L2/17-176

Re: Proposed IdnaTest.txt Revisions for v11.0
From: Mark Davis & Markus Scherer
Date: 2017-05-27
Draft: link

With the additional flags in UTS #46: Unicode IDNA Compatibility Processing, the IdnaTest.txt file is no longer able to express the variety of options that people may test for. This document proposes a somewhat different structure that should allow for that variety, providing both the result and the status codes for each conversion operation, rather than only providing the result if there are any status codes. That allow the testing structure to test for precisely the results of their combination of supported flags, by filter out status codes that correspond to a false input flag.

Old Format:

```
#  Column 1: type -           T for transitional, N for nontransitional, B for both
#  Column 2: source -         The source string to be tested
#  Column 3: toUnicode -      The result of applying toUnicode to the source,
#                             using nontransitional.
#                             A blank value means the same as the source value;
#                             a value in [...] is a set of status codes.
#  Column 4: toASCII -        The result of applying toASCII to the source,
#                             using the specified type: T, N, or B.
#                             A blank value means the same as the toUnicode value;
#                             a value in [...] is a set of status codes.
#  Column 5: idna2008Status - 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 informative values only.
…
```

New Format:

```
#  Column 1: source -          The source string to be tested
#  Column 2: toUnicode -       The result of applying non-transitional toUnicode to the source.
#                              A blank value means the same as the source value.
#  Column 3: toUnicodeStatus - A set of status codes, each corresponding to a particular test.
#                              A blank value means [].
#  Column 4: toAsciiN -        The result of applying non-transitional toASCII to the source.
#                              A blank value means the same as the toUnicode value.
#  Column 5: toAsciiNStatus -  A set of status codes, each corresponding to a particular test.
#                              A blank value means the same as the toUnicodeStatus value.
#  Column 6: toAsciiT -        The result of applying transitional toASCII to the source.
#                              A blank value means the same as the toAsciiN value.
#  Column 7: toAsciiTStatus -  A set of status codes, each corresponding to a particular test.
#                              A blank value means the same as the toAsciiNStatus value.
…
# Implementations that allow values of particular input flags to be false would remove
# the corresponding status codes listed in the table below.
#
# VerifyDnsLength:    P4
# CheckHyphens:       V2, V3
# CheckBidi:          V8
```

```
# CheckJoiners:       V7
# UseSTD3ASCIIRules:  U1
# Idna2008:          NV8
```

We would also remove the value XV8, which is not very useful in practice.

The following illustrate the differences between the old and new format. The set of examples is *not* exhaustive, but shows how there is more information available for the same examples.

Old **sample lines:**

```
T;  Faß.de;            faß.de;           fass.de
N;  Faß.de;            faß.de;           xn--fa-hia.de
B;  Bücher.de;         bücher.de;        xn--bcher-kva.de
B;  à\u05D0;           [B5 B6];          [B5 B6]
B;  a。。b;            [A4_2];           [A4_2]
```

New **sample lines:**

```
Faß.de;     faß.de;        [];       xn--fa-hia.de;     ;        fass.de;
Bücher.de;  bücher.de;     [];       xn--bcher-kva.de;  ;        ;
à\u05D0;    àא;            [B5 B6];  xn--0ca24w;        ;        ;
a。。b;      a..b;          [A4_2];   a..b;              ;        ;
```

Text:

The section http://unicode.org/reports/tr46/#Conformance_Testing will also need to be changed to reflect the above changes.

Issue:

The removal of the first field causes some of the lines to bidi-reorder. This happens already in the old format, but is much more frequent with the new one, since the first character on the line might be RTL. This does not affect the validity of the data, but can make it harder to read. One possibility would be to add a LRM at the start and end of each field. That would, however, require the test implementations to filter out those characters.