draft-ietf-iri-3987bis-05 Issues Overview

IETF 80, Praha
IRI WG Meeting
2011-03-30

Martin J. Dürst, co-Editor

Overview

- Overview/Background
- Issues in Groups:

  (discussion after each group)
  - Issues closed since last meeting
  - Query Parts
  - Bugwards Compatibility (HTML)
  - Bidirectionality
  - Weed-out
  - Other issues

Background

- IRI: Internationalized Resource Identifier, currently RFC 3987
- Internationalized (i.e. not-ASCII-only) version of URI (STD66, RFC 3986)
- Updating draft-ietf-iri-3987bis-05.txt
- List of open issues at: http://trac.tools.ietf.org/wg/iri/trac/report/1
- SVN revision log: http://trac.tools.ietf.org/wg/iri/trac/log/draft-ietf-iri-3987bis/draft-ietf-iri-3987bis.xml

IRI Examples

- http://清华大学.中国, http://清華大學.中国
- http://zh.wikipedia.org/zh/中国互联网协会
- http://بوابة.تونس
- http://ja.wikipedia.org/wiki/青山学院大学

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!)
- Issue #42: Disallow '#' in fragment part
- Issue #41: Disallow single '%'
- Issue #18: rewrite Security Considerations section (move spoofing out) (please review!)
- Issue #30: check leiri definition reference to iri syntax
- Issue #23: When to require (or not) the use of a normalizing (NFC) transcoder
- Issue #20: update acknowledgements section
- Issue #29: include tag ranges in iprivate production
- Issue #53: Remove “Design Alternatives” Appendix
- Issue #52: Update reference to Unicode to Unicode 6.0

Query Part Encodings [issues #11, #24, #40]

- Conditions:
  - Document encoding other than UTF-8/UTF-16
  - IRI with query part (e.g. <a href='…?クエリー')
- Phenomenon:
  - Query part is %-encoded based on document encoding, not UTF-8

Query Part: Scheme Dependency

- Document encoding (where available):
  - http:/https:
  - What else? [Please help!]
- UTF-8:
  - mailto:
  - What else? [Please help!]
- Unclear:
  - IMAP [Please help!]
  - XMPP [Please help!]
- Schemes without query part:
  - What? [Please help!]

Bugwards Compatibility: HTML5

- [issues #1, #2, #3 and more]
- Browsers do a lot more than what the specs require
- Browser makers want to get the spec up to speed with reality

HTML Compatibility Naming

- XML: Legacy Extended Internationalized Resource Identifier (LEIRI)
- HTML:
  - Hypertext Reference (HREF)
  - Web Address
  - Legacy Hypertext Reference (LHREF)

Bugwards Compatibility Examples

- Allow single '%'? [issue #41]
- Allow '#' in fragment part? [issue #42]
- Illegal IRI characters [issue #43]
- Many others, wide variance in implementations

- Section, appendix, separate draft?
- draft-abarth-url-00 by Adam Barth

Bidirectionality

- Adapt Bidi character restrictions to IDNA2008 [issue #25]
  - Allow combining marks at end of component (no-brainer)
  - Allow digits at end of component (probably yes, issue #28)
  - Establish non-jumping restrictions for IRIs (needs work, please help)
- Overall display strategy:
  - IDNA, RFC3987: Reordering by run, LTR
  - User/vendor pressure: Reordering by component
  - Conflict between “visual security” and “usability”

Weed-out

- Section 6: Use of IRIs [Please help reviewing!]
- Section 8: URI/IRI Processing Guidelines (Informative) [Please help reviewing!]

Other Issues (except trivial)

- #5: Distinguish IRI vs. "Presentation of IRI"?
- #15: Move comparision section to separate document?
- #22: Fix "IRIs as identity tokens MUST"
- #26: No combining marks at start of component?
- #27: Anything to say about ZWNJ/ZWJ?
- #34: Incomplete sentence
- #36: Some HTTP implementations send UTF-8 paths
- #39: Warn about wrong conversion of non-BMP characters
- #45: Secure comparisons
- #46, #47: Length limits

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 ⇔ display

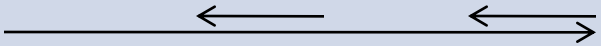
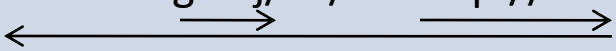
IRI Bidi Concepts

- 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

Logical: http://ab.CD/EF/gh?ij=KL#MN

Overall Directi-onality	Reordering by	Example	RFC 3987	Uni-code TR #9	Users	#
LTR →	run	http://ab.FE/DC/gh?ij=NM#LK 	okay	possible	☹	1
LTR →	component	http://ab.DC/FE/gh?ij=LK#NM 	bad	need ex-ception	☹	2
RTL ←	run	NM#LK=gh?ij/FE/DC.http://ab 	bad	possible	☹	3
RTL ←	component	NM#KL=ij?gh/FE/DC.ab//:http 	bad	need ex-ception	☺ ?	4

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