Position Paper on Metadata Standards

Alexander von Lünen

17/08/2014 02:54

Version 0.5

Introduction

This document discusses various metadata standards, their benefits and weaknesses, and which standards the FuzzyPhoto project should use. Metadata standards in the GLAM sector (Galleries, Libraries, Archives and Museums) are still quite disparate, but some standards are becoming increasingly popular.

"Harvesting Formats" vs Ontologies

A line must be drawn between so-called "harvesting formats" (HF) and full-fledged ontologies. HF in general are formats for publishing and interchanging data, independent from a specific database software etc. Ontologies, on the other hand, also *describe* the data so being published, i.e. ontologies operate at the conceptual level. Simply put, HF and ontologies differ in regard to the amount and level of metadata they incorporate. HF, as mentioned, is about publishing collection records, so metadata usually concerns things like data of creation and creator of an object, for example. Since ontologies were devised in the context of Knowledge Representation Systems, they must provide enough metadata for automated reasoning. For example, the LIDO format (see below), defines a field for the gender of an actor (such as creator or owner of an artefact). For an HF it is good enough to have different identifiers (such as male, female, unknown) to operate, whereas in an ontology there must also be rules about the "nature" of these different genders. "Nature" here referring to a set of logical constraints; for example, that "male" and "female" are mutually exclusive, i.e. when an actor is labelled as being "male", the "reasoner" (a piece of software that evaluates the ontology) can infer that the actor is not "female". This may seem trivial to humans, but in an ontology rules such as these need to be defined to allow the reasoner to work efficiently; the point here being that the strength of ontologies is to allow transitive queries. If, for example, I would query for female artists in the database, but not all of the entries have their gender field set, the only fix would be to update the gender field before I could run the query. In an ontology, however, I could define a set of rules that would let the reasoner infer that the record being looked at must be a female person, e.g. by having occupation "actress" and there being a rule in the ontology that an occupation of type "actress" refers to a female person.

So, to summarize, GLAMs might quite likely have data already that could be described as "harvesting format", usually the collection records. But it is perhaps no so likely that they have a full-blown ontology for their collection. Luckily, standards exist for both (HF and ontologies), and the remainder of his paper will discuss the best candidates and how they could be utilized in the FuzyPhoto project.

Harvesting Formats

There exists a number of HF; the main reason being that metadata about collection items have been recorded for a long time now, starting with paper cards in filing cabinets, and XML has been around for some time now acting as lingua franca of structured data exchange format.

Many different HF have been designed in the past years, usually along national and topical interests (i.e. depending on the type of collection). There seems to be a clear trend, however, to find an internationally agreeable standard. This has resulted in the LIDO (Lightweight Information Describing Objects, http://www.lido-schema.org) XML schema. Other, more national, schemas such as Museumdat by the German Museums Association (http://www.museumdat.org) have voiced their support for LIDO. It thus seems to be a promising candidate for getting an exchange format for GLAM data.

LIDO only defines a minimum set of mandatory data fields to be populated, making it rather flexible. Given that is has been developed by several museum organizations, it should be quite close to the kind of data that is frequently encountered in the GLAM sector.

Ontologies

The selection of ontologies in the GLAM sector is much more limited. For years *Dublin Core* had been the least common denominator, with various national bodies (LOC, DNB, etc) rolling their own standard. As of late, *CIDOC CRM* strives to be the gold standard in the cultural heritage sector. While quite complex, it has been closely modelled after the collection management philosophies in the GLAM area and is experiencing ever more support from it. CIDOC CRM offers a multitude of classes to describe collections and related concepts. It thus makes a good candidate for storing the project data on the search engine side.

Strategy for the project

That being said, it should be obvious that many of the project partners in the GLAM sector may very well have data in their respective proprietary collection managements systems that could be easily mapped (i.e. exported/converted) to LIDO, while it is rather unlikely that they have sufficient metadata for CIDOC CRM. There are, however, tools and manuals on the CIDOC website to transfer LIDO (or general XML data) to CIDOC CRM.

It thus makes sense to agree that the LIDO format is used as an interchange format within the FuzzyPhoto team, i.e. the project partners will deliver their data in some kind of structured data (possibly LIDO, but other formats are permissible) to the team, which will convert it to the LIDO format in order to unify the different data sets. Once in LIDO format, transforming it into other standards (such as CIDOC CRM) is straight-forward.

The plan is therefore to check with the curators a) what amount of data they have, b) in which format it is and if c) there are interfaces to get that data (data dumps into tabular structures as last resort), and finally d) how much of that data can and should be exported, i.e. do the curators want

all the data to be used, or do they think it makes sense to use it all, and what's with data that doesn't fit into the LIDO schema? These issues should be addressed in the initial phase to make sure the data is both in a unified state, and utilizes the expert knowledge of the curators as best as possible.