

Mulgara - Bug #197

SPARQL results char encoding

09/01/2009 07:59 PM - Gregg -

Status:	New	Start date:	
Priority:	Urgent	Due date:	
Assignee:	Paula Gearon	% Done:	0%
Category:	Mulgara	Estimated time:	0.00 hour
Target version:	SPARQL Query Engine		
Resolution:			

Description

The SPARQL Protocol definition says that for the HTTP binding "the whttp:outputSerialization is application/sparql-results+xml with UTF-8 encoding, application/rdf+xml with UTF-8 encoding." ([Section 2.2](#)) That's for XML results; I haven't found the equivalent requirement for JSON output, but for my application in any case full utf-8 support for json and xml is essential.

As a general matter (principle of least surprise), I think the expected behavior would be "encoding-in equals encoding-out", so if I populate a graph with utf-8 data, query results should be utf-8, no matter the output serialization. Alternatively, one could argue that the standard HTTP 1.1 Accept-Charset header should govern; since it is an HTTP binding, HTTP rules should apply.

The SPARQL Protocol definition doesn't explicitly address character encoding for the SOAP binding, but since SOAP is an HTTP protocol it should probably do utf-8 or honor the Accept-Charset header.

History

#1 - 09/03/2009 02:43 PM - Gregg -

Oops. I guess that should read something like "the charset parameter of the Accept header should govern"; e.g. "Accept: application/sparql-results+json; charset=UTF-16". Accept-Charset doesn't apply.

Also, [RFC 4627](#), "The application/json Media Type for [JavaScript](#) Object Notation (JSON)", says (Section 3, Encoding) "JSON text SHALL be encoded in Unicode. The default encoding is UTF-8." So it's not the sparql definition but the json definition that requires unicode and, effectively, utf-8.