Search

IRIs/RDFConceptsProposal

The issue before the working group are:

> Clarify the usage of IRI references for RDF resources, e.g., per SPARQL Query §1.2.4.

Proposed text changes

- Replace "URI reference" and "RDF URI reference" with "IRI" throughout
- Replace RDF Concepts Section 6.4, *RDF URI References* 🔗 with the following new text:

6.4 IRIs

An IRI (Internationalized Resource Identifier) within an RDF graph is a Unicode string [UNICODE] that conforms to the syntax defined in RFC 3987 🔗 [IRI]. IRIs are a generalization of URIs [URI]. Every absolute URI and URL is an IRI.

IRIs in the RDF abstract syntax MUST be absolute, and MAY contain a fragment identifier.

Two IRIs are equal if and only if they are equivalent under Simple String Comparison according to section 5.1 of [IRI]. Further normalization MUST NOT be performed when comparing IRIs for equality.

NOTE: When IRIs are used in operations that are only defined for URIs, they must first be converted according to the mapping defined in section 3.1 of [IRI]. A notable example is retrieval over the HTTP protocol. The mapping involves UTF-8 encoding of non-ASCII characters, %-encoding of octets not allowed in URIs, and Punycode-encoding of domain names.

NOTE: Some concrete syntaxes permit relative IRIs as a shorthand for absolute IRIs, and define how to resolve the relative IRIs against a base IRI.

NOTE: Previous versions of RDF used the term "RDF URI reference" instead of "IRI" and allowed additional characters: "<", ">", "{", "}", "|", "\", "^", "`", "'" (double quote), and " " (space). In IRIs, these characters must be percent-encoded as described in section 2.1 of [URI].

NOTE: Interoperability problems can be avoided by minting only IRIs that are normalized according to Section 5 of [IRI]. Non-normalized forms that should be avoided include:

- Uppercase characters in scheme names and domain names
- Percent-encoding of characters where it is not required by IRI syntax
- Explicitly stated HTTP default port (http://example.com:80/); http://example.com/ is preferrable
- Completely empty path in HTTP IRIs (http://example.com); http://example.com/ is preferrable
- ./ or /../ in the path component of an IRI
- Lowercase hexadecimal letters within percent-encoding triplets ("%3F" is preferable over

“%3f”)

- Punycode-encoding of Internationalized Domain Names in IRIs
- IRIs that are not in Unicode Normalization Form C [NFC]

Notable consequences

1. The characters “<”, “>”, “{”, “}”, “|”, “\”, “^”, “`”, “‘” (double quote), and “ ” (space) were allowed in URIrefs, and are not allowed in IRIs, so any data containing these characters *unescaped* is now invalid. Data containing these characters in %-encoded form is fine.

2. There was a note stating that URIrefs are compatible with the anyURI datatype. This is no longer the case as anyURI allows the characters above, but IRIs don't, so the note is simply removed.

3. A note said: “The use of %-escaped characters in RDF URI references is strongly discouraged.” This is a problem. There are many completely reasonable URIs the cannot be expressed as IRIs without %-encoding, for example this one: http://google.com/search?q=rdf%20semantics 🖉 … I removed the note, and subsumed it into another note that discourages the use of %-encoding *iff the unencoded char is allowed in an IRI*.

4. SPARQL 1.1 Query should update Section 4.1.1. Perhaps just drop the second paragraph.

This page w as last modified on 26 May 2011, at 19:13.