Bibliographic data on the Web of data or from UNIMARC record to RDF graph

Predrag Perožić
Elementary School „Fran Krsto Frankopan“, Krk (Croatia)
pperozic@yahoo.com

Abstract

This poster shows (one possible strategy for) transformation of bibliographic metadata recorded in (UNI)MARC records to RDF triplets. On the first three pictures two UNIMARC authority records from croatian union catalog are modeled according to MADS/RDF ontology. MADS/RDF is Library of Congress product and represents completely machine-understandable ontology for authorised data. On the second triplet of pictures one UNIMARC bibliographic record from croatian union catalog is modeled as frbrised bibliographic description and write down in RDF/XML syntax serialisation. This model is, more or less, my personal proposal. At the end is shown the relationship between authorised metadata (value vocabularies), bibliographic descriptions (collections of object metadata or datasets) and metadata elements sets.

Keywords: UNIMARC, RDF, FRBR, MADS/RDF

(UNI)MARC formats and Semantic Web

The Web of Data project is developing rapidly and already gathers hundreds of larger and smaller databases containing billions of RDF triplets connected in one large RDF graph (Global Giant Graph) through which it is possible to navigate from any node (resource) to any other node. However, there are still large databases that are not included in GGG. Databases whose data are not disassembled into smallest (atomic) data elements that can be machine-understandable, nor were they assigned the appropriate URIs, cannot form a RDF triplets. Their data are not readily available, it is said to have been locked in "silos" because their use requires special applications, special protocols and special authorities and knowledge of the users. Many valuable information are buried in a formats for transferring and storing data and do not contribute to the network effect of human knowledge on the web. A family of bibliographic (UNI)MARC formats is one of this "silos". Each networked data gets added value with the increase of its interconnection. However, in the case of a database where complex semantic relations between entities are presented, something more than simply an automatic translation of human readable data in RDF triplet is needed. To become a functional part of the global data network that runs on a single platform data should be more or less re-structured. It’s not always easy to do. When it comes to bibliographic data stored in one of the (UNI) MARC formats it can also mean the creation of completely new data model.

The types of resources that are intended to be described in UNIMARC/B format include virtually any cultural object. For example, if we are supposed to describe even the three-dimensional artifacts and realities (such as toys and all sorts of utility objects or artifacts), then the resource class covers all the potential objects of the real world. As we know, even the object of the natural world can occur in a context that makes it a cultural object, such as the famous antelope in a zoo that can occur as a showpiece to be cataloged.

(UNI)MARC formats are pre-web information systems in which the bibliographic databases are modeled according to the closed-world assumption. This means that the mechanism of appellation / identification of resources is unique but local. The system has no "awareness" of the existence of other resources that are out of reach. In addition, the unique identification within the format is carried out over a relatively small number of resources. At last, a large part of the data is presented informally, in natural language, because it is expected that people rather than machines are to interpret their meaning.

To respond to the emergence of new types of resources and provide new ways of linking them (UNI)MARC formats are occasionally expanded and upgraded. Theoretically, the number of fields and
subfields could be endlessly upgraded. As well as to the book that describes every day a bigger part of the world, a pages and chapters could be added. But, no matter what might be added into this format, the crucial question is whether all the bibliographic data can be represented in atomistic form that enables machine understanding. Full machine semantics is a prerequisite of widely available and arbitrarily interconnected data. This is the meaning of semantic web slogan "anyone should be able to say anything about anything". After the cry "make love not war" I believe this is the second most important phrase spoken in the past one hundred years. It's not just technology, it's a new culture.

Nevertheless, technically speaking, for this new type of functionality we do need a different technology. What is being offered from W3C’s laboratory is, first of all, different but universal data model – RDF. This model is not universal because its basic release meets everyone’s needs and wishes, but because it allows for various extensions and adjustments, so-called application profiles, which best suit the individual domains.

References and citations

This poster is abbreviated illustration of my research on the topic which was published as article in Proceedings of 15. seminar Arhivi, knjižnice, muzeji under the same title. All the numerous sources and reference are listed there.

Curriculum Vitae

I was born in 1962 in Rijeka, Croatia.
I graduated at Faculty of Philosophy in Zadar in 1988 in the field of Philosophy. My special interest was logic, philosophy of language and analitical philosophy.
From 1995 to 2007 I worked in University Library of Rijeka as a cataloguer, mostly on subject indexing. Since 2008 I work as school librarian in Elementary School „Fran Krsto Frankopan”, on the island of Krk.
In March 2011 I joined the second generation of PhD course „Knowledge Society and Information Transfer” at University of Zadar, Department of Information Sciences.