RFC Baby Steps

Adding UTF-8 Support

Tony Hansen
IETF 83
March 27, 2012
Current Restrictions

You all know this

• Line Printer Image
  – 66 lines per page
  – 72 characters per line
  – Form feed page separators
  – No overstrikes (no backspaces)
  – ASCII (7-bit) only
Alternate Versions
Currently Supported

• Alternate Renderings
  – Postscript
  – PDF
  – HTML

• Alternate Source Formats
  – XML
  – Nroff
Unicode Support

• Plethora of formats
  – UCS-2, UCS-4, UTF-16, UTF-8
  – BE vs LE, Byte Order Mark (BOM)
  – UTF-8 renders ASCII as ASCII and uses 8th bit for non-ASCII
  – UTF-8 is arguably most common

• Document Representation Used by IETF
  – \texttt{Uxxxx}, like \texttt{U2265} for GREATER THAN OR EQUAL TO
Several Proposals in the Past

• Redefine .txt files to allow UTF-8 everywhere
• Allow .txt files to allow UTF-8 in some places
  – draft-hoffman-utf8-rfcs-06 (-00 in 2005, -06 in 2010)
A Modest Proposal

• Add another source format that permits UTF-8
  – File extension of .utf8 (.utf ?)
• Generate .txt from .utf8 files
• Represent those Unicode characters in the .txt version as U<#####>
  – Does not conflict with current code point descriptions in any RFC or I-D
  – Need to relax line length restriction in .txt versions

• Complementary to other issues (HTML, PDF, ...)
• Allows .utf files to be searched just like .txt files
• Provides a test bed for possibly redefining .txt in the future
Tool Changes

• Xml2rfc, microsoft word template, nroffedit, etc. to generate UTF-8 files
• I-D upload accept .utf8 files as alternate input format
  – Could generate .txt versions directly from .utf8
  – Alternately, the .txt input could check for UTF-8 input and store those as .utf8 and generate the .txt equivalents
• tools.ietf.org/html would look for .utf8 files