RDF Validation requirements for data about products, services and companies

Mark Harrison, Auto-ID Lab, Institute for Manufacturing, University of Cambridge, UK, CB3 0FS
(on behalf of GS1 Digital initiative)

mark.harrison@cantab.net

Recently, a number of manufacturers, brand owners and retailers are showing interest in the use of Linked Data technology to make information about products, services and offerings much more discoverable on the web. GS1 [1], a global user-driven standards organization that develops global open standards to improve supply chain efficiency, has recently launched the GS1 Digital initiative to explore ways in which web technology and Linked Data in particular can help this user community to publish their own authoritative information about products and services online. The published data much be trustworthy and readily accessible to search engines and other software applications, including mobile apps. The aim is to help retailers, brand owners and manufacturers engage more effectively with current/potential consumers of their products and to help consumers find the products that best match their needs, as well as online services and support that help them in their usage of products. In this paper, we identify a number of areas in which robust validation of linked data about products would be helpful.

EU Food labelling legislation (EU 1169/2011) [2] requires information that is shown on the packaging / label of a food product to be available online prior to purchase when that product is offered for sale online. This has led many retailers to expand the amount of information shown on product-specific web pages to include not only an image and description but also lists of ingredients, known allergens, nutritional information and other accreditations, both ethical (e.g. fair trade) and environmental (sustainably sourced, organic, etc.). Although the legislation does not currently mandate that the online data must be easily machine-readable, there is an opportunity to use RDFa and appropriate web vocabularies to go beyond mere compliance with legislation and ensure that details about the product can be efficiently indexed by search engines.

For consumers with specific allergies or dietary requirements, linked data markup of product-specific web pages could enable a number of online or mobile apps that meet their needs and assist them when shopping in physical stores or online. For example, a consumer with coeliac disease could scan the bar code of a product and be warned if the product contains any incompatible ingredients, such as wheat gluten. Furthermore, the consumer could be advised of alternative products available in the same store, locally or online that don't cause a problem. Similarly, an ethical consumer who prefers to buy responsibly sourced products that have less detrimental environmental impact can scan a product and be informed about products that better match their own ethical preferences and criteria.

There are a number of aspects for which it is important to have trust that the data is authentic, complete, plausible and correctly formatted:

1) We might want to check for completeness of the data. If legislation specifies a number of product attributes whose value must be specified online, RDF validation techniques could be useful for checking that those RDF triples are present and do not contain null values. This could be used as one indicator in a test for conformance with specific legislative requirements.

2) We might want to check that correct units of measure are specified and that the numeric values seem plausible. This is especially important for nutritional information, in order to perform meaningful comparisons across products and also correctly calculate the percentage of RDA / GDA values contained within a typical serving of a product.

3) We might want to check that the values of some attributes are from controlled vocabularies or code lists. For example, units of measure might be expressed in terms of the UN ECE common code [3] such as 'GRM' for gram or perhaps more usefully a URI resource in the Quantities, Units, Dimensions and Types vocabulary [4], such as http://qudt.org/vocab/unit#Gram. Likewise, the values for product categories, attributes and vocabularies might be expected to take values from code lists in standardized vocabularies such as Global Product Classification (GPC) [5] or the United Nations Standard Product and Services Code (UNSPSC) [6]

4) We may wish to define a number of standard relationships between products and documents or services. For example, it could be helpful to define a relationship such as 'hasInstructionManual' that can consistently linked the product's global identifier (such as its barcode number (Global Trade Item Number or GTIN [7]) in URI format to a URL where its instruction manual can be found. Other plausible relationships might include 'hasWarrantyRegistrationSite', 'hasUserForum' etc. In these examples, it might not be sufficient to merely check that the RDF triple is present and contains both an attribute (predicate) and value (object); we might also want to check that the URL value does not return an HTTP 404 (Page Not Found) error [8].

5) The brand owner or manufacturer should normally be considered as the authoritative source of information about products that they produce. At present, online consumers still buy primarily from retailers rather than directly from manufacturers. All too often, the descriptions and specifications of the same product can be found to have variations when these are viewed across the websites of different retailers, which indicates that there may be some modification or manual re-entry of data. For some kinds of information, such as technical specifications, ingredient lists and nutritional information, it is preferable that correct authoritative information is published by the brand owner in a standardized format (with embedded linked data) and that this can be accessed and replicated by any retailer or reseller of a product, through inclusion within a web page. If there is a strong requirement to demonstrate and verify that such information is exactly the same as that which was published authoritatively by the brand owner or manufacturer, we might consider the use of digital signatures applied to a canonicalized representation of a specified set of triples [9]. Although different retailers might use different cascading stylesheets (CSS) [10] to subtly alter the visual appearance of such information within

a web page, the essential facts could be verified to be authentic, by comparison with the cited information from the website of the manufacturer or brand owner.

From these examples in the domain of product data, we can identify the following potential requirements for robust RDF validation:

	Validation requirement	Example
1	Cardinality constraints	Check that mandatory attributes are specified, with non-null values
2	Cardinality constraints within complex data structures such as Quantitative Values	Ensure that required units of measure are specified precisely, rather than left ambiguous.
3	Range checking for codified values. Values of specific attributes are members of a defined code list or controlled vocabulary	Ensure that values for units of measure or product categories, attributes and values take code values or URIs from a defined code list or controlled vocabulary, where this is expected.
4	Checking for current reachability of online resources (when the Object of an RDF triple is a URL)	Use RDF validation to check for broken hyperlinks
5	Canonicalization of a specified ordered set of RDF triples in order to use digital signatures to detect any deviations from a corresponding set of RDF triples from a source that is considered to be authoritative	Although anyone can say anything about anything, there is a strong desire in the industry to have trustworthy authoritative master data about products, which can be faithfully embedded / included within the web pages of retailers and resellers. There are potential liability issues that could arise from inaccuracies in product information especially if a product's formulation changes in a way that introduces an ingredient that could cause an allergic reaction

Many of these requirements have already been identified in the W3C document 'Examples of RDF Validation' [11] but we hope that this position paper explains how these are relevant to the domain of Linked Open Data for Products and Services.

Global standards have already been developed for master data about products, services and parties (companies) [12][13], enabling this information to be synchronized in a business-to-business context. The GS1 Digital initiative is also examining how the Global Data Synchronization Network (GDSN)[12] and GS1 Source [14] can be leveraged as an data source

for tools that enable companies to export their own master data about their company or their products as Linked Data, perhaps even as HTML snippets with embedded RDFa annotations, somewhat like existing snippet generator tools [15], but with the details already populated using authoritative master data that is already present in GDSN or GS1 Source.

Because the authoritative master data may be already available in XML format and could be transformed via XSLT [16] or XQuery [17] templates into RDF or HTML snippets with embedded RDFa [18] annotations, an open question is whether validation of RDF is actually necessary in our scenario, or whether it is sufficient to use existing validation mechanisms such as XML Schema Definition (XSD) [19] to validate the source data when it is in XML format, and to use W3C Provenance standards [20] to point to that source data and the subsequent transformations. We illustrate this alternative approach to validation in Figure 1 below.

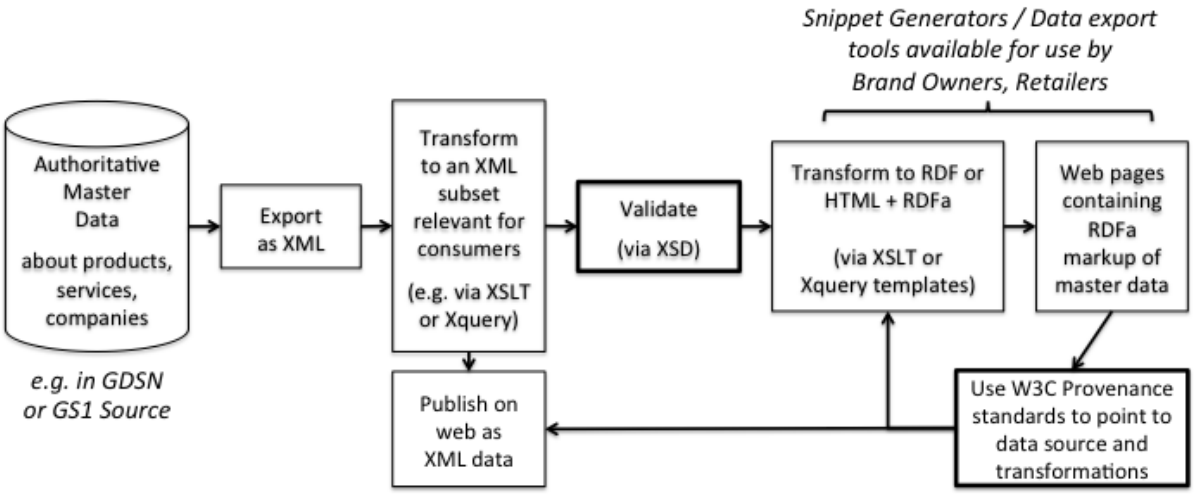

Figure 1: For master data that is already available as XML, it may be sufficient to rely on XML Schema (XSD) to validate and to use W3C Provenance standards to refer back to the XML source data and the transformations (e.g. XSLT or XQuery templates) that generated the RDF output or HTML + RDFa markup.

References

[1] GS1 http://www.gs1.org

[2] EU Regulation 1169/2011 on the provision of food information to consumers
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32011R1169:EN:NOT

[3] UN/ECE Recommendation No. 20 - Codes for Units of Measure
http://www.unece.org/fileadmin/DAM/cefact/recommendations/rec20/rec20_00cf20a1-3e.pdf

[4] Quantities, Units, Dimensions and Types (QUDT)
http://www.qudt.org/

[5] Global Product Classification (GPC)
http://www.gs1.org/gdsn/gpc
http://gpcbrowser.gs1.org

[6] United Nations Standard Products and Services Code (UNSPSC)
http://www.unspsc.org/

[7] GS1 Global Trade Item Number [GTIN]
http://www.gs1.org/barcodes/technical/idkeys/gtin

[8] Hypertext Transfer Protocol (HTTP/1.1)
http://www.w3.org/Protocols/rfc2616/rfc2616.html

[9] “Signing RDF Graphs” - J. J. Carroll
Lecture Notes in Computer Science, Vol. 2870, p369-384 (2003)
http://link.springer.com/content/pdf/10.1007%2F978-3-540-39718-2_24.pdf

[10] Cascading StyleSheets
http://www.w3.org/Style/CSS/

[11] W3C document ‘Examples of RDF Validation’
http://www.w3.org/2012/12/rdf-val/SOTA

[12] GS1 Global Data Synchronization Network (GDSN) standards
http://www.gs1.org/gsmp/kc/gdsn
http://www.gs1.org/gsmp/kc/ecom/xml/gdsn_grid

[13] GS1 Trusted Source of Data standard
http://www.gs1.org/gsmp/kc/b2c

[14] GS1 Source
http://www.gs1.org/source

[15] GoodRelations Snippet Generator
http://www.ebusiness-unibw.org/tools/grsnippetgen/

[16] XSL Transformations (XSLT)
http://www.w3.org/TR/xslt

[17] XQuery - an XML Query Language
http://www.w3.org/TR/xquery/

[18] RDFa 1.1 Primer: Rich Structured Data Markup for Web Documents
http://www.w3.org/TR/xhtml-rdfa-primer/

[19] XML Schema (XSD)
http://www.w3.org/XML/Schema

[20] W3C Provenance standards
http://www.w3.org/standards/techs/provenance