Applicability of CIDOC CRM in Digital Libraries

Cezary Mazurek, Krzysztof Sielski, Justyna Walkowska, Marcin Werla
{mazurek, sielski, ynka, mwerla}@man.poznan.pl
Poznań Supercomputing and Networking Center

Abstract

CIDOC CRM is an ontology to represent cultural heritage information, especially museum collections. This paper presents the use of CIDOC CRM as an ontology describing a digital-library-originated knowledge base. An application profile is presented with some terms taken from other vocabularies, a few CIDOC classes extensions and a small set of new properties. The application profile has been used to describe a knowledge base containing information about more than 700,000 digital publications from the Polish Digital Libraries Federation, automatically translated from a Dublin Core-based metadata schema. The paper also analyzes the advantages and disadvantages of using CIDOC CRM, and presents some statistics of the resulting digital libraries semantic database.

1. Introduction

Since 2002, the Polish National Research and Educational Network (PIONIER) has provided its users and connected institutions (e.g. research centers, universities, libraries, museums) with an advanced infrastructure which, among other things, helps in building digital libraries (DL) and facilitates efficient management of their content, making it accessible online. PIONIER DL resources (currently over 700,000 digital objects) represent a part of digitized cultural heritage, so a natural question has been posed: how well is CIDOC suited to serve as an ontology describing a digital-library-originated knowledge base?

As a part of the SYNAT [14] national research project, the Poznań Supercomputing and Networking Center (PSNC) Digital Libraries Team has been working to create the Integrated Knowledge System for science. (The sources are digital museums, libraries, archives, and scientific information systems). The first step was examining the ontologies used to describe the mentioned types of resources with hope of finding one ontology expressive enough to describe all kinds of sources. As CIDOC CRM is often mentioned in the context of intermediate representations for schema mapping, its applicability was examined. This paper presents the results of the CIDOC analysis from the DL point of view. The obtained knowledge representation is stored in an RDF repository, hereinafter referred to as the *knowledge base*.

SYNAT is a national research project aimed at the creation of a universal open repository platform for hosting and communication of networked resources of knowledge for science, education, and an open society of knowledge. One of the PSNC’s responsibilities in the project is the creation of a prototype of the Integrated Knowledge System (IKS). The IKS will become a part of a four-layer infrastructure of advanced network services: source data, distributed information services, knowledge integration and front-end services layers. The knowledge integration layer serves as middleware providing access to data from distributed information services, such as digital libraries, museums, or scientific and technical information systems. To achieve this goal, a common representation of data is necessary to which the existing heterogeneous representations and schemas can be converted.

The remaining part of this paper presents the environment of Polish digital libraries and the description schemas that they use, and finally the proposition of a digital-library-dedicated CIDOC CRM application profile, together with an analysis of the results. In particular, the next section presents the situation of Polish digital libraries and the metadata schemas they use.

2. Polish Digital Libraries

Data from Polish Digital Libraries Federation [12] was the first data loaded into the Integrated Knowledge System’s semantic database and automatically translated to the extended CIDOC CRM format described later in this paper.

2.1 Popular Metadata Formats

About 90% of Polish digital libraries use the dLibra system developed at the PSNC [18]. Most libraries internally use the MARC21 cataloguing format, but publications available online are described with a variation of Dublin Core elements. This has led to problems not only with inconsistent element sets in different libraries, but also with different interpretations of basic Dublin Core elements. In 2007, a publication appeared [6] containing guidelines for the interpretation and use of the elements.

The next subsection describes the Digital Libraries Federation, which is the next step in DL resources standardization.

2.2 Digital Libraries Federation

The PIONIER Network Digital Libraries Federation (DLF, [12]) is the next stage of the development of an infrastructure of distributed digital libraries and repositories in Poland. The DLF is a set of advanced network services based on the resources available in Polish digital libraries and repositories deployed in the Polish NREN PIONIER. The resources are created by many institutions, such as universities, libraries or museums. The Digital Libraries Federation is maintained by the Poznań Supercomputing and Networking Center.

The aggregator features of DLF have enabled digital library users to search the distributed repositories of all federated libraries from one website. The DLF does not store content; after choosing a resource of interest among the search results, the user is redirected to the resource owner's website.

As of July 2011, the number of publications with metadata aggregated by the DLF exceeded 700,000. 64 Polish digital libraries are connected in the DLF, holding content from hundreds of memory institutions. The majority of the content constitutes of newspapers and magazines, mostly historical. Most publications are in Polish, the second popular language is German.

The DLF uses a metadata schema based on Dublin Core Metadata Terms (*http://purl.org/dc/terms/*) and Electronic Thesis and Dissertation Metadata Standard [1], adding a number of proprietary elements demanded by the Polish digital libraries environment. .

2.3 Shift to Ontologies

The DLF’s effort to standardize metadata formats in Polish Digital Libraries drastically improved the quality of Polish digital libraries and the search possibilities. However, that flat metadata schemas often prove insufficient to describe DL resources.

The catalogues often mix descriptions of different objects: they describe the publication as *being in the PDF format*, having the dimension of *10x15 cm*, and *written by Adam Mickiewicz*. It is quite clear that the PDF file and the physical copy are different entities, and one can argue to which one of them the author should attributed.

Another problem is that when the cataloguer does not see a metadata element to represent a piece of information, they put it in a description field, in purely textual, natural language format. Because of this the records are difficult to organize and search.

The remaining part of this paper presents solutions developed within the SYNAT national research project to find an ontology allowing to describe digital library data together with other types of cultural heritage objects, and to automatically transform data from flat and limited metadata formats.

However, some Polish libraries have already started works (manual in large parts) to move from flat metadata formats and the MARC21 classic format to ontologies: completely proprietary or based on the FRBR guidelines [9].

The next section presents a look at CIDOC as a potential DL resources description ontology.

3. CIDOC as the Main Resource Description Format for Bibliographic Data

CIDOC CRM, a mature and carefully edited ontology that it is, is not the first natural choice to describe digital libraries collections. It was chosen as the knowledge base description format for the Integrated Knowledge

System as it is crafted to represent items of cultural heritage, which books and old prints collections definitely are, but universal enough to allow for stepping outside of the museum world.

The knowledge base is stored in an reasoning-enabled RDF repository (Ontotext's BigOWLIM [2]), so the Erlangen CRM OWL-DL implementation was chosen [7]. One of the main consequences of the choice is the rejection of CIDOC insufficient primitive types (e.g. E62 String) in favour of more expressive XSD data types [8].

The remaining part of this section presents some alternative ontologies that have been considered and provides detailed information about the application profile and its use. The application profile is made up of a subset of CIDOC, CIDOC extensions (new subclasses and new properties) and of a small number of terms from existing external vocabularies.

3.1 Considered Alternatives

Before choosing CIDOC CRM, the following alternatives were analyzed to describe digital library resources in the semantic knowledge base:

- Bibliographic Ontology [5],
- bibTeX in OWL [10],
- MarcOnt [11],
- FRBR (Functional Requirements for Bibliographic Records [9]) and derivative systems, such as RDA (Resource Description and Access) [13].

CIDOC was chosen because of the best combination of:

- *universality*: it can describe not only bibliographic resource, but also works of art and other museum items,
- *simplicity*: 90 classes and 150 relations organized in a clear hierarchy does not exceed regular librarian's cognitive possibilities, it also can be mapped automatically from other schemas with acceptable correctness probability,
- *maturity*: it has been developed, maintained and used for years,
- *popularity:* CIDOC is known and understandable in circles connected with cultural heritage object online. Even without knowledge of the extensions, a person who knows CIDOC may query the semantic knowledge base and obtain satisfying results.

3.2 Application Profile and Proposed Extensions

The list below represents the hierarchy of CIDOC and CIDOC-originated classes used to describe the resources. The added classes are always subclasses of original CIDOC classes (the same is not true in case of new properties). Their symbols are created by adding subsequent letters to the symbol of the superclass (E12a Publishing is a subclass of E12 Production). Further in this section the usage of original CIDOC classes is described and the meaning and rationale of the added subclasses is explained, together with a list of relations (properties) a given class participates in.

Indentation corresponds to the subclass relation. In cases of multiple inheritance the class is shown only once, in the first place where it was encountered. Underlined classes are those there are directly instantiated during the process of knowledge base construction.

3.2.1 Application profile's class hierarchy

The frame below (List 1) presents the hierarchy of classes from the application profile. The origin of the classes is twofold. One group is made up of a subset of original CIDOC classes, the other of subclasses added to CIDOC to better distinguish DL data.

The new classes' symbols are created by adding consecutive letters to the symbol of the superclass. (E.g. the first new subclass of E12 Production is called E12a Publishing.)

```
E1 CRM Entity
    E5 Event
        E63 Beginning of Existence
```

```
            E12 Production
                E12a Publishing
                E12b Digitalization
            E65 Creation
            E67 Birth
E64 End of Existence
    E69 Death
        E7 Activity
            E10 Transfer of Custody
            E7a Becoming Available
            E7b Becoming Unavailable
            E7c Acceptance
            E7d Copyright Acquisition
            E7e Submission
            E7f Becoming Valid
            E7g Becoming Invalid
            E8 Acquisition
    E52 Time-Span
    E53 Place
    E54 Dimension
    E39 Actor
        E21 Person
        E40 Legal Body
    E71 Man Made Thing
        E24 Physical Man Made Thing
            E84 Information Carrier
                E84a Thumbnail
            E78 Collection
        E28 Conceptual Object
            E55 Type
E55a_Degree
E55b_Education_Level
E55c_Research_Discipline
E55d_Resource_Type
E55e_Subject
E55f_User_Subject
E55g_Subject_Hierarchy
E55h_Place_Type
E55j_Subject_Type
E56 Language
E57 Material
E58 Measurement Unit
            E89 Propositional Object
                E30 Right
                    E30a Access Rights
                    E30b License
            Symbolic Object
                E41 Appellation
                    E35 Title
                    E42 Identifier
                        E42a Call Number
                        E43b Citation
```

E44 Place Appellation
E45 Address
E47 Spatial Coordinates
E50 Date
E82 Actor Appellation
E82a Person Appellation
E82b Legal Body Appellation
E73 Information Object
E33 Linguistic Object
E73a Thesis
E73b Periodical

List 1. The hierarchy of classes in the proposed digital libraries CIDOC application profile

3.2.2 Class Hierarchy Discussion

The most important concept described in the digital library oriented knowledge base is a publication. In the DL understanding of the CIDOC ontology, a publication consists of:

- a physical object (E84 Information Carrier),
- the intellectual work (E73 Information Object),
- an optional digitized copy (another instance of E84),
- a title (E35 Title),
- the text of the publication (E33 Linguistic Object).

Error! Reference source not found. presents the relations between those elements. The additional instance of E33 Linguistic Object has been introduced because the E73 Information Object is not an instance of E33 and in consequence cannot be attributed a language. The E73 Information Object and E84 Information Carrier pair is the core of the knowledge base. All other objects are connected to those instances to represent detailed information about them.

One information object can have many physical carriers. An information object represents one edition of a resource. The first edition of Shakespeare’s “Hamlet” and the second edition of its Polish translation are different information objects, related to each other with subproperties of the P205 is related to property.

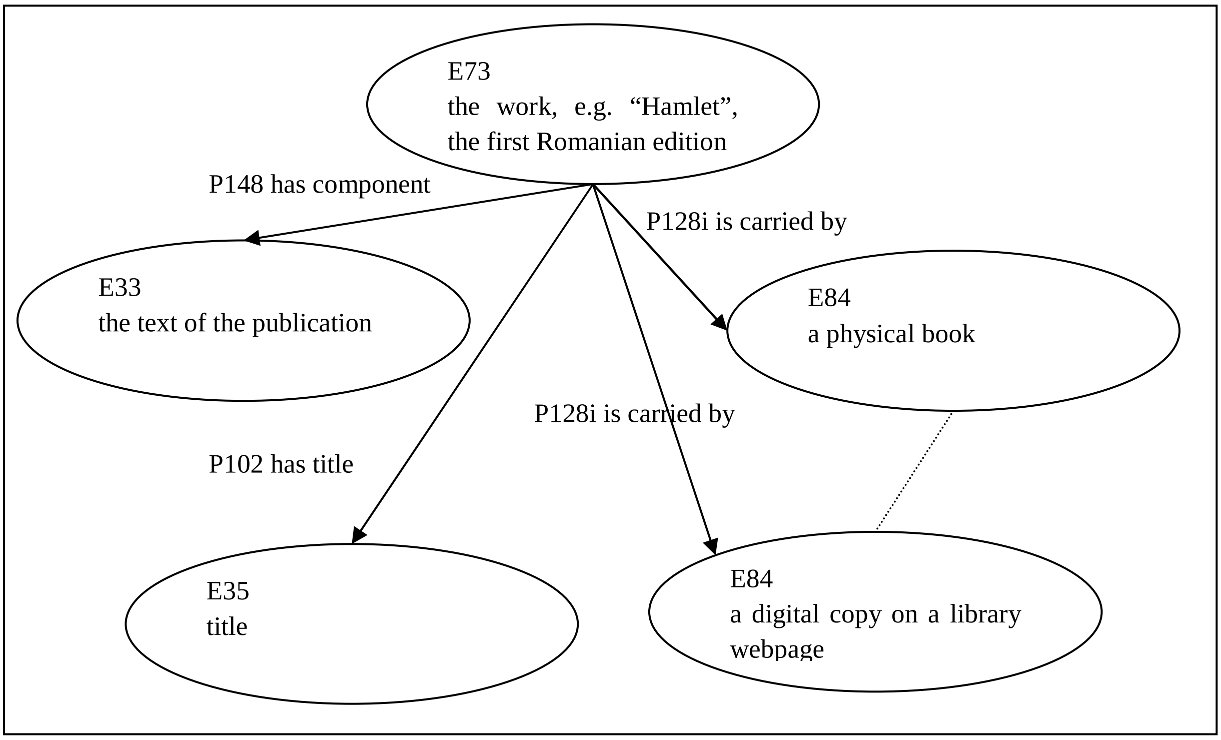
E73
the work, e.g. "Hamlet", the first Romanian edition
P148 has component
P128i is carried by
E33
the text of the publication
E84
a physical book
P128i is carried by
P102 has title
E35
title
E84
a digital copy on a library webpage

Fig. 1 The main instances describing a publication. The dotted line represents a connection through the E12b Digitalization event.

The original intention was to introduce one more level of abstraction which later was discarded as impractical – an information object without a carrier, representing all the existing editions. The advantage would be a smaller number of relations, as all editions would be only connected to this "meta-instance" and not to each other. However, this solution proved impractical after further analysis. There was no consent on what such an instance would actually represent and which properties should be attributed to it. Also, not all publications need this level of abstraction (i.e. not all have editions). Another unwanted consequence would be the complication of search queries.

In accordance with CIDOC's domain and range specifications, information objects are created (E65 Creation) by authors. A subproperty of the P14 carried out by has been introduced to represent co-authors, such as translators, editors or reviewers. In contrast to information objects, information carriers are produced (E12 Production). In the digital library world the production is either the publishing or the digitization, hence the two new subclasses of E12.

The type of the information object (e.g. a book, a journal) is represented by a special subclass of E55 Type, discussed in more detail in section 4.2.5. Nonetheless, two new subclasses have been introduced whose handling is different: the E73a Thesis and E73b Periodical. E73a is defined as the domain of a special set of new properties relevant only for academic works (see 4.4). E73b represents periodical as a whole – it is an information object grouping all issues of a journal or newspaper, so also participates in different relations than a regular E73 Information Object.

Only one subclass has been added to the E84 Information Carrier class: the E84a Thumbnail. A thumbnail is a miniature image representing (P62_depicts) a publication (the publication's information carrier). It may be used to display the resource on the search query results list.

As far as different kind of activities are concerned, the E10 Transfer of Custody and E8 Acquisition classes serve the same purpose they do when describing museum collections – they represent the exchange of physical objects (prints) between institutions. The newly added activities, E7a to E7g, represent milestones in the "life cycle" of publications and documents. They were inspired by different dates given in the DCMI Metadata Terms *date* elements and its subelements.

Other minor extensions in the class hierarchy are two separate types of E82 Actor Appellation: the E82a Person Appellation and E82b Legal Body Appellation. The separation was introduced because person appellation has different properties than group's or institution's. Two new types of identifiers (E42) have been introduced: E42a Call Number and E43b Citation. A call number identifies a publication in the collection of the keeping institution. A citation is a bibliographic citation, i.e. the way to unambiguously refer to the publication in another text (there are different citation formats).

3.2.3 Added Properties

As stated before, all classes added to the ontology are extensions (subclasses) of existing classes. This is only partly true about the added properties: some of them are subproperties of existing CIDOC properties (code names are superproperties' code names with added subsequent letters), but others are completely separate (code numbers from 200 up), with no similar elements in CIDOC. The complete list of added relations is presented in Table 1.

Table 1. The complete list of relations added to CIDOC CRM to represent DL data

Name	Extends	Domain	Range	Comments / scope note
P212 has display uri	-	E84 Information Carrier	xsd:anyURI	This is a rare situation in which a URI does not identify a resource in the ontology/knowledge base. It is only an address at which a representation of the resource can be viewed. It MIGHT be the identifier of the information carrier.
P212a is shown at	P212	E84 Information Carrier	xsd:anyURI	Mapped from europeana:isShownAt [**Error! Reference source not found.**] ("this element will be active in the portal and will provide the link to the digital object in full information context on the provider website")
P212b is shown by	P212	E84 Information Carrier	xsd:anyURI	Mapped from europeana:isShownBy [**Error! Reference source not found.**] ("this element will be active in the portal and will provide a link to the digital object on the provider website")
P220 has begin	-	E52 Time-Span	xsd:date	Begin of time span.
P221 has end	-	E52 Time-Span	xsd:date	End of time span.
P3a has tag	P3	E1 CRM Entity	xsd:string	User defined tag, mapped from dlf:userTag
P3b has abstract	P3	E73 Information Object	xsd:string	Book or paper abstract, mapped from dcterms:abstract.
P3c has table of contents	P3	E73 Information Object	xsd:string	Table of contents of a book, mapped from dcterms:tableOfContents
P3d has provenance info	P3	E73 Information Object	xsd:string	Mapped from dcterms:provenance. In future the content should be parsed to create instances of E8_Acquisition or E10_Transfer_of_Custody
P3e has unstored info	P3	E84 Information Carrier	xsd:string	Mapped from europeana:unstored. This element has been created in order to allow providers to retain all important information that cannot otherwise be mapped to ESE. The contents of this element are indexed and searched but the values do not show in the display. Care should be taken not to map several fields with similar data to avoid distorting the weighting.
P3f has ancillary	P3	E82a Person Appellation	xsd:string	Ancillary info about person, e.g. "saint", "king", "junior"

info				
P3g has first name	P3	E82a Person Appellation	xsd:string	Person's first name
P3h has last name	P3	E82a Person Appellation	xsd:string	Person's last name
P3j has information object description	P3	E73 Information Object	xsd:string	Information object description (dc:description)
P3k has information carrier description	P3	E84 Information Carrier	xsd:string	Information carrier description (in case dc:description does not concern the information object). Advanced linguistic analysis is necessary to detect such a situation automatically. Normally dc:description in digital libraries concerns the intellectual contents of the publication, but sometimes it also contains information about the state of the physical copy (“stained”, “bullet went through”).
P200 was sponsored by (P200i sponsored)	-	E12b Digitization	E39 Actor	Names the sponsor of the digitization process for digital information carriers that were created based on a physical one.
P205is related to	-	E73 Information Object	E73 Information Object	The most general level of resource relation. Mapped from dc:relation if no more specification is given. Contrary to most of its subproperties, it is symmetric.
P206 has version (P206i is version of)	P205	E73 Information Object	E73 Information Object	Mapped from dcterms:hasVersion
P207 has format (P207i is format of)	P205	E73 Information Object	E73 Information Object	Mapped from dcterms: hasFormat
P208 replaces (P208i is replaced by)	P205	E73 Information Object	E73 Information Object	Mapped from dcterms:replaces
P209 requires (P209i is required by)	P205	E73 Information Object	E73 Information Object	Mapped from dcterms:requires
P210 conforms to (P210i is conformed to by)	P205	E73 Information Object	E73 Information Object	Indicates the norm which an information object conforms to. Mapped from dcterms:conformsTo. The norm is also an information object.
P211 is provided by (P211i provides)	-	E84 Information Carrier	E39 Actor	If an actor provides an information carrier it means that they are responsible for providing the resource but are not the current carrier of the resource (but, for example, an online metadata aggregator).

P14a carried out with contribution by (P14ia contributed to)	P14	E7 Activity	-	Represents contribution to object creation, mapped from dc:contributor
P67a has temporal coverage (P67ai is temporal coverage of)	P67	E73 Information Object	E53 Place	Mapped from dcterms:temporal (or dc:coverage).
P67b has spatial coverage (P67bi is spatial coverage of)	P67	E73 Information Object	E52 Time-Span	Mapped from dcterms:spatial (or dc:coverage).
P213 see also earlier form (P213i see also later form)	similarTo	E55g Subject Hierarchy	E55g Subject Hierarchy	Represents earlier form of KABA [16] record e.g. “Akademia Krakowska” is earlier form of “Uniwersytet Jagielloński”. It is not transitive! (See discussion in 4.2.5)
P214 see also broader term (P214i see also narrower term)	similarTo	E55g Subject Hierarchy	E55g Subject Hierarchy	Represents the “broader term” relation between KABA records. It is not transitive! (See discussion in 4.2.5)

3.2.4 Other Ontologies Assimilated in the Application Profile

Three external ontologies are present in the semantic knowledge base. One is the WGS84 Geo Positioning (World Geodetic System, *http://www.w3.org/2003/01/geo/wgs84_pos*) used to represent information about the geographic coordinates. The coordinates are obtained from the Geonames.org service, and they are used in the knowledge base to allow searching by geographic proximity. The CIDOC’s class E47 Spatial Coordinates could be used to hold the information, but this is against the WGS definition, which attributes coordinates to a place (location), and not the place’s appellation.

The second external vocabulary source is OpenVocab (http://open.vocab.org/), from which only one term was taken: the *similarTo* property, used the superproperty of four subject heading hierarchy relations (see 4.2.5).

Finally, the Electronic Thesis and Dissertation Metadata Standard [1] is used in combination with the E73b Periodical class and the E55a Degree type hierarchy (described in the previous section) to convey information about theses. The terms used are: *degree* (the academic degree that a thesis is required for), *education level* (as in the Bologna declaration, see E55a in 4.2.5), *research discipline* (E55b in 4.2.5), and *degree grantor* (the institution that confers the degree).

3.2.5 E55 Type External Hierarchies

According to the CIDOC specification, “E55 Type is the CRM’s interface to domain specific ontologies and thesauri. These can be represented in the CRM as subclasses of E55 Type, forming hierarchies of terms.”

The subclasses of E55 Type relevant in the digital library application have been listed in 4.2.1. This section explains the purpose of each type and the origin of the hierarchies.

The E55 Degree hierarchy represents the degrees associated with E73b Thesis instances. The E55b Education Level groups the corresponding degrees, while the E55c Research Discipline is the hierarchy of disciplines that degrees may be conferred in. Those types have been coded manually, the first two based on the Bologna declaration, the last one has been modeled on the official ministerial hierarchy of Polish research disciplines.

The E55d Resource Type is a hierarchy of resource types based on the DCMI type vocabulary, but extended with types traditionally used in digital libraries.

Classes E55e Subject, E55f User Subject, and E55g Subject Hierarchy are used to represent publications subjects: one of the most important feature used to group and search resources. Subjects have been divided into two categories. The E55g Subject Hierarchy has been built by transforming the KABA [16] subject headings into CIDOC-compatible format. The transformation process and the numerous problems associated with it (starting with the mere size of the hierarchy) have been described in [15]. Originally the dedicated CIDOC's P127 has broader term and P127i has narrower term relations were used to connect the subjects, but problems occurred due to the transitive nature of the relation. As KABA has been created manually by humans, it contains errors. Long subject headings hierarchies cannot be trusted: one can learn from KABA that "lighthouse is life", for instance. Problems of this type are also present in other subject heading hierarchies [17]. This is why the P213 and P214 intransitive relations have been introduced (see the bottom of the added properties table in 4.2.3). The E55f User Subject class groups subjects that were used in the digital libraries metadata, but cannot be directly mapped to KABA (though relations between KABA subjects and user subjects may be introduced based on the similarity of terms and the KABA grammar).

The E55j Subject Type hierarchy is used to further organize KABA subjects (which means that this is a "metatype", a type of type).

The E55h Place Type hierarchy is based on the Geonames.org feature classes and feature codes that define location types.

The next section summarizes the process of mapping flat metadata to CIDOC, discusses the advantages of such transformation, and also names the problems encountered in the process.

3.3 Metadata Translation to CIDOC

The detailed rules of translating PLMET, the Digital Libraries Federation metadata schema based on Dublin Core, have been discussed in [14]. The knowledge base creation process is comprised of the following stages:

- harvesting metadata by means of the OAI-PMH protocol,
- semantic cleaning of data: deleting irrelevant elements, copying or moving wrongly assigned data to the correct elements,
- normalization of dates, names, etc.,
- mapping the flat metadata schema to CIDOC,
- enriching data with information from external services,
- detection of relations among knowledge base entities,
- validation of the knowledge base and contraction removal.

3.3.1 Added Value

Using CIDOC CRM instead of a simple set of metadata elements opens up new search and resource discovery possibilities. In a regular digital library only publications (in the sense given in 4.2.2) could be returned as search results, and the search was based on pure text, possibly limited to the contents of a given set of elements.

In the semantic knowledge base, the user can search for any entity represented as an instance of a CIDOC class. The resources are tightly connected: one can move from a digital copy to other copies of the same publication, to other editions, to other works of the same author, to works about the author, and so on. As the data is checked and supplemented with external sources, typos are detected, different author appellations are correctly associated with the relevant author, and it is much easier for a user to be satisfied with the data.

Using and ontology, especially one that assimilates other hierarchies (the E55 Type class) helps organize the data. Often users refer to the same resource with different names: names in different languages (e.g. *Poznań* - *Posen*), synonymous names for types (*czasopismo* - *periodyk*), different inflection forms (g*azeta* - *gazety*). In the knowledge base such resources are properly merged and structured.

It is possible to ask complicated queries that would out of question without a proper ontology. For instance, the user can ask for “all books about insects written by Polish authors living in the 16th century”. The amount of time necessary to obtain the equivalent result in a library system without an ontology is incomparable.

Section 4.3 presents some facts about the knowledge base resulting from the translation of the LDF metadata to CIDOC.

3.3.2 Known Problems

The mapping from flat schemas to CIDOC may be complicated and the rules have to be determined manually. However, as the enrichment and relation detection steps in the Integrated Knowledge System have been separated and are performed after the mapping process, they can be applied to data coming from any schema.

It would be difficult to cataloguers and digital librarians to describe their data in CIDOC, as they are not used to this kind of conceptualization. A lot of work has been put into describing the more than 700,000 resources in the DLF and nobody will transform them to CIDOC manually, hence the automatic mapping step is inevitable. Still, the IKS is planned to have a special “wizard-type” interface by which users will be able to describe their resources unambiguously by means of CIDOC. Also, errors detected in the mapping and post-processing steps may and will be reported to the original hosting institutions, in hope that this kind of feedback will improve the quality of their data. A number of errors have been detected during the automatic processing of KABA (the E55g Subject Hierarchy) and a significant portion of them have already been corrected by the NUKAT centre.

It is worth noting that one of the biggest confusions during the mapping/knowledge base design process was the case of multiple appellations of the same entity (e.g. a place). CIDOC gives you two alternatives:

- setting a number of appellations to the same entity, identifying one of them as the main appellation,
- setting only one appellation to an entity, and connect others this appellations by means of the P139 has alternative form property.

At first the second solution seemed more user-friendly, as it easier to build SPARQL or SeRQL queries without having to consider the alternative appellations chain. Then, after analysis of the first solution, a risk has been identified: the P48 has preferred identifier does not have equivalents for places, people etc. This means that if a person build a query supposed to return a place using the E44 Place Appellation class and the P87 is identified by property for place appellations – the query would return all appellations (if any) except the preferred one, as the P48 is not a subclass of P87. Finally the second alternative was chosen as less likely to mislead users of the knowledge base who decide to write their own queries instead of using the provided IKS interface.

Some of the first users of the knowledge base (including those writing user interfaces) complain about the appellations layer, as they are used to systems in which the (String) name is attributed directly to the resource. However, as a user can have a number of appellations with an internal structure (name, surname) and its own metadata (language, years of validity), this is the correct solution.

A final concern is that the proposed extension is not *valid* with respect to properties. As stated above, all added classes are subclasses of those defined in the CIDOC specification, but some added properties are not subproperties of those from the CIDOC document. This issue will be addressed in future.

In the next section some interesting numbers are given concerning the knowledge base resulting from mapping the DLF data to the CIDOC application profile described above.

3.4 Knowledge Base Statistics

The numbers given in this section concern a knowledge base resulting from mapping 500,000 DLF publications. The knowledge base contains 148,253,680 RDF triples, including 23,040,700 explicit and 125,212,980 implicit (deduced) triples. Of the total number of 148,253,680 triples, 28,597 are ontology triples, and 6,031,068 represent KABA [15,16] hierarchy.

Subjects of the publications are of three main types. They are either KABA subjects (153,073 instances of the is about property, avg. 16.2 publications / used KABA subject), user subjects (235,155 property instances, avg. 4.3), or places (100,906 property instances, avg. 28.5).

The number of information objects is 2,082,076 of which 773,682 represent publications. There are 934,571 information carriers.

4. Conclusions

CIDOC has a steep learning curve for digital librarians, with a different conceptualization and a number of concepts nonexistent in the DL world. Thus, this paper proposes an application profile of CIDOC sufficient for DL use. It also names additional elements to be considered in the specification. The proposed classes are only specifications (subclasses) of existing ones, but a number of new properties had to be added to meet the set-up goals. CIDOC has been used to describe 700,000 DL publications (mapped automatically from a DL schema) and the results are promising.

The main conclusion is that CIDOC, even though created to represent museum collections, is expressive enough to describe information coming from digital libraries, with only minor extensions. There are more detailed ontologies, like FRBR, to describe library resources, but their level of complication requires manual verification of records, which would be inapplicable in the Integrated Knowledge System, aggregating large amounts of heterogeneous data. The author’s hope is that the proposed CIDOC application profile and extensions will be taken into consideration by the CIDOC community – especially that museums also often hold old prints.

Acknowledgements

The SYNAT project is funded by the Polish National Center for Research and Development (grant no SP/I/1/77065/10) and coordinated by ICM (University of Warsaw).

References

1. Atkins, A., Fox, E., France, R., and Suleman, H.: ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations, 1.2 edition. http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html (2008)
2. Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., Velkov, R. OWLIM: A family of scalable semantic repositories. In: Semantic Web – Interoperability, Usability, Applicability, http://www.semantic-web-journal.net (2010)
3. Clayphan, R. (Ed.): Europeana Semantic Elements Specification, Version 3.3.1, 24/01/2011. https://version1.europeana.eu/c/document_library/get_file?uuid=a830cb84-9e71-41d6-9ca3-cc36415d16f8&groupId=10602 (2011)
4. Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff, M.:. Definition of the CIDOC Conceptual Reference Model, 5.0.2 edition, June 2005. *http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf* (2005)
5. D'Arcus, B., and Giasson, F.: Bibliographic Ontology Specification, *http://bibliontology.com/specification* (2009)
6. Domowicz, I., Kalota, T., Koty ska, E., Łukaszewicz, J., Raczy ski, R.,, Szala, M., and Zgli ska-Adamska, D (Eds.). ePoradnik redaktora zasobów cyfrowych. Interpretacja schematu Dublin Core wraz z materiałami pomocniczymi dla redaktorów zasobów cyfrowych Biblioteki Cyfrowej Uniwersytetu Wrocławskiego. Biblioteka Uniwersytecka we Wrocławiu (2007)
7. Görz, G., Oischinger, M., Schiemann, B.: An Implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL. In: Proceedings of CIDOC 2008 — The Digital Curation of Cultural Heritage. ICOM CIDOC, Athens (2008)
8. Hohmann, G., Scholz, M.: Recommendation for the representation of the primitive value classes of the CRM as data types in RDF/OWL implementations. *http://erlangen-crm.org/docs/crm-values-as-owl-datatypes.pdf*
9. IFLA Study Group on the Functional Requirements for Bibliographic Records: Functional Requirements for Bibliographic Records, http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records (2009)
10. Kartsonakis, E., Kriara, L., and Papadakis, G.: BibTeX In OWL, *http://www.csd.uoc.gr/~hy566/bibtex/bibtex.pdf* (2008)
11. Kruk, S.R.: Semantic Digital Libraries - Improving Usability of Information Discovery with Semantic and Social Services (2010)
12. Lewandowska, A. Mazurek, C.,Werla, M.: Enrichment of European Digital Resources by Federating Regional Digital Libraries in Poland. In: 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14-19, 2008. Proceedings Series: LNCS, Vol. 5173, pp. 256--259 (2008)
13. Library of Congress Working Group on the Future of Bibliographic Control: Testing Resource Description and Access (RDA), *http://www.loc.gov/bibliographic-future/rda/* (2011)
14. Mazurek, C., Sielski, K., Stroi ski, M., Walkowska, J., Werla, M., W glarz, 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
15. Mazurek, C, Sielski, K, Walkowska, J., and Werla, M: KABA Subject Heading Language as the Main Resource Subject Organization Tool in a Semantic Knowledge Base, manuscript (2011)
16. NUKAT, the National Union Catalog, http://www.nukat.edu.pl/
17. Spero, S.: LCSH is to Thesaurus as Doorbell is to Mammal: Visualizing Structural Problems in the Library of Congress Subject Headings. In: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications (DCMI '08). Dublin Core Metadata Initiative, pp. 203--203. (2008)
18. Werla, M.: Metadane dokumentów w bibliotekach cyfrowych. XVII edycja seminarium w cyklu Digitalizacja - Problemy Cyfryzacji Dokumentów Pi mienniczych w Bibliotekach, Muzeach i Archiwach