IETF 81 - Using the RFC 3986 RegExp for Parsing

Julian Reschke, greenbytes

Problem Statement

- References in web documents contain invalid URIs
- User Agents need to process them anyway
- RFC 3986's normative ABNF doesn't help with references that do not conform to it
- RFC 3986's Appendix B *does* work for any input, though
- Feedback in the W3C HTML WG was: "not normative enough"

Proposal

A standards-track document that...

- defines terminology for invalid URIs and their components
- defines parsing of "any" reference into the five URI components
- defines resolving "any" reference against "any" base reference

where...

- parsing is based on either the regular expression or a relaxed ABNF (similar to RFC 3987(bis))
- resolution is identical to RFC 3986 Section 5, applied to potentially invalid components

Problems Solved

- No needless specification of a separate parsing algorithm that may or may not conform to RFC 3986.
- Broken references can be parsed into components, and fixed based on the componentization (such as: non-ASCII in query parameters, etc).

Problems Not Solved

- Pre-processing (such as whitespace-stripping or splitting into lists of references).
- Scheme-dependent post-processing (such as fragment handling in "data" URIs, or handling of "\" in "file" URIs).

...but these things *could* be added or layered on top of this.

Links

- Test cases: <http://greenbytes.de/tech/tc/uris/>
- Internet Draft: <http://greenbytes.de/tech/webdav/#draft-reschke-ref-parsing>