From MARC21 and Dublin Core, through CIDOC CRM:
First Tenuous Steps towards Representing Library Data in FRBRoo

Cezary Mazurek, Krzysztof Sielski, Justyna Walkowska, Marcin Werla
Poznań Supercomputing and Networking Center
ul. Noskowskiego 12/14, 61-704 Poznań , Poland
{mazurek, sielski, ynka, mwerla}@man.poznan.pl

1. Introduction

As a part of works undertaken by Poznań Supercomputing and Networking Center (PSNC) in the frame of the SYNAT[1] project, a semantic database was built, containing information obtained from Polish digital cultural heritage institutions such as museums, libraries, archives, and scientific information systems. Different metadata formats used by those institutions have been mapped to CIDOC CRM and FRBRoo to provide interoperability. The developed knowledge base will be used to provide advanced searching and browsing features in a prototype portal dedicated to researchers and hobbyists interested in Polish cultural and scientific heritage.

In the paper we discuss the issues concerning (automatic) mapping of MARC and PLMET metadata records to FRBRoo, comment on FRBRoo's compatibility with both FRBR and with CIDOC CRM, and discuss the consequences FRBRoo has had on the Semantic Web knowledge base. The next section provides an overview of the source data used in the knowledge base building process and its mapping to CIDOC CRM. Section 3 describes further transition to FRBRoo. The paper ends with discussion on most interesting issues encountered during the mapping development, followed by some final conclusions.

2. Source data and mapping to CIDOC CRM

A large portion of the data for the described knowledge base comes from aggregated metadata describing objects from digital collections of Polish libraries, museums and archives. Currently, the main data sources are Polish Digital Libraries Federation (DLF, http://fbc.pionier.net.pl/) and NUKAT – a union catalogue of Polish research libraries (http://www.nukat.edu.pl/).

The Polish Digital Libraries Federation provides metadata in PLMET and ESE (Europeana Semantic Elements) schemas. Both of them extend Dublin Core with some additional tags from well-

[1] SYNAT is a research project aimed at the creation of universal open repository platform of networked resources of knowledge for science, education, and open society. It is funded by the National Center for Research and Development (grant no SP/I/1/77065/10).

established namespaces (such as Dublin Core Metadata Terms or Electronic Thesis and Dissertation Metadata Standard) and proprietary tags. Those schemas should be regarded as flat, i.e. each metadata record consists of only one level elements with no direct cross-reference records.

NUKAT library catalogue describes records in MARC XML format. This schema is more complete and less ambiguous since it organizes information into hierarchically split atomic pieces, but is more complex (related data can often be found in seemingly unrelated elements) and not very human-readable (many types of information are encoded). NUKAT catalogue consists of bibliographical records (describing books, periodicals, maps, movies etc.) and authority records (describing subject headings or organization/person names) which are cross-referenced.

We chose to map all aggregated source data to CIDOC CRM to have a coherent representation of resources no matter whether they originate from libraries, museums, archives or other institutions. For the first major portion of library data (provided by DLF) this approach was sufficient, even though automatic mapping to CIDOC CRM proved to be a challenge due to differences in the interpretation of fields in the flat metadata schema in particular institutions [5].

The resulting knowledge base is represented in RDF format (stored in Owlim SE triplestore) using an OWL DL 1.0 implementation of the CIDOC CRM ontology: Erlangen CRM. We also introduced some extensions to this ontology [5], most of which target at accurate representation of geographical data which allows us to use advanced geo-spatial reasoning provided by Owlim. In order to map source data from various formats to RDF, a dedicated tool jMet2Ont (http://fbc.pionier.net.pl/pro/jmet2ont/) has been developed along with appropriate mapping rules for MARC XML and PLMET formats.

3. Moving to FRBRoo

In [5] we discussed the applicability of CIDOC CRM in a semantic knowledge base containing information about cultural heritage resource, including resources from digital libraries. The details of the mapping from the Dublin Core based schema used in the Polish Digital Libraries Federation were presented in [4]. However, to fully represent the properties of cultural heritage objects in the knowledge base we had to provide extensions of some of the CIDOC CRM classes, especially when it comes to events (subclasses of E5 Event).

The next level of analysis revealed that a number of necessary subclasses we have introduced have already been defined in FRBRoo, an ontology that extends CIDOC CRM with entities from the FRBR model [3]. We have refrained from using FRBR at the early stages of the SYNAT project, mostly because we were afraid that the often simplistic records from digital libraries would be impossible to translate into the 4-tier model of FRBR, where a publication consists of: Work (*a* *distinct intellectual or artistic creation*), Expression (*the intellectual or artistic realization of a Work*, e.g. a translation or edition), Manifestation (*the physical embodiment of an Expression*, i.e. the set of all copies), and Item (*a single exemplar of a manifestation*). In metadata records of digital libraries it is often difficult to differentiate between the physical and electronic resource, and the FRBR levels introduce additional complexity.

The physical vs. electronic resource problem has been also put into spotlight recently by works related to Europeana [2]. The Polish Digital Libraries Federation is a national metadata aggregator which passes the metadata on to Europeana. The new Europeana metadata schema called Europeana Data Model (EDM) introduces two disjoint classes called Provided Cultural Heritage Object (ProvidedCHO) and Web Resource. The former represents the physical resource, the latter its digitized representation (obviously the situation is more complicated with born-digital resources). Accordingly, the PLMET (new DLF metadata schema) guidelines call for descriptions of the original resources, as they are more significant to library readers searching for information.

The shift from proprietary CIDOC CRM extensions to FRBRoo was catalyzed by the inclusion of NUKAT (the union catalogue of Polish research libraries) data in the knowledge base. NUKAT describes their data with MARC 21, a format which is not very human-friendly, but represents information in a more detailed and unambiguous way than PLMET and other Dublin Core Application Profiles. Another added value from switching to FRBRoo (and thus adding the Work level, not present in our original Information Object/Information Carrier two-tier representation) is a more natural way to describe periodicals (with the F18 Serial Work class).

Finally, we have created two mapping definitions: PLMET to FRBRoo and MARC 21 (as used in Poland) to FRBRoo. The mapping rules have been externalized to XML compatible with the jMet2Ont tool – this way they are not hardcoded in a computer program and can be assessed by library experts. The mapping definition files are too big to be included in this paper, but below in Figure 1 you will find an example of a PLMET record from the Digital Library Federation, and Figure 2 presents a graphical form of the same data mapped to FRBRoo.

The next section summarizes some of the most interesting issues we encountered while mapping to FRBRoo and using the ontology in our knowledge base.

4. Practical difficulties

The first issue is the way of representing contributors. The author is the one who conceives the Work, but often the library metadata record also lists contributors, such as translators, editors, illustrators and so on. We have not reached any agreement yet on whether all of them should be treated as entities responsible for the creation of the FRBOoo Expression, or maybe some of them actually contribute to the FRBRoo Work.

```
<plmet:record>
  <dc:title xml:lang="pl">
    Zadania ochronne ubrania strażackiego przeznaczonego do akcji
    przeciwpożarowej
  </dc:title>
  <dc:title xml:lang="en">
    Protection tasks of firefighter clothes destined for firefighting
    action
  </dc:title>
  <dc:creator>inż. Mariusz Jaworski</dc:creator>
  <dc:date>2011</dc:date>
  <dc:type>artykuł</dc:type>
  <dc:language>polski</dc:language>
  <dc:subject>ubranie strażackie</dc:subject>
  <dc:description xml:lang="pl">
    Opracowanie porusza trzy zagadnienia związane z (...)
  </dc:description>
  <dc:description xml:lang="en">
    Elaboration rises three problems connected with firefighter clothes.
    (...)
  </dc:description>
  <dc:publisher xml:lang="pl">CNBOP</dc:publisher>
  <dc:identifier>oai:czytelnia.cnbop.pl:269</dc:identifier>
</plmet:record>
```

Figure 1. Example metadata record in PLMET schema with DC elements, encoded in XML.

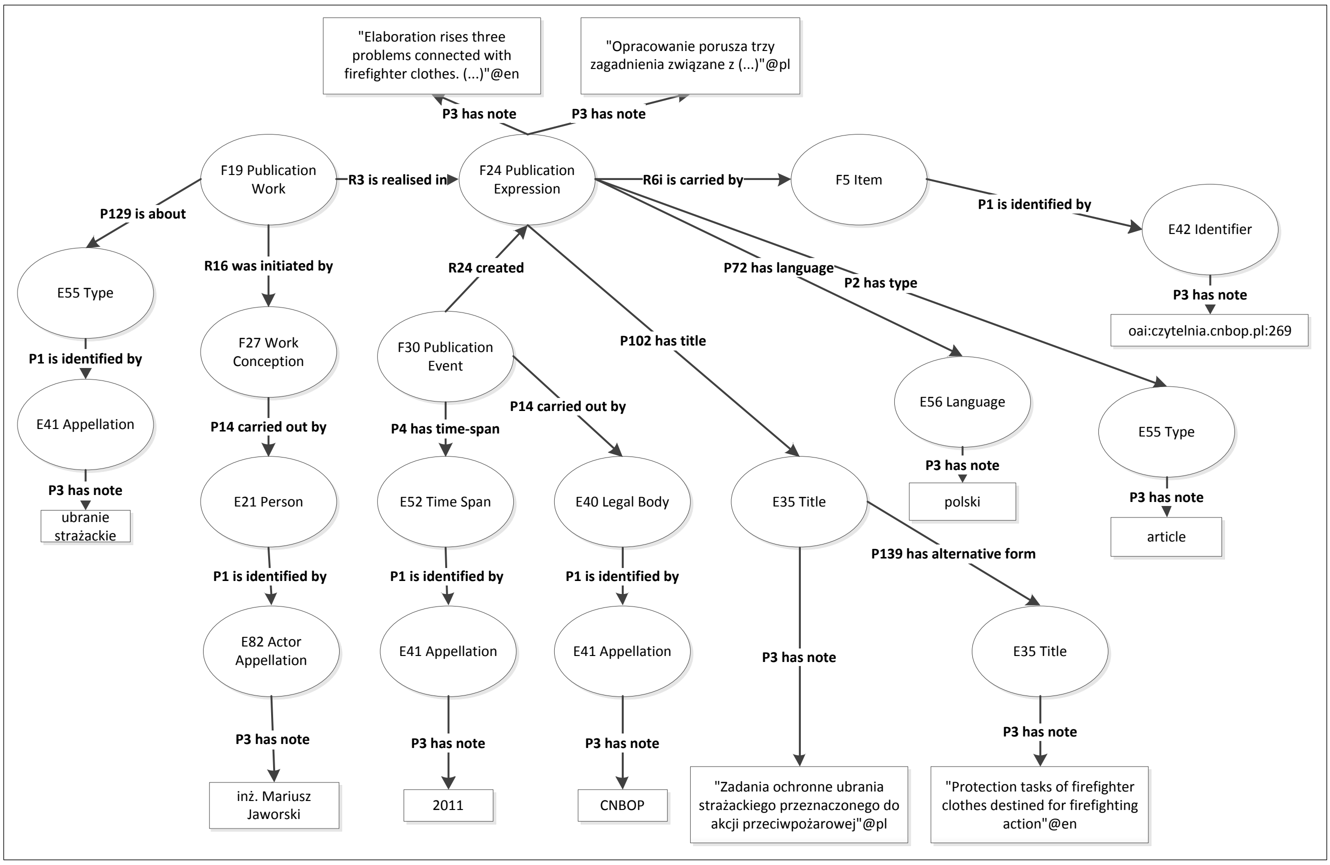

Figure 2. Graphical representation of data from PLMET record (see Figure 1) mapped automatically to FRBRoo with the use of jMet2Ont tool.

There is another, more technical but severe problem with contributors. The FRBRoo [1] specification calls for describing the contributor with the *P14 carried out* property with *P14.1 in the role of* property. However, this is a property of a predicate instance (this particular *P14 carried out by* instance has the particular role), something not allowed in RDF or OWL. To solve this, we could have introduced distinct *P14 carried out* subproperties for each contribution role but this would lead to unexpected ontology growth since there is no closed enumeration of all possible contribution roles. We decided to use another solution which involves creating subevents of the creation event: a F28a Contribution (subclass of F28 Expression Creation) has been added to the ontology and is used in the following manner: F28 Expression Creation *P9 consists of* F28a_Contribution *P14 carried out by* E39 Actor, F28a_Contribution *P2 has_type* E55 Type.

One more technical issue is the fact that FRBRoo features deep class and property hierarchies. After introducing elements of this ontology, we noticed the sudden growth in the number of triples in our knowledge base which was taking place during the forward reasoning, performed as the last stage of knowledge base construction process. Therefore we decided to apply elements of backward reasoning in the class hierarchy calculations, to avoid introducing into the repository repetitive triples about a class instance belonging to every class higher in the hierarchy.

5. Conclusions

The FRBR conceptualization makes talking about library resources and understanding their structure much easier. The FRBRoo ontology is an interesting solution to apply in repositories where museum and library resources metadata are combined as different examples of cultural heritage items. However, the automatic translation of existing digital library metadata records is not always straightforward, because often they do not draw a clear line between the four FRBR levels or between physical and digital resources. Also, it is still a challenge how to present such structured information to digital repositories end-users.

References

[1] Bekiari, Ch., Martin Doerr, M., Le Boeuf, P. (2012): FRBR: object-oriented definition and mapping to FRBRer (Version 1.0.2), http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0.2.pdf

[2] Definition of the Europeana Data Model elements, Version 5.2.3, 24/02/2012, http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22

[3] IFLA Study Group on the Functional Requirements for Bibliographic Records (2008): Functional Requirements for Bibliographic Records. Final Report, http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf

[4] Mazurek, C., Sielski, K., Stroiński, M., Walkowska, J., Werla, M., Węglarz, J. (2011): Transforming a Flat Metadata Schema to a Semantic Web Ontology. The Polish Digital Libraries Federation and CIDOC CRM Case Study. In: Proceedings of the Nineteenth International Symposium on Methodologies for Intelligent Systems, ISMIS 2011, Warsaw, Poland, Lecture Notes in Artificial Intelligence, 6804, Springer-Verlag 2011

[5] Mazurek, C., Sielski, K., Walkowska, J., Werla, M. (2011) Applicability of CIDOC CRM in Digital Libraries, CIDOC 2011, Sibiu.