draft-ietf-iri-3987bis-08

Issues Overview

IETF 82, Taipei
IRI WG Meeting
2011-11-17
Martin J. Dürst, co-Editor
Overview

• Overview/Background
• Issues Closed Recently
• Open Issues
• Requests for Review
Background

• IRI: Internationalized Resource Identifier, currently [RFC 3987](http://tools.ietf.org/html/rfc3987)
• Internationalized (i.e. not-ASCII-only) version of URI ([STD66, RFC 3986](http://tools.ietf.org/html/rfc3986))
• Updating [draft-ietf-iri-3987bis-08.txt](http://tools.ietf.org/html/draft-ietf-iri-3987bis-08.txt)
• List of open issues at: [http://trac.tools.ietf.org/wg/iri/trac/report/1](http://trac.tools.ietf.org/wg/iri/trac/report/1)
IRI Examples

- http://بوابة.تونس/
Please Don’t Forget

• URIs/IRIs are a META-syntax
• Many pieces with different requirements get thrown together
• URIs/IRIs can be:
  – Absolute, complete from scheme to fragment id
  – Relative, just one or a few pieces
  – User-oriented (short, memorable)
  – Back-end (long, complicated)
Issues Closed Since Last Meeting

• @@@Issue #35: Allow generic scheme-independent IRI to URI translation (please review!)
Issues Closed Recently

(time-frame: Oct/Nov 2011)

- Issue #24: sort out issues between IRIs and HTML form upload
- Issue #6: Allow "best practices for BIDI IRIs" to be separate document
- Issue #101: Section 5.1 should only list "garbage"
- Issue #102: ABNF notation needs improvement
- Issue #38: Handling of backslash ("¥")
- Issue #43: How to with illegal and disallowed IRI characters
- Issue #46: Normative length limits on IRIs or components theroff
- Issue #66: Allow %-encoding to punycode conversion for ireg-name?
- Issue #34: Incomplete sentence in Section 3.7: "and in general there This section gives"
- Issue #39: Warn about mistaken conversion for non-BMP characters
- Issue #86: Add reference to RFC 5137
- Issue #90: Pct-encoding of non-ASCII chars
- Issue #95: Update reference to UCS standard (ISO 10646)
- Issue #9: "URL" can be used in "formal" documents
- Issue #87: Mention that octets in <> are denoted in hex
- Issue #94: Add reference to RFC 6082
Open Issues:
Document Status

• Issue #105: recycle at Proposed or do not allow characters previously allowed

• Proposal: Now that we have only Proposed and Full, recycle at Proposed anyway
Issue #3: verify additional errata reported by Ian Hickson on draft-duerst-iri-bis-06

- Proposal:
  - Split up into smaller issues
  - Check where they belong (3987bis or iri-processing)
  - Address individually
Terminology

• Issue #85: Align terminology in 3987bis with 3536bis
  – Proposal: Accept in principle, defer to editors

• Issue #5: Separate IRI from "presentation of IRI" as concepts
  – Background: Proposal by Larry is for overreaching change, not wordsmithing
IRI <-> URI Conversion

- Issue #13: fixes to IRI parsing algorithm for clarity (!?)
- Issue #14: update URI to IRI to match IRI to URI
- Issue #92: Sections 3.4 & 3.5 should be moved into Section 3.6 of 3987bis
Character-Related Issues  
(Section 6, Characters Disallowed or Not Recommended in IRIs)

• Caution: This section was a subsection of LEIRI, carefully negotiated with W3C XML Core WG
• Issue #93: clarify which characters are disallowed vs. not recommended
• Issue #103: u+FFFE and u+FFFF missing in section 6
• Issue #104: Characters are still excluded from URIs
• Alternatives: Move back to LEIRI section OR rewrite carefully to be general
Character-Related (other)

• Issue #89: **Pct-encoding of what? (what chars should be pct-encoded in IRIs)**
  – Lots of people have questions in this area
  – Solicit some text

• Issue #26: **disallow combining characters at start of a component**
  – “Component” only defined in bidi document
  – Discourage is good, disallow may be impossible

• Issue #27: **do we need to say anything special about ZWNJ and ZWJ?**
  – Proposal: Move to Comparison document
Issue #91: Mapping of <iрег-name>: preferred way?

- Example: IRI: http://résumé.org
- URI: http://xn--rsum-bad.example.org OR http://r%C3%A9sum%C3%A9.example.org
- Repetition of Issue #35: Allow generic scheme-independent IRI to URI translation (closed)
- Text in RFC 3986 about %-encoding in <iрег-name> (end of Section 3.2.2)
- RFC 6055: IAB Thoughts on Encodings for Internationalized Domain Names
Domain Name Part:

Issue #44: Reference Unicode TR 46, and if yes, how?

• Please discuss
Other Issues

• Issue #22: *Fix "IRIs as identity tokens MUST"*
  – Proposal: Go back to mail from Roy Fielding
  – May actually apply to comparison document now

• Issue #88: *A number of ABNF issues in 3987bis*
  – Proposal: Editors to check carefully
Editorial

• Issue #47: **Practical length limits on IRIs or components thereof**
  – Proposal: Use proposed text
Please Help Reviewing

• Section 4: Use of IRIs
• Section 7: URI/IRI Processing Guidelines (Informative)
End of Presentation

Following slides are “just in case”
Bidi(rectionality) Basics

• Arabic, Hebrew,... scripts read TFEL2THGIR
  (in examples, we use ESAC REPPU for right-to-left)
• Storage is in logical order (parsing,... is easy)
• Display for running text is specified by Unicode TR 9
  – Directionality of punctuation follows surrounding letters
  – In computer syntax, stuff gets thrown around
Bidi IRI Goals

• Easily readable (for native readers)
• Easy to display (ideally no deviation from TR 9)
• Consistent conversion logical $\Leftrightarrow$ display
IRI Bidi Concepts

• Component: String between syntax characters
  – Domain name label
  – Path component
  – Query parameter name/value
  – …

• Component directionality:
  Each component clearly one way, to avoid letters jumping punctuation

• Run: Same-directionality component sequence
# Bidi IRI Ordering Alternatives

<table>
<thead>
<tr>
<th>Overall Directionality</th>
<th>Reordering by</th>
<th>Example</th>
<th>RFC 3987</th>
<th>Unicode TR #9</th>
<th>Users</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTR →</td>
<td>run</td>
<td><a href="http://ab.FE/DC/gh?ij=NM#LK">http://ab.FE/DC/gh?ij=NM#LK</a></td>
<td>okay</td>
<td>possible</td>
<td>😞</td>
<td>1</td>
</tr>
<tr>
<td>LTR →</td>
<td>component</td>
<td><a href="http://ab.DC/FE/gh?ij=LK#NM">http://ab.DC/FE/gh?ij=LK#NM</a></td>
<td>bad</td>
<td>need exception</td>
<td>😞</td>
<td>2</td>
</tr>
<tr>
<td>RTL ←</td>
<td>run</td>
<td>NM#LK=gh?ij/FE/DC.<a href="http://ab">http://ab</a></td>
<td>bad</td>
<td>possible</td>
<td>😞</td>
<td>3</td>
</tr>
<tr>
<td>RTL ←</td>
<td>component</td>
<td>NM#KL=ij?gh/FE/DC.ab//:http</td>
<td>bad</td>
<td>need exception</td>
<td>😞 ?</td>
<td>4</td>
</tr>
</tbody>
</table>

- Worst-case example, shows main design choices
- Conflict between users (and user-oriented vendors) and security concerns