Europeana and RDF data validation
A short expression of interest from a vocabulary owner

Position paper for the W3C RDF Validation Workshop – 10-11 September 2013

Antoine Isaac
Europeana
aisaac@few.vu.nl

Europeana [1] has been developing a data model (EDM) [2] for several years. This model draws heavily on the semantic web and linked data visions, but it appears that the RDF technology stack alone, besides the usual complains about scalability and complexity issues, does not catch all our business requirements, especially with respect to controlling the quality of the data we receive and distribute.

Data validation is indeed a crucial aspect of Europeana's business. Europeana aims to enable services to give access to millions of digitized cultural items, from very heterogeneous contexts. These services are based on the metadata provided by our partners, which is also provided as CC0 to any interested third party, notably via an API.

A viable service will require appropriate metadata to be provided. Especially
(i) pointers and descriptions of digitized content on the web representing cultural artifacts, including the copyrights and re-use conditions of this digitized content;
(ii) descriptive metadata for the cultural object (paintings, books...), including title, creator, etc.

We aggregate metadata from hundreds of providers, in a two-step process where data from individual providers is gathered by a first layer of 'aggregators' (projects, thematic or national initiatives...) before being sent to Europeana itself. This imposes the constraints on our EDM model to be communicated to our partners in a clear, efficient way, both in the form of (i) human-readable documentation and (ii) machine-readable specs.

RDF is a very useful model for enabling richer forms of data. And to support semantic web application that would use EDM, we provide with an RDFS/OWL version of our vocabulary [3]. However, the lack of consensual complete RDF validation technology stack proves a hindrance to further RDF adoption. Especially, even though our data constraints are simple, they cannot be captured in an appropriate way because of the open world assumption that currently grounds standard RDF data modeling solutions.

We thus decided to keep to a "traditional" option to data ingestion and validation, based on XML and XML Schema. At the time, Europeana's existing, tools as well as many other tools being worked with in our network, were based on XML, so one cannot say that the lack of data validation mechanism was the only rationale for our decision. Yet, as there was just no suitable (standardized) alternative available, we would have needed to go for a non-RDF solution anyway next to an RDF data pipeline. RDF clearly became less a priority for us then, at least with respect to the

core data manipulation chain. We created our XML Schema [4] so that valid data according to our schema also happens to be valid RDF/XML data, which allows RDF triple stores to load it. But this is quite a meager consolation in the eye of a semantic web enthusiast.

Realizing that XML-based solutions were clearly sub-optimal to meet our requirements makes things even worse. Expressing some of our (cardinality) constraints in basic XML Schema forced us to also enforce an ordering of elements in the XML data we manipulate, which was absolutely not required (and even quite harmful) from our data model perspective. Besides, turning to non-basic XML Schema mechanisms could only capture some constraints. We use Schematron rules as a complement to the XML Schema constraints, in our XML Schema, which creates some burden for the tool developers and Europeana data providers who want to build on this schema [5].

There is thus quite some awkwardness to only be able to rely on XML validation for checking the correctness of EDM data. For us there is clearly a level of data modeling that, while not belonging to "pure" open world-based considerations, does not belong either to the syntactic (file) level (i.e., "object records" as represented by self-contained XML elements with ordered sub-elements).

The level we are after is most probably the one of "dataset level", where we want to evaluate the completeness and correctness of a body of facts. Many in the community have already recognized this need, and the RDF Validation Workshop will certainly discuss the main approaches available (Stardog ICV [6], SPIN [7]).

Also, the ability to re-use and adapt sets of constraints on the same vocabulary elements seems an important requirement. EDM already has two "profiles" (one for our data providers, and one for the data maintained and distributed by Europeana after the first data provision step [8,9]). We can foresee that our partner aggregators will further extend EDM, and such extensions will probably need to de-activate some constraints or add new ones. In such a perspective, the notion of "application profiles" from the Dublin Core initiative [10] is promising. But as of now it lacks proper implementation. A well-defined, consensual RDF data validation stack could provide the basis for realizing this vision from the metadata community.

This paper, and the Europeana initiative in general, will lack the resource to submit an appropriate framework and implementation for expressing and sharing constraints, and tools that validate data based on these constraints. But as vocabulary owners, we are willing to test the proposals of other participants of the RDF Validation workshop against our requirements, as preparation for the workshop. The kind of constraints we have (cardinality, testing the membership of a statement's object resource to a vocabulary, etc) and their number should make it a manageable effort. This could be merged with their own contribution if they want.

References

[1] http://europeana.eu
[2] http://pro.europeana.eu/edm-documentation

[3] http://europeanalabs.eu/browser/europeana/trunk/corelib/corelib-solr-definitions/src/main/resources/eu/rdf, also accessible from http://www.europeana.eu/schemas/edm/
[4] http://europeanalabs.eu/browser/europeana/trunk/corelib/corelib-solr-definitions/src/main/resources/eu
[5] See "EDM validation in Oxygen" at http://pro.europeana.eu/edm-documentation, http://pro.europeana.eu/documents/900548/804dae77-8d00-4eb9-8ecd-41e290128b11
[6] http://stardog.com/docs/sdp/icv-specification.html
[7] http://spinrdf.org
[8] http://europeanalabs.eu/wiki/EDMObjectTemplatesProviders
[9] http://europeanalabs.eu/wiki/EDMObjectTemplatesEuropeana
[10] http://dublincore.org/documents/singapore-framework/