Transforming a Flat Metadata Schema to a Semantic Web Ontology: The Polish Digital Libraries Federation and CIDOC CRM Case Study

Cezary Mazurek, Krzysztof Sielski, Maciej Stroiński,
Justyna Walkowska, Marcin Werla, and Jan Węglarz

Poznań Supercomputing and Networking Center, ul Z. Noskowskiego 12/14,
61-704 Poznań, Poland
{mazurek,sielski,stroins,ynka,mwerla,weglarz}@man.poznan.pl

Abstract. This paper describes the transformation of the metadata schema used by the Polish Digital Libraries Federation to the CIDOC CRM model implemented in OWL as Erlangen CRM. The need to transform the data from a flat schema to a full-fledged ontology arose during preliminary works in the Polish research project SYNAT. The Digital Libraries Federation site offers aggregated metadata of more than 600,000 publications that constitute the first portion of data introduced into the Integrated Knowledge System - one of the services developed in the SYNAT project. To be able to perform the desired functions, IKS needs heavily linked data that can be searched semantically. The issue is not only one of mapping one schema element to another, as the conceptualization of CIDOC is significantly different from that of the DLF schema. In the paper we identify a number of problems that are common to all such transformations and propose solutions. Finally, we present statistics concerning the mapping process and the resulting knowledge base.

Keywords: CIDOC CRM, digital libraries, metadata, ontologies, RDF repositories, semantic web, thesauri.

1 Introduction

SYNAT is a national research project aimed at the creation of universal open repository platform for hosting and communication of networked resources of knowledge for science, education, and open society of knowledge. It 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).

Poznań Supercomputing and Networking Center (PSNC) is one of the key SYNAT partners. One of the PSNC 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

R. Bembenik et al. (Eds.): Intelligent Tools for Building a Scient. Info. Plat., SCI 390, pp. 153–177.
springerlink.com

necessary, to which the existing heterogeneous representations and schemas can be converted.

As the system is dedicated mostly to researchers in the field of humanities, and the first aggregated resources will be the Digital Libraries Federation data, the CIDOC Conceptual Reference Model [4] (implemented in OWL as Erlangen CRM / OWL [11]) has been chosen as the main description ontology. The obtained knowledge representation is stored in an RDF repository, hereinafter referred to as the *knowledge base*.

The remaining part of this paper describes the conversion process and highlights the most important research and technical issues and results obtained so far. We summarize the mission of the Polish Digital Libraries Federation and describe the metadata it aggregates. We discuss the applicability of the CIDOC CRM ontology and justify its choice as the main ontology describing the contents of the knowledge base. We also dedicate some space to describe the auxiliary ontologies and vocabularies that support CIDOC in the knowledge base. Later on we move to discussing the biggest challenges in the process of mapping a flat metadata schema to an event-centric ontology in order to create the knowledge base, such as a very different conceptualization, the need for nesting external type hierarchies (sometimes large, complicated, and even internally inconsistent) and for data enrichment. We propose a modular Knowledge Retrieval System architecture and describe the mapping process. In the last part of this paper we present the results and statistics, and also name some remaining problems and future tasks.

2 Digital Libraries Federation

The PIONIER Network Digital Libraries Federation (DLF, [15]) is the next stage of the development of an infrastructure of distributed digital libraries and repositories in Poland [17]. The name of the DLF corresponds to its nature - it 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 mission of the DLF is to:

1. facilitate the use of resources from Polish digital libraries,
2. increase the visibility of Polish resources in the Internet,
3. provide advanced network services based on the digital libraries' content to Internet users and digital libraries creators.

Thanks to the aggregator features of DLF, digital libraries users have been given the possibility to search over 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.

The DLF harvests data through the OAI-PMH protocol. The updates are performed nightly. The DLF publishes the harvested metadata further through the OAI-PMH protocol. The added functionalities include e.g. duplicate and similar resources

detection, system of persistent identifiers and a Shibboleth-based mechanism of networked readers' profile[8].

As of April 2011, the number of publication in DLF exceeds 600,000. 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.

Polish digital museums are also interested in joining the Federation. The first one which actually joined (in April 2011) was the National Museum in Warsaw.

The DLF's metadata schema is described in section 3. More information about the contents is given in section 7.

3 DLF Metadata

The Digital Libraries Federation metadata can be obtained via the OAI-PMH protocol. The basic metadata schema used by the Federation in the Dublin Core Metadata Element Set, but the DLF is now switching to a new, richer metadata schema called PLMET. This schema has been created in accordance with the demands and suggestions of creators digital libraries cooperating with the Federation. The mapping described here and performed by the prototype is prepared for the new schema, but also works on the current one (a subset of PLMET elements).

PLMET is comprised of a number of tags from different established namespaces and a few proprietary tags. The used namespaces are:

- Dublin Core Metadata Element Set, http://purl.org/dc/elements/1.1/, *dc*
- Dublin Core Metadata Terms, http://purl.org/dc/terms/, *dcterms*
- Electronic Thesis and Dissertation Metadata Standard, http://www.ndltd.org/standards/metadata/etdms/1.0/, *etdms*

The proprietary namespace is referred to as *plmet*.

Additionally, the DLF exposes data in the Europeana Semantic Elements schema (http://www.europeana.eu/schemas/ese/) which will be referred to as *ese* [3]. Besides a selection of Dublin Core Metadata Terms elements, *ese* includes also several proprietary elements whose values can be generated automatically, mostly on the basis of the *dc* elements. Those proprietary elements do not form part of the PLMET schema, but they are used in the mapping process as additional information.

The schema contains all *dc* elements, that is: *title, creator, contributor, subject, coverage, description, publisher, date, type, format, source, language, relation, rights*.

It also contains the following *dcterms* elements (so called qualifiers of elements from *dc*): *alternative* (title), *spatial, temporal* (coverage), *abstract, tableOfContents* (description), *available, created, dateAccepted, dateCopyrighted, dateSubmitted, issued, modified, valid* (dates), *extent, medium* (formats), *hasPart, isPartOf, hasVersion, isVersionOf, hasFormat, isFormatOf, references, isReferencedBy, replaces, isReplacedBy, requires, isRequiredBy, conformsTo* (relations), *bibliographicCitation* (identifier), *accessRights, license* (rights), *rightsHolder, provenance.*

The used elements from *etdms* [1] are: *degree, degree.name, degree.level, degree.discipline, degree.grantor* (*degree* is a part of the description)

The proprietary *plmet* elements are userTag (subject), *placeOfPublishing* (description), callNumber (identifier), *locationOfPhysicalObject* (provenance), *digitisation* and *digitisationSponsor*.

The *ese* elements that are not part of PLMET but are offered by the DLF and used in the mapping are *unstored, object, provider, type, rights, dataProvider, isShownBy, isShownAt*.

The semantics of the elements are discussed in more detail section 7.

4 CIDOC CRM as the Knowledge Base Target Ontology

To achieve the goals connected with Semantic Web, Linked Data and intelligent search and resource discovery, an ontology had to be chosen that would suffice to describe all types of data (including the DLF data described above) that might appear in the Integrated Knowledge System, dedicated to researchers in the field of humanities. There have been proposals to create a proprietary ontology, but finally CIDOC Conceptual Reference Model (CRM, [4]) was chosen due to the following reasons:

1. It is a widely recognized standard, so data from the IKS knowledge base will be understandable for wider audiences.
2. Substantial effort has been dedicated to the creation of CIDOC CRM, so the ontology is rather mature and stable.
3. CIDOC is an ontology for describing cultural heritage, and this kind of information will prevail in a humanity-research portal [20].
4. CIDOC is very expressive and is sometimes used as an intermediate representation while mapping between different metadata schemas [16],[10].

Attempts to map the Dublin Core Metadata Element Set (which is a subset of the DLF's PLMET schemta) to CIDOC CRM for museum collections have been described e.g. in [14], [7].

This combination of features allows us to use CIDOC as the main ontology for resource description in the knowledge base. This is a more convenient solution than trying to integrate different, possibly disjoint ontologies for different types of resources.

Still, some extensions (mostly subclasses and subproperties) of CIDOC have been proposed. Additional types of information are described with a number of other standard, smaller ontologies and vocabularies, described shortly in 4.3.

4.1 CIDOC CRM

CIDOC Conceptual Reference Model [4] is an ontology for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. The

ontology defines 86 classes and 137 unique properties. It is event-centric: events such as creation or acquisition are crucial to the domain description.

CIDOC CRM is mostly used to describe museum collections, but as publications (both physical and digital) are also artifacts of cultural heritage, they can be described with the (slightly extended) ontology. Another reason of choosing CIDOC in the SYNAT project, and not a purely bibliographic ontology like http://bibliontology.com/, was the fact that both the data from digital libraries (or catalogues) and from digital museums have to be integrated in SYNAT, together with many other forms of resources.

By convention, class names in CIDOC CRM start with E (e.g. E21_Person), property names start with P (e.g. P25_moved). Inverse properties, if present, are marked with i (e.g. P25i_moved_by). If in any of the examples in this paper another letter appears after the number (e.g. E55b_Education_Level), it means that the class or property is a subclass or subproperty of a CIDOC class introduced for the purposes of the described knowledge base (in this case: E55_Type).

4.2 Erlangen CRM / OWL

The Erlangen CRM / OWL [11] is an OWL-DL 1.0 implementation of the CIDOC Conceptual Reference Model. A OWL2 (RL) version is planned soon.

It has been originally created at the Friedrich-Alexander-University of Erlangen-Nuremberg in cooperation with the Department of Museum Informatics at the Germanisches Nationalmuseum Nuremberg and the Department of Biodiversity Informatics at the Zoologisches Forschungsmuseum Alexander Koenig Bonn. It is currently maintained by Martin Scholz, Georg Hohmann and Mark Fichtner.

The Erlangen CRM is an interpretation of the CIDOC CRM in a logical framework attempting to be as close as possible to the text of the CIDOC specification. However, there are some CIDOC classes that are not present in the Erlangen implementation. These are mostly classes representing primitive types (e.g. E60_Number, E62_String). They are supposed to be replaced in OWL with typed (XSD) literals [12].

4.3 Other Ontologies and Vocabularies

Other ontologies and vocabularies used in the knowledge base are:

- the Geonames ontology (http://www.geonames.org/ontology),
- ISO 639 language codes (http://downlode.org/rdf/iso-639/schema#),
- Dublin Core Metadata Initiative Type Vocabulary (http://dublincore.org/2010/10/11/dctype.rdf#),
- WGS84 Geo Positioning (World Geodetic System, http://www.w3.org/2003/01/geo/wgs84_pos).
- OpenVocab (http://open.vocab.org/)

The Geonames ontology is used to represent place types (e.g. city, lake). Language codes are used to standardize language descriptions. The DCMI Type Vocabulary is a basis to describe types of works (books, images, audio files). WGS84 is used to hold information about geographic coordinates, which can be used to disambiguate locations (e.g. if there is a number of places with the same name it can be checked which one is closest to another place mentioned in the description) and also to perform advanced queries, as the RDF repository we use offers Geo-spatial Extensions. The method of connecting Erlangen CRM / OWL with WGS84 is presented in Fig. 1.

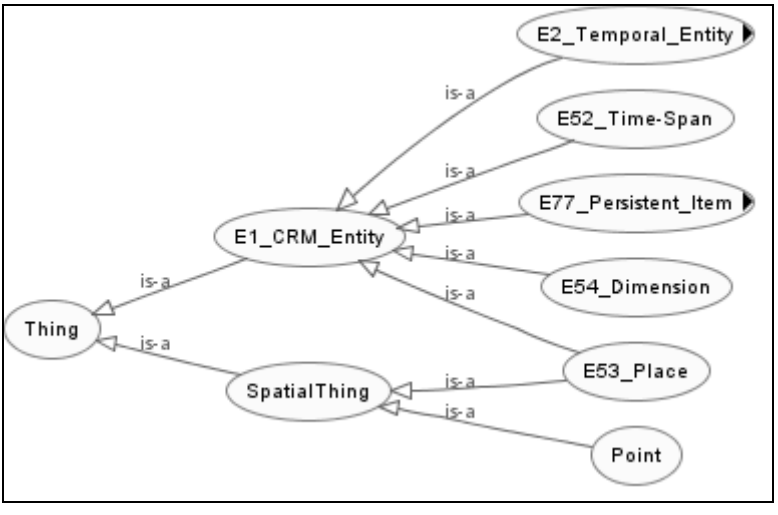

Fig. 1. Combining Erlangen CRM / OWL with WGS84

5 Knowledge Base Construction Steps

At the current stage of the prototype development, the CIDOC-conformant knowledge base is constructed from the DLF metadata (600,000 records mapped by the DLF to the PLMET schema). Other sources of data will be added in future. Knowledge is stored in an RDF repository and can be queried using RDF query languages such as SPARQL.

Fig. 2. illustrates the components participating in the process of knowledge base construction. The functional components are marked as circles and ellipses, the rectangles are data sources and databases. The dashed components are planned, but not yet implemented or connected as part of the prototype. The processing steps necessary to transform the Digital Libraries Federation's metadata into CIDOC CRM and create a truly semantic knowledge base are described below.

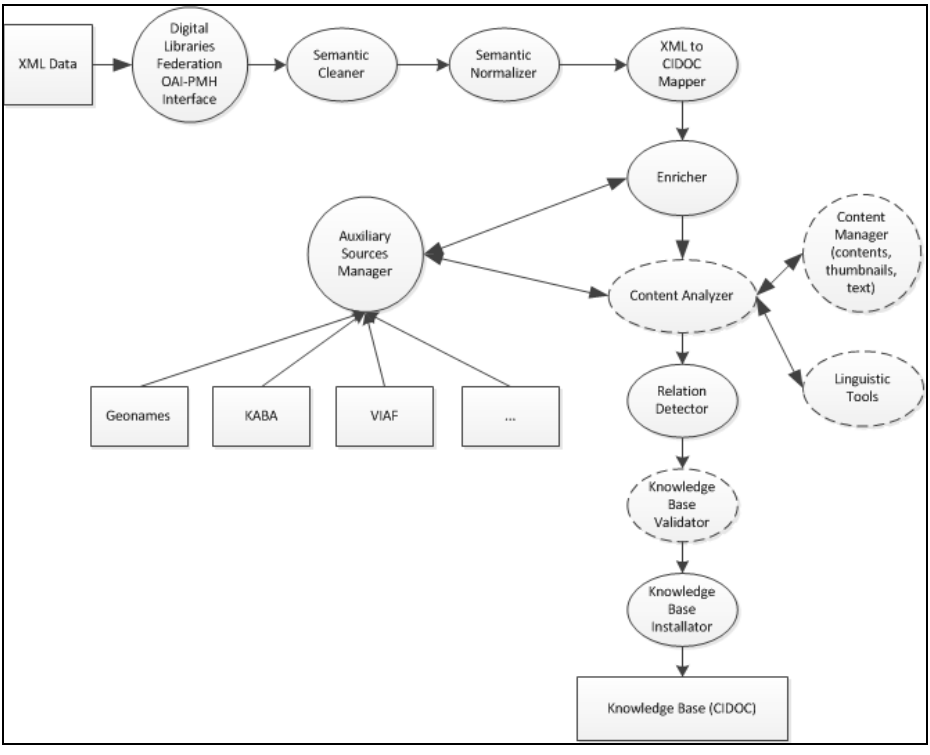

Fig. 2. Components participating in the knowledge base construction process

The processing steps are as follows:

1. The incoming metadata,[1] still in the flat XML source format, is semantically cleaned. Cleaning is done on the record level. As a result some of the record's elements may be deleted (if they contain irrelevant information, e.g. "no date" entry in the "publishing date" element), moved (if they have been assigned incorrectly by cataloguers), or copied to another element, if their meaning is wider.
2. The next step is semantic normalization. Normalization is done on the level of one element. Text elements representing entities such as dates or people are transformed into structures with established format (e.g. all dates are converted to structure representing range YYYY-MM-DD – YYYY-MM-DD).
3. The cleaned and normalized metadata are mapped to a so called *intermediate* CIDOC representation. The data is in CIDOC, but it has not yet been enriched or internally related. Because advanced reasoning has been removed from this stage, the mapping can be performed by one of the existing mapping tools, such as AnnoCultor (http://annocultor.eu/).

[1] Depending on the current state of the system and the knowledge base, the input constitutes of either the complete set of metadata from all sources, or just those records that have recently been added or updated.

4. The next stage is data enrichment. External sources, such as Geonames or VIAF, are called to see if they possess information about the entities referred to in the metadata. If so, two operations are performed:
   a) the entity is linked to the external resource by a URI,
   b) useful additional information (e.g. alternative names, geographic coordinates) is copied to the knowledge base.
5. The operations in step 4. are performed on information contained in the metadata. Similar operations should be performed on the contents of the resource (and also on longer natural language texts in some metadata elements) based on a number of linguistic tools. This step has not yet been implemented.
6. During the relation detection step, related resources (e.g. different editions of the same book) are connected to each other with CIDOC relations. This step may reduce the number of triples in the RDF repository, because some resources are discovered to represent the same entity. Because of this, up to this point the reasoning engine is turned off, as when it is turned on it is very costly to remove triples (the knowledge base contents have to be recalculated).
7. After the relations are detected, the reasoner is turned on and any possible contradictions are detected. Finding and fixing the reason for a contradiction is an advanced step that has not been implemented in the prototype.
8. Finally, the knowledge base is installed. In production environment it will be deployed on a cluster of nodes. The installer will be responsible for ensuring that during the installation process users of the system have access to at least one functional node.

6 Main Challenges in the Mapping Process

This section names some of the most difficult aspects of the transformation from the DLF metadata schema to the chosen ontology.

6.1 Conceptual Separation of Information Object and Information Carrier

One of the first challenges that need to be faced when transforming a flat digital libraries metadata schema to CIDOC is the conceptual separation between two classes: E73_Information_Object and E84_Information_Carrier. To quote [4]:

[The E73_Information_Object*] class comprises identifiable immaterial items, such as a poems, jokes, data sets, images, texts (...) that have an objectively recognizable structure and are documented as single units. An* E73_Information_Object *does not depend on a specific physical carrier.*

The E84_Information_Carrier, on the other hand:

(...) comprises all instances (...) that are explicitly designed to act as persistent physical carriers for instances of E73_Information_Object.

At first it might seem difficult to decide which metadata elements relate to the E73 and which to the E84. However, careful analysis inevitably reveals that this kind of

separation is highly desired in the digital libraries world. Very often in the catalogs one can find information that a resource is *in the DjVu format*, *measures "15x10 cm"*, and *was written by Jan Kowalski*. Obviously, the DjVu file has no spatial dimension, the original book is printed on paper and has nothing to do with the DjVu format, and the author created a composition that is independent from the carrier.

To conclude, what we have here is one E73_Information_Object (the work itself) and two E84_Information_Carriers (physical and digital). A straightforward consequence is that one E73_Information_Object can have a number of E84_Information_Carriers.

Fig. 3. presents the ideal outcome of transforming a number of metadata records describing related copies from digital libraries and catalogues.

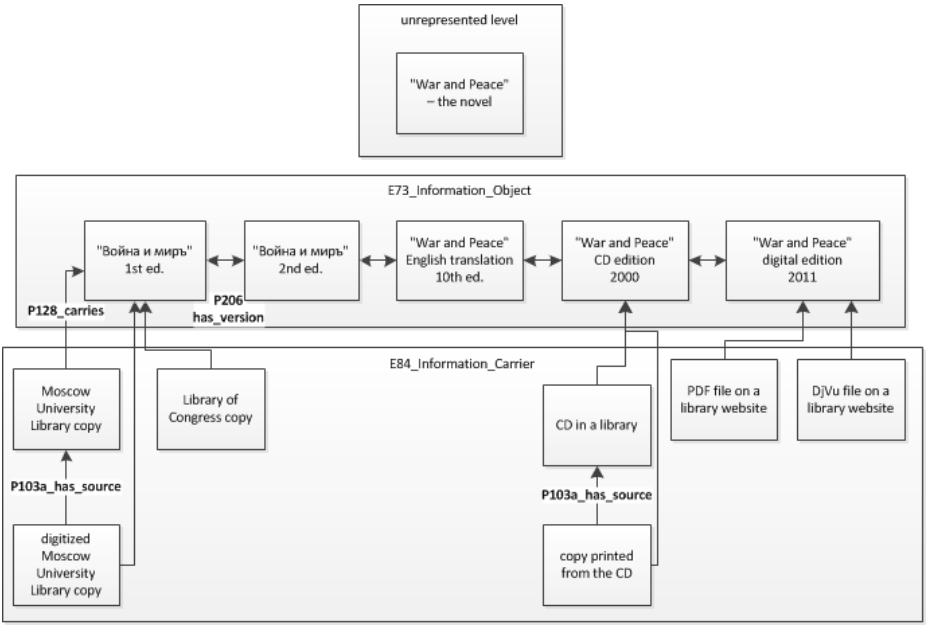

Fig. 3. Representation of different editions of the same work and different copies of the same editions in the knowledge base, based on CIDOC CRM

6.2 E55 Type Hierarchies

CIDOC's E55_Type class is an interface to domain specific ontologies and thesauri. External ontologies and classifications are represented as subtypes of E55 (built-in subclasses include E56_Language, E57_Material and E58_Measurement_Unit). The instances of each subclass represent concepts (and not individuals!), and are connected by means of the P127_has_broader_term property.

The following external hierarchies (presented also in Fig. 4) are used in the described translation:

- E55a_Degree (to describe the degree associated with a thesis),
- E55b_Education_Level (as in the EU's Bologna declaration),
- E55c_Research_Discipline (the official hierarchy of Polish research disciplines),
- E55d_Resource_Type (resource types based on the DCMI type vocabulary),
- E55e_Subject (divided into E55_Subject_Hierarchy and E55f_User_Subject),
- E55g_Subject_Hierarchy (the resource subject hierarchy, based on KABA, a Polish thesaurus of subject headings, based on Library of Congress Subject Headings),
- E55f_User_Subject (subjects that are not from KABA and are not recognized as people, places or other entities)
- E55h_Place_Type (based on the Geonames classes).

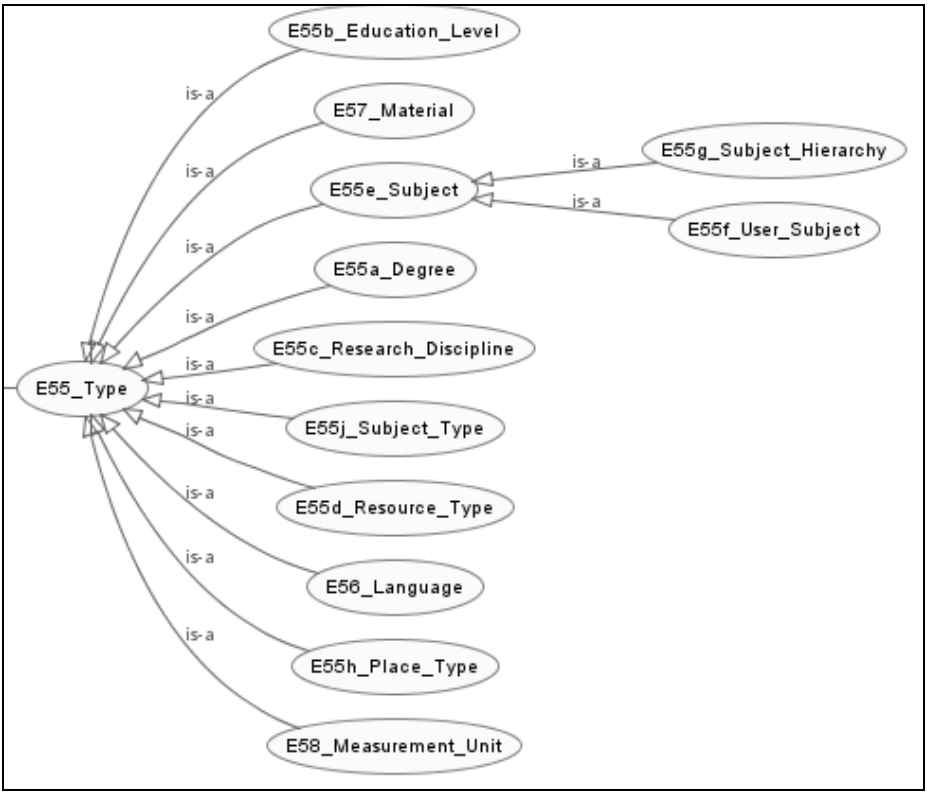

Fig. 4. The E55_Type hierarchies in the knowledge base

6.3 Identities and URIs

One of the goals of translating a flat schema to an ontology is to detect and show resource relations. One of the difficulties is that the related resources are often described in natural language. The text from the elements of a metadata record (the author, related publications, publication place etc.) itself represents knowledge base entities. These entities have to be recognized, created, enriched with external information, and merged with existing entities if they are already present in the knowledge base (e.g. other books of the same author are present).

6.4 Data Enrichment

To allow for more efficient searching and to offer users more valuable information, and also to facilitate disambiguation, the information obtained from the metadata records is enriched with data from auxiliary sources. The functional components of the Knowledge Retrieval System are shown in 5.

The auxiliary sources are heterogeneous and have to be accessed and processed in dedicated ways. Listed below are the auxiliary sources used in the prototype implementation together with their descriptions. With many of the sources the main difficulty is that there is more than one result with the given name, so some reasoning is necessary to choose the most probable option.

Geonames is a geographical database. It contains over 10 million geographical names and consists of 7.5 million unique features of 2.8 million populated places and 5.5 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes.

It can be accessed as a web service, but it is also possible to download the data and use it locally. The place descriptions contain information about name, alternate names, country, place type, population, elevation, time zone and geographic coordinates.

TERYT is the National Official Register of Territorial Division of Poland. It contains identifiers and names of units of territorial division, identifiers and names of localities, statistical regions and census enumeration areas, identification of addresses of streets, real estates, buildings and dwellings.

Most towns present in TERYT are also described in Geonames. However, Geonames do not offer information about territorial division, so it is useful to combine data from these sources. However, it is not a straightforward task to provide a one-to-one mapping between places in TERYT and in Geonames.

VIAF (Virtual International Authority File) is an international authority file. It is a joint project of several national libraries and operated by the Online Computer Library Center (OCLC). It gathers data from a number of national institutions and offers (via a web service) information about people and institutions.

NUKAT is one of the institutions participating in VIAF. It is the union catalogue of Polish scientific and academic libraries. It also maintains an authority file (also accessible through VIAF) and the KABA **subject headings system** described below. As NUKAT is a partner in the SYNAT project, we are able to access their resources directly.

6.5 KABA

KABA is used differently than other auxiliary sources in the way that it is completely incorporated into the knowledge base to allow reasoning.

In order to describe publications (E73_Information_Objects) with precise subjects we decided to incorporate the KABA Subject Headings into our knowledge base. KABA is a Polish subject indexing system built and maintained by NUKAT. Subject headings systems are used for cataloging of library collections by subjects. KABA is based on three such systems [18]: Library of Congress Subject Headings (LCSH), National Library of France Subject Headings (RAMEAU) and University of Laval in Quebec Subject Headings (RVM).

KABA consists of almost 900,000 records represented in the MARC21 authority format [19]. The MARC21 elements most significant from the Integrated Knowledge System's point of view include the heading (the 1XX fields in MARC21 format), alternative versions of the heading (4XX) and its relations with other records (5XX). The relations include: broader term, narrower term (these two can be regarded as hierarchical relations), earlier form of heading and later form of heading. Eleven types of records can be distinguished, such as: geographical name, person name, or corporate name. Some potentially significant information in record definitions had to be ignored as it is formulated in natural language. An example of it is field 360 with natural language explanation of complex relationships between records.

The idea was to convert KABA Subject Headings into a thesaurus compatible with CIDOC. A new subclass of E55_Type (see 6.2) has been introduced to represent this structure: E55g_SubjectHierarchy. Hierarchical relations were mapped as the P127_has_broader_term and P127i_has_narrower_term CIDOC relations. Other relations were mapped using the OpenVocab's (http://open.vocab.org/) similarTo symmetric property.

The first obtained version of the thesaurus lacked many hierarchical relations which are not explicitly defined in records, but can be inferred from the grammar of subject headings. Rules for inferring such relations have been proposed in [5]. We implemented all four rules described in the paper. The major rule is based on the fact that a subject heading may contain determiners (*subdivisions*) that bring more detail to the topic. A broader term can be built from a subject heading containing such subdivisions by removing some of them. For instance, the broader term *Malarstwo* (*Painting*) can be inferred from *Malarstwo -- techniki* (*Painting -- techniques*). This rule has produced over 200,000 new relations in the knowledge base (KABA explicitly defines about 230,000 relations). The reasoning engine also induced a number of relations (implicit RDF triples) calculating the closure of symmetric and inverse relations.

Still, there are some concerns about the quality of the resulting thesaurus. KABA has been created manually and as such contains errors in record definitions: typographical errors, duplicates and specification incompatibilities. A similar experience of converting LCSH into a thesaurus revealed mistakes in hierarchy relation definitions which lead to apparently wrong conclusions, such as "*a doorbell is a mammal*" [22]. Such mistakes can result in structural inconsistencies in the hierarchy of subjects (i.e. cycles). The resulting thesaurus has been examined with the

Tarjan algorithm[2] to find all cycles. An example of such cycle is *Inżynieria Ropy Naftowej* (*Petroleum engineering*) and *Ropa naftowa -- produkcja i handel* (*Petroleum --industry and trade*) declared as broader terms of each other. 270 such cycles were discovered. The majority consist of two elements. In some cases a record is defined to be a broader term of itself.

The actual number of wrongly defined relations is unknown. Cycles form a special case that can be detected automatically, but the majority of errors and logical inconsistencies can be found only by human verification.

7 DLF Metadata Schema Mapping Description

This section describes the most important points and issues of the transformation, also addressing the non-standard way in which digital library editors use the metadata elements.

One of the most crucial decisions that have to be made while mapping a DLF metadata record is deciding which elements describe the E73_Information_Object and which the E84_Information_Carrier, and whether there is one E84 or two (i.e. the physical copy of the book and the digitized version).

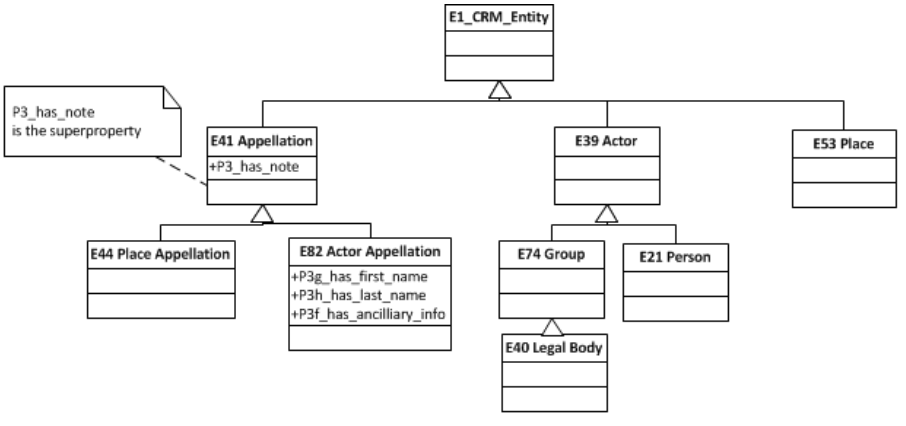

Fig. 5. The hierarchy of entities and their corresponding appellations

It is also important to realize that CIDOC treats names (*appellations*) in a special way. A person does not have a name and surname: a person has an appellation, or even a number of appellations of which one may be declared as preferred. Details and examples are illustrated in Fig. 5 and Fig 6. This is significantly different than what can be found in flat metadata schemas, but is fully justified. People, places and works of art may have different names in different languages, domains or historical periods.

[2] The Tarjan algorithm finds strongly connected components (such components contain at least one cycle) [23].

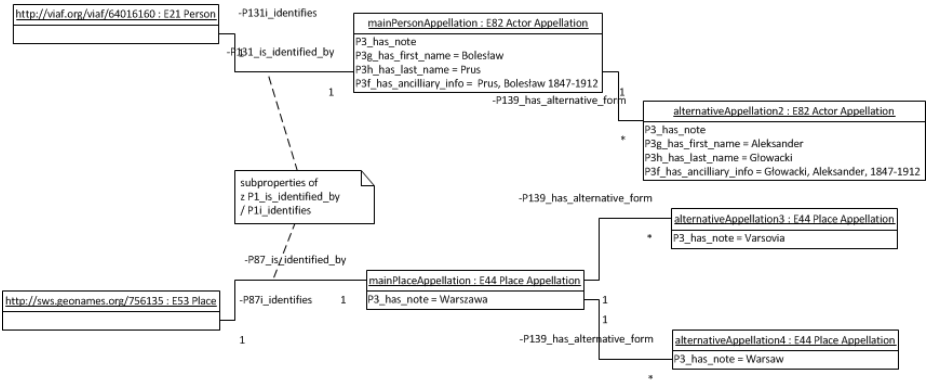

Fig. 6. Example of entities and their appellations

The names themselves (the appellations) may have a complicated structure, and grouping those structures in different appellation instances facilitates understanding and reasoning.

7.1 Titles

The *dc:title* is the title of the work, i.e. the E73_Information_Object. It can be easily mapped to CIDOC's E35_Title (both an E33_Linguistic_Object and E41_Appellation[3]) associated with the E73 by means of the P102_has_title property. P139_has_alternative_form allows for connecting the *dcterms:alternative* title to the main title.

7.2 Creators

Both *dc:creator* and *dc:contributor* are instances of E39_Actor. Creator performs the E65_Creation event of the E73_Information_Object. An E73 comes into existence through an E65_Creation, while an E84_Information_Carrier through an E12_Production.

The contents of the creator and contributor elements often consist of more than just the name string, for instance:

Kowalski, Jan (1950- ; historyk) Red.

Often, the name is followed by parentheses in which dates and other additional (e.g. domain of activity, especially when there are more people with the same name) information is given. Information about role in the creation process is usually given after the parentheses. This information is used in the mapping process to connect the actor instances not only with their respective appellations, but also with the E67_Birth and E69_Death events. The creator is connected to the creation event with the P14i_performed property. For the contributor, different roles (editor, advisor, ...) are proposed, represented by subproperties of P14i.

[3] The OWL specification allows multiple inheritance.

7.3 Subjects and Coverage

The *dc:subject* is one of the trickiest elements, as a resource can be *about* anything (the CIDOC specification agrees, defining the range of the P129_is_about property as the most general E1_CRM_Entity). Fortunately, the majority of Polish librarians are accustomed to using the KABA subject headings. However, it is possible that, for instance, a resource is about another one. A possible option is to use the WordNet ([9],[6]) thesaurus to represent all possible concepts.

The current implementation tries to map the subjects to KABA (see 6.4) or resources of types found in VIAF (people, institutions) or Geonames (places). If it fails, a new user subject (see 6.2) is created.

One of the challenges addressed in the SYNAT project is automatic subject detection based on the contents of the resource.

The *plmet:userTag* is represented with a subproperty of P129 with the assumption that *tags* are even more informal than user-created subjects. The tags in this metadata element come from the readers, i.e. the users of a digital library, and not from librarians or cataloguers (as is the case for the subjects forming the E55e_Subject hierarchies, see 6.2).

The *dc:coverage* and its subelements *dcterms:spatial* and *dcterms:temporal* are also connected to what the work is about, only they cover more specified types (time and place). It is important to remember that the coverage describes the subject of the resource and not aspects of its availability (e.g. when/where copies of particular book were distributed). Therefore in the mapping process coverage is related to the E73_Information_Object instance by means of the CIDOC P67_refers_to property. In case of spatial coverage the range (i.e. the object of the triple representing the relation) is an E53_Place, in case of temporal property it is E52_Time-Span.

In case of old publications the place instance may represent a historical place, significantly different from the modern one, although having the same name (e.g. compare borders of Poland before and after the Second World War).

7.4 Descriptions

The *dc:description*, *dcterms:abstract* and *dcterms:tableOfContents*, all related to the E73, constitute of plain text and at this point it would be difficult to propose a better mapping than just copying the contents to subproperties of CIDOC's datatype property[4] P3_has_note. In some cases a more meaningful table of contents might be built based on different kinds of relations (*dc:relation* and its subelements, especially *dcterms:hasPart*). Future works allowing to achieve more meaningful mapping should include deep natural language processing.

7.5 Publishing

The *plmet:placeOfPublishing* element names the E53_Place in which the resource (this time this is the physical copy, i.e. the E84_Information_Carrier) was published.

[4] A datatype property in OWL is a property whose range (object) is a literal and not a resource identified with a URI.

The publishing is an event represented by a subclass of E12_Production (as stated before, in CIDOC the carriers are produced and the information objects are created).

Another of the DLF's proprietary elements, *plmet:digitisationSponsor*, names the institution that paid for the digitization process (a subclass of E12_Production). The product of the digitization is a new E84_Information_Carrier, based on another E84 instance representing the physical resource. The publications whose original form is digital are not subject to digitization. In their case the only E84 instance represents the digital copy.

The *dc:publisher* is associated with the same publishing event as *plmet:placeOfPublishing*. The E39_Actor (usually an institution, i.e. E40_Legal_Body) is connected to the event by means of the P14i_performed property.

7.6 ETD-MS Elements

The translation of the ETDMS thesis elements is quite straightforward. It is based on the E55_Type hierarchies described in 6.2. The degree grantor is an institution, represented in CIDOC as an instance of E40_Legal_Body.

7.7 Dates

It is advised in the PLMET specification that the *dc:date* element be left empty, and the subelements be filled instead. Otherwise it is difficult to determine what the date represents. As CIDOC is an event-based ontology, the respective dates are not mapped to static values, but are instead translated to events (some of which are built-in CIDOC events, like E65_Creation, and some are new subclasses of E7_Activity). If *dc:date* is the only given date, depending on the type of the resource it is assumed to be either a creation (e.g. manuscripts) or publishing (e.g. old prints) date.

7.8 Types

The mapping of the *dc:type* to CIDOC and its consequences are described in [13].

Most resources in Polish digital libraries are periodicals and books. Periodicals have to be treated in a distinct manner. Typical library catalogues contain one record per periodical title, while in digital libraries each issue of the title is described. Often a periodical issue described in digital library metadata has two titles. One of them is the title of the periodical as a whole (e.g. *Kurier Warszawski*), the other contains the particular issue information (e.g. *Kurjer Warszawski / [red. L. A. Dmuszewski]. 1827, nr 121*). In the proposed CIDOC representation a new subtype of E73 has been created to represent a periodical title, to which all issues are related (with P148_has_component[5]). Sometimes advanced parsing of the whole record has to

[5] For some time we considered an alternative solution in which the periodical was not an instance of E73_Information_Object, but a type of E78_Collection. This solution proved to be inapplicable because the E78_Collection as specified in CIDOC aggregates physical resources.

be used to detect the part representing the title of the whole periodical, as librarians not always adhere to standards or create their personal standards.

7.9 Formats

The *dc:format* and its subelements (*dcterms:extent, dcterms:medium*) describe the dimensions and medium of the resource. It is important to realize that those are physical features (P45_consists of and P43_has_dimension) of the E84_Information Carrier, and not of E73_Information Object (not to be confused with *dc:coverage*).

A unit should be associated with a dimension (P91_has_unit). In case of book copies there might be centimeters or the number of pages, in case of recordings these might be minutes.

7.10 Identifiers

The *dc:identifier* element should be a URI which can also be the URI of the relevant individual in the RDF repository. Often the resource has a number of URIs, e.g. different ones in its owner's and an aggregator's databases. The *plmet:callNumber* and *dcterms:bibliographicCitation* are also ids. One id should be set as preferred one – we propose the URI from the data provider. The *bibliographicCitation* is a subject of controversy. For one, large parts of it could potentially be generated automatically from the resource metadata (e.g. a paper from a conference volume). Another problem is that there is a number of different citation formats, although PLMET recommends to use PN-ISO 690 standard here.

7.11 Language

Representing the *dc:language* is straightforward, with one remark: the domain[6] of the P72_has_language property is E33_Linguistic_Object, and neither the information object nor carrier are linguistic objects. The E33_Linguistic_Object has to be connected with the E73 by means of the P148_has_component property before the language property is set.

It is worth noting that E35_Title also is a subclass of E33. If the digital library metadata contains information about title language (for instance expressed as xml:lang) the information can be introduced to the knowledge base.

Languages in the knowledge base are described using one of the E55_Type hierarchies (see 6.2.).

7.12 Relations between Resources

The *dc:source*, *dc:relation* and its subelements represent relations between resources. New properties (and their inverse counterparts) had to be added to the ontology to represent some relations, some (e.g. P67_refers_to) were already present.

---

[6] In OWL domain constrains the types that can be the subject of a triple in which the given property is the predicate.

The *dcterms:conformsTo* relation turned out to be controversial. It seems to be of a different kind than other relations - it relates a resource to a standard or norm, and not to another resource. This problem was solved with the decision to regard the standard as a new E73_Information_Object. At some point the actual text of the standard might be introduced into the repository as a regular resource.

During the mapping and knowledge retrieval process described in 5, a module called the Relation Detector is responsible for connecting related resources. The relations are most often described in plain text. The Mapper, if it is to be kept simple and possibly based on existing tools, should not be responsible for this task.

7.13 Rights

CIDOC is prepared for modeling rights information (license type, rights owner) with the E30_Rights class and the properties P104_is_subject_to and P105_rights_held_by. Additional information about rights may be extracted from Europeana elements, provided they are available (see 7.15).

7.14 Provenance

The *dcterms:provenance* element contains *a statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation*. There are two proprietary DLF subelements *plmet:locationOfPhysicalObject* and *plmet:digitization*. The digitization describes the person or organization who performed the digitization (which, as stated earlier, is a subclass of E12_Production). The location of physical object is either a place, in which case the P55_has_current_location property is used, or an institution, in which case we use P50_has_current_keeper.

If the provenance element contains a natural language description, at this point it is copied as-is.

7.15 Europeana Elements

The Digital Libraries Federations is one of the providers of data to Europeana [21], an aggregator portal that groups and displays information about digital resources of European museums, libraries, archives and audio-visual collections. Europeana is in the process of switching to a linked data schema, but for now the official data exchange schema is ESE, which is an extension (a profile) of the Dublin Core Metadata Terms schema.

The ESE proprietary tags are designed for automated processing and therefore quite strictly specified, so obtaining their contents from the DLF adds value to the proposed mapping. The kind of information that can be extracted from these tags and introduced into the knowledge base is summarized below. It is important to note that the Europeana tags concern mostly the E84_Information_Carrier:

- *ese:object* – this is the URI of a thumbnail of the resource. In the knowledge base the thumbnail is a subclass of the E84 class.
- *ese:provider* – names the source from which Europeana obtained data (this might be an aggregator service).

- *ese:provider* – this is the original source, i.e. the keeper of data, mapped with the P50_has_current_keeper property.
- *ese:type* – the type of the resource, as in E55d_Resource_Type (Europeana's list of allowed types is more limited). This, for consistency reasons, it mapped to the E73_Information_Object.
- ese:rights – valuable right information: the URI to a document describing the license. Mapped with P104_is_subject_to.
- *ese:isShownBy* – URI of a high quality presentation of the described resource (E84). This kind of information is useful to knowledge base clients.
- ese:isShownAt - URI of a presentation of the described resource in its full information context (E84).
- *ese:unstored* – this is additional information that cannot be otherwise assigned to the available tags. The best thing we can do with it is save it as-is with hope of applying linguistics tools in future.

8 Results and Statistics

The prototype implementation of the Knowledge Retrieval System is functional and has been tested on a large sample of the DLF metadata. The tests have been performed on a desktop computer with the following characteristics: Intel Core 2 Quad CPU, Q9650 3.00GHz, Windows 7 64b, 4GB RAM. In future the system is supposed to work on a cluster of dedicated servers with a significantly higher amount of RAM, so we expect shorter times even with larger amounts of data.

The RDF repository used to hold the knowledge base is Ontotext's BigOWLIM (http://www.ontotext.com/owlim/editions) to which we have been granted a research license. BigOWLIM is both an RDF/OWL repository and a reasoning engine, one of the few that are able to perform the reasoning using their persistence mechanisms and not fully in the limited RAM memory, which is a critical condition if the reasoning engine is to be used in a system with millions or billions of triples.

This section presents statistics concerning the mapping and reasoning process, the contents of the resulting knowledge base, and some information about sample queries.

8.1 Knowledge Base Construction Process

Table 1 presents the times of the knowledge base creation with different amounts of starting DLF data. A *record* is a description of one publication in the DLF metadata schema. The first row of the table (*0 records*) contains data about the preparation processes in which all the ontologies and the KABA hierarchy are loaded. The results of this process are constant, hence it does not need to repeated every time the knowledge base is created. This is why the initialization time is not counted as part of the *processing time* in the subsequent rows.

The knowledge base creation process is described in section 5. The process is comprised of data cleaning and normalization, mapping from DLF metadata schema to CIDOC, relation detection, data enrichment, and performing reasoning (i.e. computing the full closure) on the triples.

Explicit triples are triples that have been introduced to the repository as a result of the mapping, enrichment, and relation detection operations. Implicit triples are triples generated by the reasoner (using BigOWLIM's owl-horst-optimized rule set, [2]). The total number of triples in the repository is the sum of explicit and implicit triples.

There are two different numbers of implicit triples in the 500,000 records case. The reason for this is that after close inspection of the KABA data in the repository we discovered that the P127_has_broader_term relation used to build E55_Type hierarchies was not declared transitive in Erlangen CRM. After our request this has been corrected in a newer Erlangen version, hence the higher number of triples (now all parts of the *broader term* chain are connected with implicit triples, there also has been a number of other similar changes).

Table 1. DLF records processing times and the number of resulting RDF triples

No. of Records	Processing Time	Explicit RDF triples	Implicit RDF triples
0	2.5 hours	6,059,665	30,710,518
10,000	7 minutes	6,443,246	32,907,960
50,000	30 minutes	7,928,790	40,849,942
100,000	70 minutes	9,808,073	51,177,846
500,000	6 hours	23,040,700	125,212,980
			181,209,531[7]

We find the processing times on the limited hardware quite satisfactory. However, we have detected two performance problems with BigOWLIM (or, possibly, reasoning engines in general). One is the unacceptable reasoning time for long chains of OWL properties that are both transitive and symmetric (especially when long chains of transitive properties with inverse properties work fine). The other is the fact that if we populate the repository with no reasoning at all (the empty rule set) and turn the reasoning on later, the processing time is significantly longer than with instant reasoning (e.g. 13 minutes for 10,000 records, when instant reasoning together with the knowledge base construction takes only 7 minutes).

This is an important issue for us, because after mapping the XML schema to Erlangen CRM we want to perform some operations on the repository with the reasoning turned off. If the reasoning is turned on, the triple deletion operation becomes very costly, as the full closure has to be recomputed and all triples inferred from the deleted one have to be deleted as well. Some triples have to be deleted in the relation detection process, e.g. when two different instances are discovered to in fact represent the same entity. At this point it seems that it would be more efficient to perform the operations with the reasoning turned off, serialize the triples, and load them with reasoning turned on than to simply turn reasoning on in a working repository.

[7] This is the result for the newest version (2011-04-04) of Erlangen CRM in which on our request the P127_has_broader_term property has been declared transitive (which is in accordance with the CIDOC CRM specification).

8.2 Knowledge Base Contents

In accordance to the Semantic Web and Linked Data guidelines, in the knowledge base we are trying to reuse existing ontologies and vocabularies and connect our data (by means of URIs) to external recognized sources. In the prototype implementation we are enriching our data with information from VIAF (Virtual International Authority File), NUKAT (Polish National Union Catalogue) authority data, Geonames (a worldwide geographical database), TERYT (Central Statistical Office's administrative division register). We also try to map subjects to KABA, which is a subject headings language (and dictionary) based on the LCSH (Library of Congress Subject Headings).

We use the external sources not only to identify the entities in the knowledge base, but also to add new information that can be useful to its clients. For example, when we obtain location data from Geonames, we are able to unify all references to this location with different names (e.g. names in different languages), we download the coordinates information to allow for geographic proximity searching, and we remember all the alternative names, so that a query for the place using any one of them returns expected results.

Table 2. Contents of the knowledge base – links to external resources

No. of Records	VIAF Persons	NUKAT Persons	Other Persons	VIAF Institutions	NUKAT Institutions	Other Institutions	Geonames Places	TERYT Places	Other Plcs	KABA Subjects
10,000	518	37	860	139	10	304	422	10	3	805
50,000	1869	107	2,833	406	*29*	*968*	*1.277*	*47*	*3*	*1,966*
100,000	5,016	204	6,625	760	*70*	*2.385*	*4.230*	*277*	*6*	*4,378*
500,000	25,467	4,862	62,673	2,825	*242*	*15.396*	*8.901*	*608*	*77*	*16,353*

Table 2 shows the number of entities recognized in external data sources. A reason for optimism is that most people, places, institutions and subjects are repeated in the repository, which allows us to connect similar resources to each other. An interesting issue is the fact that there are some entities (people and institutions) that have been found in NUKAT and have not been found in VIAF which contains also the NUKAT information. One reason for this is that the copy of NUKAT data we received is newer than what the VIAF web service offers access to. Beside there are other two possible explanations: the web service API returns less good results than our local Lucene index query, or our local query should be more restrictive and exclude some results.

Locations are searched for both in Geonames and in TERYT, but Geonames have higher priority, as the service contains more information. The places that have been found in TERYT and not in Geonames are very small Polish villages.

Table 3 presents information about publications subjects. The *Total Publ.* column shows the number of distinct E73_Information_Object instances whose subject is of the given type (place, KABA subject, user subject). The two final columns show the

Table 3. Publication subjects distribution

Subject Type	Total Publ.	Subject Instances	Avg. Publ./Subj.
Place	249,129	8,742	28.5
KABA Subject	265,320	16,353	16.2
User Subject	328,389	75,844	4.3

number of used subject instances for each subject type and the average number of publications on one subject instance.

The number of KABA records used as a subject (Table 3) and the number of mapped KABA records (Table 2) are equal, because KABA records are only used as subjects. The number of mapped places and the number of places used as subjects are not equal, because places may play different roles in the knowledge base (they might for instance represent the place of publication).

8.3 Knowledge Base Queries

The knowledge base is to be used by the Integrated Knowledge System portal to allow for semantic (and more efficient) search of resources. The queries can be asked in SPARQL or SeRQL languages. Table 4 contains information about sample SeRQL queries to the knowledge base generated by processing 500,000 DLF records (over 200 million triples: 23 million explicit and 181 million inferred). The response times (which contain data retrieval times) are not fully satisfactory for a production system, but, as stated before, the impressive repository was held and queried on a desktop machine.

Table 4. Sample query times

Query	Time	Cached Time[8]	No. of Results
All places with their names	15s	199ms	9663
Titles of books about Warsaw	771ms	7ms	1273
Works of a person with surname “Mozart”	17ms	3ms	221
All places whose names start with “Nowy”	886ms	24ms	27

As an example below is the SeRQL listing to perform the last query (*all places whose names start with “Nowy”*):

```
SELECT geoId FROM
  {s} luc:luceneLiteralIndex {"nowy*"},
  {geoId} rdf:type {cidoc:E53_Place} ;   cidoc:P1_is_identified_by {}
cidoc:P3_has_note {s}
USING NAMESPACE
  cidoc = <http://erlangen-crm.org/current/>,
  luc = <http://www.ontotext.com/owlim/lucene#>
```

[8] There are several components in OWLIM that make use of caching (e.g. full-text search indices, predicate list, tuple indices). When a query is asked for the first time some parts of the result are cached, so similar queries asked afterwards are completed in shorter times.

9 Conclusions and Future Work

The automatic conversion of 500,000[9] metadata records described by the DLF schema (PLMET) to a Semantic Web ontology based on CIDOC CRM proved possible (within reasonable time) and brings promising results. It opens new possibilities for search and discovery of resources (e.g. displaying information about a person and all related resources, and not only textual fields for a given publication record), that will be used by the Integrated Knowledge System Portal. The knowledge base is enriched by additional information coming from external sources and contains links to other pieces of information that are not explicitly needed by the knowledge base or the portal, but can be obtained by users navigating in the Semantic Web environment. As the resources are identified by unique URIs, it is also easy for external services to link to the knowledge base.

There is still a large number of problems that should be solved to make the knowledge base truly useful. One problem is that external sources (e.g. Geonames) often return more than one possible result for a given string. At this point we use heuristics (e.g.: the most populated, but preferably a town rather than a region, etc.) but more context information should be included in the reasoning process (as it is possible that somebody is in fact referring Warsaw, Ohio, and not Warsaw, the capital of Poland).

Another issue is that of historical names of places. If a book from 1920 is about *Poland*, it may refer to towns that are now parts of another country. Works dedicated to the creation and management of a historical names database fall beyond the scope of the SYNAT project. However, the existence of such a database would be useful, so a separate project is being considered to handle this task.

An important (and challenging) position in our task list is the inclusion of linguistic tools (Fig. 2). One of the partners is the SYNAT project is Wrocław University of Technology, the creator of the Polish WordNet (plWordNet) [6] which would allow for further enrichment of resources and resource navigation possibilities.

Also, a very interesting task is the design of a convenient user interface to introduce data to the IKS portal. We want users to unambiguously describe the published resources (from their personal digital libraries) by means of an ontology (the extended CIDOC ontology). Instead of a very long metadata form, the users should be offered a dynamic "wizard" to introduce the data corresponding to the type of the resource being published. Linking to other resources (both internal and external to the IKS system) should be facilitated with innovative UI solutions.

Finally, the Digital Libraries Federation data is only the first step (although an important one) in acquiring and understanding cultural heritage information from heterogeneous sources. It is an inspiring challenge to create a knowledge base of interconnected information of different types: digital libraries, museums, archives, catalogues, etc.

[9] There are now more than 600,000 publications in DLF, but the tests have been performed on data obtained in mid-February 2011.

References

1. Atkins, A., Fox, E., France, R., Suleman, H.: ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations, 1.2 edn. (2008), http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html
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 (2010), http://www.semantic-web-journal.net
3. Clayphan, R. (ed.): Europeana Semantic Elements Specification, Version 3.3.1, January 24 (2011), https://version1.europeana.eu/c/document_library/get_file?uuid=a830cb84-9e71-41d6-9ca3-cc36415d16f8&groupId=10602
4. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.: Definition of the CIDOC Conceptual Reference Model, 5.0.2 edn. (June 2005), http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf
5. Daćko, D., Józefowska, J., Ławrynowicz, A.: An ontology based semantic library catalogue. In: Proceeding of the 3rd Language & Technology Conference, Poznań, pp. 109–113 (2007)
6. Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M.: Polish WordNet on a Shoestring. In: Proceedings of Biannual Conference of the Society for Computational Linguistics and Language Technology, Universität Tübingen, Tübingen, April 11–13, pp. 169–178 (2007)
7. Doerr, M.: Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM. Technical Report, 274, ICS-FORTH, Heraklion, Crete (July 2000), http://www.cidoc-crm.org/docs/dc_to_crm_mapping.pdf
8. Dudczak, A., Heliński, M., Mazurek, C., Mielnicki, M., Werla, M.: Extending the Shibboleth Identity Management Model with a Networked User Profile. In: Proceedings of the 1st International Conference on Information Technology, Gdańsk, May 18-21, pp. 179–182 (2008)
9. Fellbaum, C.: WordNet. An Electronic Lexical Database. MIT Press (1998)
10. Gill, T.: Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model. First Monday 9(5) (2004), http://firstmonday.org
11. 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)
12. 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
13. Kakali, K., Doerr, M., Papatheodorou, C., Stasinopoulou, T.: Wp5 - task 5.5 dc.type mapping to cidoc/crm. Technical report. Department of Archives and Library Science. Ionian University (2007)
14. Koutsomitropoulos, D.A., Solomou, G.D., Papatheodorou, T.s.: Metadata and Semantics in Digital Object Collections: A Case-Study on CIDOC-CRM and Dublin Core and a Prototype Implementation. Journal of Digital Information 10(6) (2009), http://journals.tdl.org/jodi/

15. Lewandowska, A., Mazurek, C., Werla, M.: Enrichment of European Digital Resources by Federating Regional Digital Libraries in Poland. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 256–259. Springer, Heidelberg (2008)
16. Lourdi, I., Papatheodorou, C., Doerr, M.: Semantic Integration of Collection Description. Combining CIDOC/CRM and Dublin Core Collections Application Profile. D-Lib Magazine 15(7/8) (2009), http://www.dlib.org/
17. Mazurek, C., Stroiński, M., Węglarz, J., Werla, M.: Metadata harvesting in regional digital libraries in PIONIER Network. Campus-Wide Information Systems 23(4), 241–253 (2006)
18. NUKAT, the National Union Catalog, http://www.nukat.edu.pl/
19. Paluszkiewicz, A.: Format rekordu kartoteki haseł wzorcowych: zastosowanie w Centralnej Kartotece Haseł Wzorcowych NUKAT. In: Formaty, Kartoteki, SBP, Warszawa, vol. 17 (2009)
20. Pasin, M., Motta, E.: Ontological Requirements for Annotation and Navigation of Philosophical Resources. In: Synthese, September 28, pp. 1–33. Springer, Netherlands (2009)
21. Purday, J.: Think culture: Europeana.eu from concept to construction. The Electronic Library 27(6), 919–937 (2009)
22. 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 2008), Dublin Core Metadata Initiative, pp. 203–203 (2008)
23. Tarjan, R.E.: Depth-First Search and Linear Graph Algorithms. SIAM Journal on Computing 1(2), 146–160 (1972)