dealing with confusables, and for characters that should be excluded from identifiers. Note that the recommended exclusions are a superset of those in [DNA2008].

4.2 ToASCII

The operation corresponding to ToASCII of [RFC3490] is defined by the following steps:

Input

- A prospective domain_name expressed as a sequence of Unicode code points
- A boolean flag: CheckHyphens
- A boolean flag: CheckBidi
- A boolean flag: CheckJoiners
- A boolean flag: UseSTD3ASCIIIRules
- A processing_option with one of two values:
  - Transitional_Processing for transitional handling of Deviation characters, or
  - Nontransitional_Processing otherwise
- A boolean flag: VerifyDnsLength

Processing

1. To the input domain_name, apply the Processing Steps in Section 4 Processing, using the input processing_option and the boolean flags CheckHyphens, CheckBidi, CheckJoiners, and UseSTD3ASCIIIRules. This may record an error.
2. Break the result into labels at U+002E FULL STOP.
3. Convert each label with non-ASCII characters into Punycode [RFC3492], and prefix by “xn--”. This may record an error.
4. If VerifyDnsLength flag is true, then verify DNS length restrictions. This may record an error. For more information, see [STD13] and [STD3].
   1. The length of the domain name, excluding the root label and its dot, is from 1 to 253.
   2. The length of each label is from 1 to 63.
5. If an error was recorded in steps 1-4, then the operation has failed and a failure value is returned. No DNS lookup should be done.
6. Otherwise join the labels using U+002E FULL STOP as a separator, and return the result.

Implementations are advised to apply additional tests to these labels, such as those described in Unicode Technical Report #36, Unicode Security Considerations [UTR36] and Unicode Technical Standard #39, Unicode Security Mechanisms [UTS39], and take appropriate actions. For example, a label with mixed scripts or confusables may be called out in the UI. Note that the use of Punycode to signal problems may be counter-productive, as described in [UTR36].

4.3 ToUnicode
The operation corresponding to ToUnicode of [RFC3490] is defined by the following steps:

Input

- A prospective domain_name expressed as a sequence of Unicode code points
- A boolean flag: CheckHyphens
- A boolean flag: CheckBidi
- A boolean flag: CheckJoiners
- A boolean flag: UseSTD3ASCIIRules

Processing

1. To the input domain_name, apply the Processing Steps in Section 4 Processing, using Nontransitional_Processing and the boolean flags CheckHyphens, CheckBidi, CheckJoiners, and UseSTD3ASCIIRules. This may record an error.
2. Like [RFC3490], this will always produce a converted Unicode string. Unlike ToASCII of [RFC3490], this always signals whether or not there was an error.

Implementations are advised to apply additional tests to these labels, such as those described in Unicode Technical Report #36, Unicode Security Considerations [UTR36] and Unicode Technical Standard #39, Unicode Security Mechanisms [UTS39], and take appropriate actions. For example, a label with mixed scripts or confusables may be called out in the UI. Note that the use of Punycode to signal problems may be counter-productive, as described in [UTR36].

4.4 Preprocessing for IDNA2008

The table specified in Section 5, IDNA Mapping Table may also be used for a pure preprocessing step for IDNA2008, mapping a Unicode string for input directly to the algorithm specified in IDNA2008.

Preprocessing for IDNA2008 is specified as follows:

Apply the Section 4.3, ToUnicode processing to the Unicode string.

Note that this preprocessing allows some characters that are invalid according to IDNA2008. However, the IDNA2008 processing will catch those characters. For example, a Unicode string containing a character listed as DISALLOWED in IDNA2008, such as U+2665 (♣) BLACK HEART SUIT, will pass the preprocessing step without an error, but subsequent application of the IDNA2008 processing will fail with an error, indicating that the string is not a valid IDN according to IDNA2008.

4.5 Implementation Notes

A number of optimizations can be applied to the Unicode IDNA Compatibility Processing. These optimizations can improve performance, reduce table size, make use of existing NFKC transform mechanisms, and so on. For example:

- There is an NFC check in Section 4.1, Validity Criteria. However, it only needs to be applied to labels that were converted from Punycode into Unicode in Step 3.
- A simple way to do much of the validity checking in Section 4.1, Validity Criteria is to reapply Steps 1 and 2, and verify that the result does not change.
Because the four label separators are all mapped to U+002E ( . ) FULL STOP by Step 1, the parsing of labels in Steps 3 and 4 only need to detect U+002E ( . ) FULL STOP, and not the other label separators defined in IDNA [RFC3490].

Note that the input domain_name string for the Unicode IDNA Compatibility Processing must have had all escaped Unicode code points converted to Unicode code points. For example, U+5341 ( 十 ) CJK UNIFIED IDEOGRAPH-5341 could have been escaped as any of the following:

- &amp;x5341; an HTML numeric character reference (NCR)
- \u5341 a Javascript escapes
- %E5%8D%81 a URI/IRI %-escape

Examples are shown in Table 2, Examples of Processing:

<table>
<thead>
<tr>
<th>Input</th>
<th>Map</th>
<th>Normalize</th>
<th>Convert</th>
<th>Validate</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloß.de</td>
<td>bloss.de</td>
<td>=</td>
<td>n/a</td>
<td>ok</td>
<td>Transitional: maps uppercase and eszett</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Nontransitional: maps uppercase</td>
</tr>
<tr>
<td>xn--blo-7ka.de</td>
<td>=</td>
<td>=</td>
<td>bloß.com</td>
<td>ok</td>
<td>Punycode is not mapped, so ß never changes (whether transitional or not).</td>
</tr>
<tr>
<td>u¨.com</td>
<td>=</td>
<td>ü.com</td>
<td>n/a</td>
<td>ok</td>
<td>Normalize changes u + umlaut to ü</td>
</tr>
<tr>
<td>xn--tda.com</td>
<td>=</td>
<td>=</td>
<td>ü.com</td>
<td>ok</td>
<td>Punycode xn--tda changes to ü</td>
</tr>
<tr>
<td>xn--u--ccb.com</td>
<td>=</td>
<td>=</td>
<td>u¨.com</td>
<td>error</td>
<td>Punycode is not mapped, but is validated. Because u + umlaut is not NFC, it fails.</td>
</tr>
<tr>
<td>a1.com</td>
<td>error</td>
<td>error</td>
<td>error</td>
<td>error</td>
<td>The character &quot;1.&quot; is disallowed, because it would produce a dot when mapped.</td>
</tr>
<tr>
<td>xn--a-ecp.ru</td>
<td>xn--a-ecp.ru</td>
<td>=</td>
<td>=</td>
<td>a1..ru</td>
<td>error</td>
</tr>
<tr>
<td>xn--0.pt</td>
<td>xn--0.pt</td>
<td>=</td>
<td>=</td>
<td>error</td>
<td>error</td>
</tr>
<tr>
<td>日本語。J P</td>
<td>日本語.jp</td>
<td>=</td>
<td>=</td>
<td>n/a</td>
<td>ok</td>
</tr>
<tr>
<td>🇺.us</td>
<td></td>
<td></td>
<td></td>
<td>n/a</td>
<td>ok</td>
</tr>
</tbody>
</table>

5 IDNA Mapping Table

For each code point in Unicode, the IDNA Mapping Table provides one of the following status values:

- **valid**: the code point is valid, and not modified.
- **ignored**: the code point is removed: this is equivalent to mapping the code point to an empty string.
- **mapped**: the code point is replaced in the string by the value for the mapping.
- **deviation**: the code point is either mapped or valid, depending on whether the processing is transitional or not.
- **disallowed**: the code point is not allowed.
  - **disallowed_STD3_valid**: the status is disallowed if `UseSTD3ASCIIRules=true` (the normal case); implementations that allow `UseSTD3ASCIIRules=false` would treat the code point as valid.
  - **disallowed_STD3_mapped**: the status is disallowed if `UseSTD3ASCIIRules=true` (the normal case); implementations that allow `UseSTD3ASCIIRules=false` would treat the code point as mapped.
If this status value is **mapped, disallowed_STD3_mapped** or **deviation**, the table also supplies a mapping value for that code point.

A table is provided for each version of Unicode starting with Unicode 5.1, in versioned directories under [IDNA-Table]. Each table for a version of the Unicode Standard will always be backward compatible with previous versions of the table: only characters with the status value **disallowed** may change in status or mapping value. Unlike the IDNA2008 table, this table is designed to be applied to the entire domain name, not just to individual labels. That design provides for the IDNA2003 handling of label separators. In particular, the table is constructed to forbid problematic characters such as U+2488 ( 1. ) DIGIT ONE FULL STOP, whose decompositions contain a "dot".

The Unicode IDNA Compatibility Processing is based on the Unicode character mapping property [NFKC_Casefold]. Section 6, **Mapping Table Derivation** describes the derivation of these tables. Like derived properties in the Unicode Character Database, the description of the derivation is informative. Only the data in IDNA Mapping Table is normative for the application of this specification.

The files use a semicolon-delimited format similar to those in the Unicode Character Database [UAX44]. The field values are listed in Table 2b, **Data File Fields**:

<table>
<thead>
<tr>
<th>Num</th>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Code point(s)</td>
<td>Hex value or range of values.</td>
</tr>
<tr>
<td>1</td>
<td>Status</td>
<td>valid, ignored, mapped, deviation, disallowed, disallowed_STD3_valid, or disallowed_STD3_mapped</td>
</tr>
<tr>
<td>2</td>
<td>Mapping</td>
<td>Hex value(s). Only present if the status is ignored, mapped, deviation, or disallowed_STD3_mapped</td>
</tr>
<tr>
<td>3</td>
<td>IDNA2008 Status</td>
<td>There are two values: NV8 and XV8. NV8 is only present if the status is valid but the character is excluded by IDNA2008 from all domain names for all versions of Unicode. XV8 is present when the character is excluded by IDNA2008 for the current version of Unicode. These are not normative values.</td>
</tr>
</tbody>
</table>

**Example:**

```
0000..002C ; disallowed                       # NULL..COMMA
002D ; valid                                  # HYPHEN-MINUS
...                                           
0041 ; mapped ; 0061                          # LATIN CAPITAL LETTER A
...                                           
00A1..00A7 ; valid ; ; NV8                    # INVERTED EXCLAMATION MARK..SECTION SIGN
00AD ; ignored                                # SOFT HYPHEN
...                                           
00DF ; deviation ; 0073 0073                  # LATIN SMALL LETTER SHARP S
```

http://www.unicode.org/reports/tr46/tr46-18.html
6 Mapping Table Derivation

The following describes the derivation of the mapping table. This description has nothing to do with the actual mapping of labels in Section 4, Processing. Instead, this section describes the derivation of the table in Section 5, IDNA Mapping Table. That table is then normatively used for mapping in Section 4, Processing.

The derivation is described as a series of steps. Step 1 defines a base mapping; Steps 2, 3, and 4 define three sets of characters. Step 5 will modify the base mapping or the sets of characters as needed to maintain backward compatibility. The mapping and sets are all used in Step 6 to produce the mapping and status values for the table. Step 7 removes characters whose NFD form would be invalid. Each numbered step may have substeps: for example, Step 1 consists of Steps 1.1 through 1.2.

The computation is done twice, once with UseSTD3ASCIIRules=true, and once with UseSTD3ASCIIRules=false. Code points that are disallowed with UseSTD3ASCIIRules=true, but valid or mapped with UseSTD3ASCIIRules=false, are given the special values disallowed_STD3_valid and disallowed_STD3_mapped.

If a Unicode property changes in a future version in a way that would affect backward compatibility, a grandfathering clause will be added to Step 5 to maintain compatibility. For more information on compatibility, see Section 5, IDNA Mapping Table.

Step 1: Define a base mapping

This step specifies a base mapping, which is a mapping from each Unicode code point to sequences of zero or more code points. The value resulting from mapping a particular code point C is called the base mapping value of C. The base mapping value for C may be identical to C.

1. Map the following exceptional characters:
   a. Map label separator characters to U+002E ( . ) FULL STOP:
      • U+FF0E ( . ) FULLWIDTH FULL STOP
      • U+3002 ( 　) IDEOGRAPHIC FULL STOP
      • U+FF61 ( 。) HALFWIDTH IDEOGRAPHIC FULL STOP
   b. Map all Bidi_Control characters to themselves

2. Map each other character to its NFKC_Casfold value [NFKC_Casfold].

Unicode 6.3 adds Bidi_Control characters that were not present in Unicode 3.2. To preserve the intent of IDNA2003 in disallowing Bidi_Control characters rather than just ignoring them, Step 1.1.b was added. This step causes Step 6.3 to disallow all Bidi_Control characters.

Step 1.1.b only affects 5 new characters added in Unicode 6.3. It would also impact any new Bidi_Control characters in future versions of the standard.

Step 2: Specify the base valid set

The base valid set is defined by the sequential list of additions and subtractions in Table 3, Base Valid Set. This definition is based on the principles of IDNA2003. When applied to
the repertoire of Unicode 3.2 characters, this produces a set which is closely aligned with IDNA2003.

### Table 3. Base Valid Set

<table>
<thead>
<tr>
<th>Formal Set Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{Changes\textunderscore\textunderscoreNFKC\textunderscoreCasefolded})</td>
<td>Start with characters that are equal to their \textit{NFKC\textunderscoreCasefold} value. This criterion excludes uppercase letters, for example, as well as characters that are unstable under NFKC normalization, and default ignorable code points. Note that according to Perl/Java syntax, (\text{\textbackslash P}) means the inverse of (\text{\textbackslash p}), so these are the characters that \textit{do not} change when individually mapped according to \textit{NFKC\textunderscoreCasefold}.</td>
</tr>
<tr>
<td>(\text{\textbackslash p(c) - \text{\textbackslash p(z)})</td>
<td>Remove Unassigned, Controls, Private Use, Format, Surrogate, and Whitespace.</td>
</tr>
<tr>
<td>(\text{\textbackslash p(Block=Ideographic\textunderscoreDescription_Characters})</td>
<td>Remove ideographic description characters.</td>
</tr>
<tr>
<td>(\text{\textbackslash p(ascii) + [\texttt{u002Da-zA-Z0-9]})</td>
<td>If UseSTD3ASCIIRules = True: Remove disallowed ASCII; '-' is valid.</td>
</tr>
<tr>
<td>(\text{\textbackslash p(ascii) - [\texttt{u002E}])</td>
<td>If UseSTD3ASCIIRules = False: Add all ASCII except for &quot;.&quot;</td>
</tr>
</tbody>
</table>

### Step 3: Specify the base exclusion set

Form the base exclusion set in the following way:

1. Start with the empty set.
2. Add each code point C such that:
   a. According to IDNA2003, C is neither prohibited nor unassigned nor a label separator (that is, it is either valid or mapped), \textit{and}
   b. According to IDNA2003, C has a different mapping than C's base mapping value specified in Step 1.
3. Add each code point C such that:
   a. According to IDNA2003, C is prohibited, and
   b. either C is in the base valid set, or every code point in C’s base mapping value is in the base valid set.

For example, for Unicode 5.2 and 6.0, the base exclusion set consists of the list that follows. The subheads (like “Case Changes”) are informational, and do not represent the principle for excluding the characters listed under them.

**Characters that have a different mapping in IDNA2003 (Step 3.2 above)**

- Case Changes
  - U+04C0 (І) CYRILLIC LETTER PALOCHKA
  - U+10A0 (Ї) GEORGIAN CAPITAL LETTER AN...U+10C5 ( Dtype) GEORGIAN CAPITAL LETTER HOE
  - U+2132 (⇧) TURNED CAPITAL F
  - U+2183 (☺) ROMAN NUMERAL REVERSED ONE HUNDRED
- Normalization Changes (CJK Compatibility Characters)
  - U+2F868, U+2F874, U+2F91F, U+2F95F, U+2F9BF
- Default ignorable Changes
  - U+3164 HANGUL FILLER
  - U+FFA0 HALFWIDTH HANGUL FILLER
  - U+115F HANGUL CHOSEONG FILLER
  - U+1160 HANGUL JUNGSEONG FILLER
  - U+17B4 KHMER VOWEL INHERENT AQ
  - U+17B5 KHMER VOWEL INHERENT AA
  - U+1806 (¬) MONGOLIAN TODO SOFT HYPHEN

**Characters that are disallowed in IDNA2003 (Step 3.3 above)**

- Bidi_Control characters
  - U+200E LEFT-TO-RIGHT MARK..U+200F RIGHT-TO-LEFT MARK
  - U+202A LEFT-TO-RIGHT EMBEDDING..U+202E RIGHT-TO-LEFT OVERRIDE
- Invisible operators
  - U+2061 FUNCTION APPLICATION..U+2063 INVISIBLE SEPARATOR
- Replacement characters
  - U+FFFC OBJECT REPLACEMENT CHARACTER
  - U+FFFD (�) REPLACEMENT CHARACTER
- Musical symbols
  - U+1D173 MUSICAL SYMBOL BEGIN BEAM..U+1D17A MUSICAL SYMBOL END PHRASE
- Format characters (deprecated)
  - U+206A INHIBIT SYMMETRIC SWAPPING..U+206F NOMINAL DIGIT SHAPES
- Tags (deprecated)
  - U+0001 LANGUAGE TAG
○ U+E07F CANCEL TAG
  • Other tags
    ○ U+E020 TAG SPACE..U+E07E TAG TILDE

Step 4: Specify the deviation set

This is the set of characters that deviate between IDNA2003 and IDNA2008.

  • U+200C ZERO WIDTH NON-JOINER
  • U+200D ZERO WIDTH JOINER
  • U+00DF ( β ) LATIN SMALL LETTER SHARP S
  • U+03C2 ( ζ ) GREEK SMALL LETTER FINAL SIGMA

Step 5: Specify grandfathered changes

This set is currently empty. Adjustments to the above sets or base mapping will be made in this section if the steps would cause an already existing character to change status or mapping under a future version of Unicode, so that backward compatibility is maintained.

Step 6: Produce the initial status and mapping values

For each code point:

1. If the code point is in the deviation set
   • the status is deviation and the mapping value is the base mapping value for that code point.
2. Otherwise, if the code point is in the base exclusion set or is unassigned
   • the status is disallowed and there is no mapping value in the table.
3. Otherwise, if the code point is not a label separator and some code point in its base mapping value is not in the base valid set
   • the status is disallowed and there is no mapping value in the table.
4. Otherwise, if the base mapping value is an empty string
   • the status is ignored and there is no mapping value in the table.
5. Otherwise, if the base mapping value is the same as the code point
   • the status is valid and there is no mapping value in the table.
6. Otherwise,
   • the status is mapped and the mapping value is the base mapping value for that code point.

Step 7: Produce the final status and mapping values

After processing all code points in previous steps:

1. Iterate through the set of characters with a status of valid. Any whose canonical decompositions (NFD) are not wholly in the valid set, make disallowed.
2. Iterate through the set of characters with a status of mapped. Any with mapping values whose canonical decompositions (NFD) are not wholly in the valid set, make disallowed.
3. Recursively apply these actions until there are no more status changes.
For example, for Unicode 5.2 and 6.0, the set of characters set to disallowed in Step 7 consists of the following:

- U+2260 (≠) NOT EQUAL TO
- U+226E (<\) NOT LESS-THAN
- U+226F (>) NOT GREATER-THAN
- U+FE12 (°) PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC FULL STOP

Note that characters such as U+2488 (1.) DIGIT ONE FULL STOP are disallowed by Step 6.3.

7 IDNA Comparison

Table 4, **IDNA Comparisons** illustrates the differences between the three specifications in terms of valid character repertoire. It omits the ASCII-repertoire code points, all code points unassigned in the latest version of Unicode, as well as control characters, private-use characters, and surrogate code points. It also includes labels separators that are valid or mapped. The table separates the Unicode 3.2 characters from those encoded later, because they have a special status in IDNA2003. It also separates buckets where UTS #46 and IDNA2008 behave the same from those where they behave differently.

Each row in the table defines a bucket of code points that share a pattern of behavior across the three specifications. The columns provide the following information:

- The column titled **Count** shows the number of characters in each bucket.
- The columns titled **IDNA2003**, **UTS46**, and **IDNA2008** show the status of the characters in each bucket for the respective specifications.
  - Deviations are modified in Transitional Processing, but not modified in Nontransitional Processing; see Section 4, **Processing**.
  - IDNA2003 allows unassigned code points in lookup but not registration. These are in the section of the table under "Unicode 4.0 to Latest", and marked as **LookupValid**.
  - IDNA2008 uses several subcategories that are grouped together here for comparison. Characters marked as **Valid** are those that are CONTEXTJ, CONTEXTO, and PVALID in IDNA2008*. Other characters are marked as **Disallowed**.

  *This list of Valid characters for Unicode 4.0 and beyond is calculated as the union of characters with values CONTEXTJ, CONTEXTO, and PVALID under any version of Unicode from Version 5.2 or later. The union of valid characters over versions of Unicode is used for comparison because IDNA2008 does not guarantee backward compatibility over different versions of Unicode.*

- The column titled **Comment and Examples** describes the correlation between the specifications and provides illustrative characters.

<table>
<thead>
<tr>
<th>Count</th>
<th>IDNA2003</th>
<th>UTS46</th>
<th>IDNA2008</th>
<th>Comment and Examples</th>
</tr>
</thead>
</table>

http://www.unicode.org/reports/tr46/tr46-18.html
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>86,676</td>
<td>Valid</td>
<td>Valid</td>
<td>Valid</td>
</tr>
<tr>
<td>b</td>
<td>431</td>
<td>Disallowed</td>
<td>Disallowed</td>
<td>Disallowed</td>
</tr>
<tr>
<td>c</td>
<td>48</td>
<td>Valid</td>
<td>Disallowed</td>
<td>Disallowed</td>
</tr>
<tr>
<td>d</td>
<td>8</td>
<td>Mapped</td>
<td>Disallowed</td>
<td>Disallowed</td>
</tr>
<tr>
<td>e</td>
<td>4,640</td>
<td>Mapped / Ignored</td>
<td>Mapped / Ignored</td>
<td>Disallowed</td>
</tr>
<tr>
<td>f</td>
<td>3,254</td>
<td>Valid</td>
<td>Valid</td>
<td>Disallowed</td>
</tr>
<tr>
<td>g</td>
<td>4</td>
<td>Mapped / Ignored</td>
<td>Display: Valid Lookup: Mapped / Ignored</td>
<td>Valid</td>
</tr>
<tr>
<td>Unicode 4.0 to Latest (UTS46 = IDNA2008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>35,722</td>
<td>LookupValid</td>
<td>Valid</td>
<td>Valid</td>
</tr>
<tr>
<td>i</td>
<td>61</td>
<td>LookupValid</td>
<td>Disallowed</td>
<td>Disallowed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unicode 4.0 to Latest (UTS46 ≠ IDNA2008)</th>
</tr>
</thead>
<tbody>
<tr>
<td>j</td>
</tr>
<tr>
<td>k</td>
</tr>
</tbody>
</table>

**Caution:** There are 26,568 Unicode characters whose IDNA2008 values are not official. That is, the IDNA2008 values listed here and in the data tables are preliminary. They are computed according to the rules in [IDNA2008], but the IETF has not yet updated RFC 5892 for any characters added to the Unicode Standard since Version 6.3.

The IETF may also change their IDNA2008 values for any particular version of Unicode, and even do so retroactively. Thus the IDNA2008 values for any version of Unicode are also not guaranteed to be backward-compatible with the IDNA2008 for previous versions of Unicode. For example, even as of the release date for Unicode 8.0, it is as yet unclear what the IETF position is on which characters should be valid for IDNA2008 for Unicode 7.0. (Only a very small number of characters added in Unicode 7.0 have raised concerns within IETF.)

**Review Notes:**

1. TODO Update the h-k values if there are any changes in data before release

A more detailed online listing of differences is found at [DemoIDNChars] and [DemoIDN]. The implications for confusability can be seen at [DemoConf].

### 7.1 Implications for Implementers

Table 4, [IDNA Comparisons](http://www.unicode.org/reports/tr46/tr46-18.html) can also be used to categorize implications for implementers.

With the exception of Row (g), if any characters are Mapped/Ignored in any specification—Rows (d), (e), (k)—then in the other specifications they are either Mapped/Ignored in precisely the same way, or they are Disallowed. This prevents domain names from being mapped differently on different browsers: either the characters map to the same result, or
they do not work. Row (k) is unproblematic in this regard, assuming that registries follow one of the specifications, because characters like U+023A (Å) will not be valid in registered labels.

The only exceptions are the four problematic Deviations in Row (g), which require more complex handling because of their treatment in IDNA2008, as discussed earlier. Transitions for Deviation characters will depend upon how registries handle IDNA2008 going forward, and how soon IDNA2003 clients are retired. Outside of the transition from IDNA2003 to IDNA2008, the UTS #46 Nontransitional Processing should be used, thus preserving Deviation characters.

This presupposes that IDNA2008 implementations do not use custom, incompatible mappings: that is, that they do not take advantage of the fact that arbitrary mappings are allowed in IDNA2008, and choose a mapping that is incompatible with IDNA2003 or UTS #46. This pertains to any of Rows (e), (f), (j), (k). If custom mappings were used by any significant client base, it would result in serious problems for security and interoperability. For more information, see the [IDN_FAQ].

With the exception of the above issues, implementation is straightforward:

- Rows (a) and (b) are unproblematic. All three specifications behave identically.
- Rows (c) and (d) are unproblematic. They contain characters that are allowed under IDNA2003, but are disallowed in UTS #46 because their mappings would be different after Unicode 3.2, based on the Unicode Standard mappings. This treatment also matches IDNA2008. Those mappings were stabilized some time ago, so mappings will not change in the future; see [Stability]. Fortunately, in-depth analysis of Web content indicates these characters are quite rare: their presence in domain names in web pages cannot be distinguished from noise (unlike the Deviation characters in Row (g)).
- Rows (e) and (k) are unproblematic. Ideally, implementations will map these characters in IDNA2008, producing precisely the same results as in UTS #46, and the same results for Unicode 3.2 characters as IDNA2003.
- Rows (f) and (j) are symbols and punctuation that are disallowed in IDNA2008, but allowed transitionally in UTS #46. Row (j) contains post-Unicode 3.2 characters that are handled in UTS #46 according to IDNA2003 principles. These symbols and punctuation will transition smoothly as registries discontinue support for them.
- Rows (h) and (i) are unproblematic. The characters have the same status in IDNA2008 and UTS #46.

8 Conformance Testing

A conformance testing file (IdnaTest.txt) is provided for each version of Unicode starting with Unicode 6.0, in versioned directories under [IDNA-Table]. It only provides test cases for UseSTD3ASCIIIRules=true.

To test for conformance to UTS46, on each line, perform the toASCII and to Unicode operations on the source string, with the indicated type. The results must match what is given in the toUnicode and toASCII columns, except for errors. In the case of errors, an implementation only needs to record that there is an error; it need not reproduce the results in […], which are only informative.

8.1 Format
The file is UTF8, with certain characters escaped using the `\{XXXX}` convention for readability. Columns (c1, c2,...) are separated by semicolons. Leading and trailing spaces and tabs in each column are ignored. Comments are indicated with hash marks. The columns contain the following information:

1. type - T for transitional, N for nontransitional, B for both
2. source - the source string to be tested
3. toUnicode - the result of applying toUnicode to the source, using the specified type from Column 1; or an error list
4. toASCII - the result of applying toASCII to the source, using nontransitional; or an error list
5. NV8/XV8 - an informative field:
   1. NV8 is present if some character in the toUnicode value, after mapping, could not be in a valid domain name under any version of IDNA2008.
   2. XV8 is present when the character could not be in a valid domain name under the current version of Unicode.

Review Notes:

1. TODO Fix the description of IdnaTest to make it clearer exactly how it expects the test to be processed. (And make sure that there are no “extra rules”.)
2. TODO Fix a problem in the test where a sequence of two label is not properly tested for the BIDI rules.

Because the [IDNA2008] values for any particular version of Unicode may change at any point in time—due to the factors mentioned in the Caution statement of Section 7 IDNA Comparison—the NV8 and XV8 values may deviate from the future [IDNA2008] values for Unicode 8.0.

For readability, if the value of toUnicode or toASCII is the same as source, the column will be blank.

An error in toUnicode or toASCII is indicated by an error list of the form [...]. In such a case, the contents of that list are error codes based on the step numbers in UTS46 and IDNA2008:

- Pn for Section 4 Processing, step n
- Vn for Section 4.1 Validity Criteria, step n
- An for Section 4.2 ToASCII step n
- Bn for Bidi Rule #n from Section 2, The Bidi Rule, in Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA) [IDNA2008]
- Cn for ContextJ tests in, Appendix A.n in The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) [IDNA2008].

Thus C1 = Appendix A.1. ZERO WIDTH NON-JOINER, and C2 = Appendix A.2. ZERO WIDTH JOINER. The CONTEXTO tests are optional for client software, and not tested here.

However, these particular error codes are only informative; the important feature is whether or not there is an error.

**Acknowledgments**
Mark Davis and Michel Suignard authored the bulk of the text of this document, under direction from the Unicode Technical Committee. For their contributions of ideas or text to this specification, the editors thank Julie Allen, Matitahu Allouche, Peter Constable, Craig Cummings, Martin Dürst, Peter Edberg, Deborah Goldsmith, Laurentiu Iancu, Gervase Markham, Simon Montagu, Lisa Moore, Eric Muller, Simon Sapin, Murray Sargent, Markus Scherer, Jungshik Shin, Shawn Steele, Erik van der Poel, Chris Weber, and Ken Whistler. The specification builds upon [IDNA2008], developed in the IETF Idna-update working group, especially contributions from Matitahu Allouche, Harald Alvestrand, Vint Cerf, Martin J. Dürst, Lisa Dusseauilt, Patrik Fältström, Paul Hoffman, Cary Karp, John Klensin, and Peter Resnick, and also upon [IDNA2003], authored by Marc Blanchet, Adam Costello, Patrik Fältström, and Paul Hoffman.

References


The most interesting studies cited there (originally from Mike Beltzner of Mozilla) are:

- Decision Strategies and Susceptibility to Phishing by Downs, Holbrook & Cranor
- Why Phishing Works by Dhamija, Tygar & Hearst
- Do Security Toolbars Actually Prevent Phishing Attacks by Wu, Miller & Garfinkel
- Phishing Tips and Techniques by Gutmann.

[DemoDNChars] http://unicode.org/cldr/utility/list-unicodeset.jsp?a=\p{age%3D3.2}\p{cn}\p{cs}\p{co}&abb=on&g=uts46+idna+idna2008

[IDNA2003] The IDNA2003 specification is defined by a cluster of IETF RFCs:

- IDNA [RFC3490]
- Nameprep [RFC3491]
- Punycode [RFC3492]
- Stringprep [RFC3454].

[IDNA2008] The IDNA2008 specification is defined by a cluster of IETF RFCs:
• Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework  

• Internationalized Domain Names in Applications (IDNA) Protocol  

• The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)  

• Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)  

There is also an informative document:

• Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale  

[IDNA-Table] http://www.unicode.org/Public/idna
[NFKC_Casefold] The Unicode property specified in [UAX44], and defined by the data in DerivedNormalizationProps.txt (search for "NFKC_Casefold").


3492, March 2003.
http://ietf.org/rfc/rfc3492.txt


[Stability] Unicode Consortium Stability Policies
http://www.unicode.org/policies/stability_policy.html

http://www.rfc-editor.org/std/std3.txt

http://www.rfc-editor.org/std/std13.txt

[UAX29] UAX #29: Unicode Text Segmentation
http://www.unicode.org/reports/tr29/

[UAX31] UAX #31: Unicode Identifier and Pattern Syntax
http://www.unicode.org/reports/tr31/

[UAX44] UAX #44: Unicode Character Database
http://www.unicode.org/reports/tr44/

[Unicode] The Unicode Standard
For the latest version, see:
http://www.unicode.org/versions/latest/

[UTR36] UTR #36: Unicode Security Considerations
http://www.unicode.org/reports/tr36/

[UTS18] UTS #18: Unicode Regular Expressions
http://www.unicode.org/reports/tr18/

http://www.unicode.org/reports/tr39/

Modifications

The following summarizes modifications from the previous revisions of this document.

Revision 18

- Proposed update for Version 10.0.
- Section 4 Processing

http://www.unicode.org/reports/tr46/tr46-18.html
- Added three new parameters for processing: CheckHyphens, CheckBidi, and CheckJoiners. These parameters allow implementations to reflect current practice in browsers.
- Added two review notes related to testing joiners, for emoji and UAX 31.

- **Section 7 IDNA Comparison**
  - Updated the table of values.
  - Moved the Caution note below the table, included the number of “missing” IDNA2008 characters (26,568), and reworded for clarity.

- **Section 8 Conformance Testing**
  - Added two review notes with TODOs for fixes to the test file and its documentation.

Previous versions and their modifications can be accessed with the "Previous Version" link in the header.

© 2017 Unicode, Inc. All Rights Reserved. The Unicode Consortium makes no expressed or implied warranty of any kind, and assumes no liability for errors or omissions. No liability is assumed for incidental and consequential damages in connection with or arising out of the use of the information or programs contained or accompanying this technical report. The Unicode Terms of Use apply.

Unicode and the Unicode logo are trademarks of Unicode, Inc., and are registered in some jurisdictions.