Dati.CulturaItalia: a Use Case of Publishing Linked Open Data Based on CIDOC-CRM

Sara Di Giorgio[1], Achille Felicetti[2], Patrizia Martini[1], and Emilia Masci[3]

[1]Central Institute for the Union Catalogue of Italian Libraries (ICCU)
of the Italian Ministry of cultural heritage, activities and Tourism
(MiBACT), Rome, Italy
[2]PIN, VAST-LAB, Prato, Italy
[3] MIUR, Italy
{sara.digiorgio,patrizia.martini}@beniculturali
{achille.felicetti}@pin.unifi.it
{emilia.masci}@gmail.com

Abstract. In this paper we describe the pilot project dati.culturaitalia.it, which started in 2012 to build up a Linked Open Data (LOD) Service that will progressively make available open datasets from the web-portal CulturaItalia [1], the Italian national aggregator for Europeana [2]. CIDOC-CRM Ontology was used for transformation and representation of data widely pertaining to the cultural domain. RDF triples mapped into Erlangen CRM were then enriched with links to URIs identifying instances of internationally established RDF resources for geographic names, and instances of authority files for personal and corporate names, such as GeoNames and Virtual International Authority File (VIAF). CulturaItalia is the Portal of Italian Culture, promoted by the Italian Ministry of cultural heritage, activities and tourism (MiBACT), in which cultural institutions from all sectors and levels (national, regional and local) are involved. CulturaItalia also plays an important role for the development of Europeana, making available cooperative networks and agreements and coordinating technical activities leading to the establishment of Europeana environment.

Keywords: System interoperability, Data integration, Cross domain portal, CIDOC-CRM, Metadata Crosswalks, Linked Open Data, Semantic Web

1 Introduction

CulturaItalia [1], is the portal of the Italian Culture on-line since April 2008, managed by the Italian Ministry of cultural heritage, activities and tourism (MiBACT) through the Central Institute for the Union Catalogue of Italian Libraries (ICCU) [2]. The Web-portal indexes the main cultural databases and

[1] http://www.culturaitalia.it
[2] http://www.europeana.eu

gathers the metadata to Europeana, the public digital library promoted by the European Community. CulturaItalia is targeted to general users, by offering them a service for retrieving information on Italian culture from one access-point, and to more expert users, such as the operators in the cultural field, who can take advantage of a high-quality showcase to promote their own digital resources. CulturaItalia makes the digital resources interoperability possible, through a cross-domain Application Profile (PICO AP: PICO is the acronym for "Portale della Cultura Italiana On-line"), based on the Dublin Core Metadata Initiative technical guidelines. The Portal gives access to a rich "metadata" collection, which gathers and organizes information arriving from the various providers participating in the project. Users can discover different kinds of digital resources, describing the country's extensive cultural heritage (museums, photographs, libraries, archives, galleries, exhibitions, monuments, audio-visual works, etc.). The pilot project dati.culturaitalia.it started in 2012 with the aim to build up a Linked Open Data (LOD) Service that will progressively make available open datasets from the Web-portal. The application was designed by the CulturaItalia team with the technical and scientific support of Scuola Normale Superiore, and was developed by Meta s.r.l., to allow the resources aggregated by CulturaItalia to be involved into large semantic networks after exposing, sharing and connecting data according to LOD principles. A first release of this service is available on-line since 2013 [3] as a section, or sub-portal, of CulturaItalia dedicated to LOD. It presently makes available as LOD the Thesaurus PICO, adopted by the portal for facilitating the browsing of a variety of resources in its domain, and a selection of metadata sets from the Portal. The CulturaItalia team has chosen CIDOC-CRM, in the implementation of Erlangen CRM/OWL, to foster the interoperability in the cultural heritage sector. In the perspective of a future integration with the bibliographic heritage of the Open Catalogue of the National Librarian System (OPAC SBN), managed by ICCU, the Institute implemented, in 2014, a mapping activity, with the support of a team from VAST-LAB (PIN), to convert resources from OPAC SBN, encoded in UNIMARC format, in FRBRoo, adopting the CIDOC-CRM model.

2 CulturaItalia Application Profile and Thesaurus

CulturaItalia manages a catalog - called Index - which gathers and indexes metadata provided by the partners. The original data remain on the Web-site of the provider, to which the final user is redirected by CulturaItalia, through links, thus allowing to retrieve the original and complete information. For example, in the case of a photograph, in the CulturaItalia Index the preview image (thumbnail) is visible, together with some identifying data, and a link to the provider's website allows the user to visualize the photograph in its original format, accompanied by the complete information and services, in order to get the full benefit of the item. The resources in the Index are classified on the basis of the PICO Thesaurus, designed to manage and organize heterogeneous information, from

[3] http://dati.culturaitalia.it/

different cataloging systems. Browsing the Index, the user consults the metadata through a hierarchical classification of terms (facets). CulturaItalia is an "open" system: it grows up and develops together with the continuous enrichment of its metadata Index, through the metadata harvesting according to OAI-PMH, a protocol which allows the harvesting of metadata from content providers to one or more harvesters, adding services as indexing system or automatic classification. The Portal harvests metadata from different repositories and exports metadata to other national and international portals and repositories. At present CulturaItalia aggregates over 3 million metadata from 32 public and private partners including thematic aggregators, such as Internet Culturale, the portal of Italian Libraries, also created and managed by ICCU. Internet Culturale plays a key-role in guiding the libraries in the production of standardized digital cultural resources and metadata, according to the Italian standards. Metadata published in Internet Culturale are automatically transferred to CulturaItalia, and then, if the providing libraries agree, to Europeana.

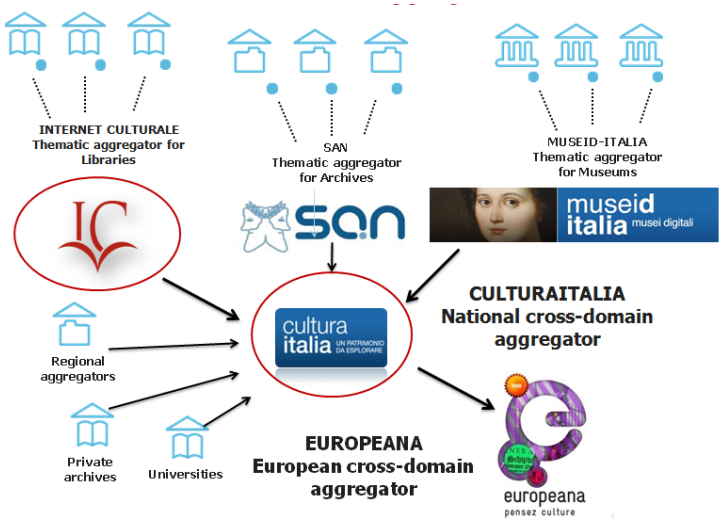

Fig. 1. CulturaItalia aggregation workflow

All content from external data sources are integrated in CulturaItalia in the form of metadata, thanks to the OAI-PMH harvesting protocol and published in the portal using a specific Application Profile (PICO AP), based on the international standard language, Dublin Core, that can describe, in a single scheme, every type of cultural resource, both physical and digital. Following DCMI recommendations, Scuola Normale Superiore di Pisa (SNS), which supports the ICCU working group engaged in the development of CulturaItalia, defined an application profile which joins DC Element set, Qualified DC terms and some further refinements and encoding schemes conceived for the applica-

tion of CulturaItalia. The PICO AP combines in one metadata schema all DC Elements, all DC Element Refinements and Encoding Schemes from the Qualified DC, and other refinements and encoding schemes specifically conceived to retrieve information pertaining to Italian culture. This Application Profile could be further expanded for harvesting possible unexpected contents in the future, by adding Refinements and Encoding Schemes that could be suitable for data retrieval. The PICO AP can be consulted at: http://purl.org/pico/picoap1.0.xml. Schemas used for the PICO AP are published on a PURL, under the domain PICO: `http://purl.org/pico/1.1/pico.xsd` and `http://purl.org/pico/1.1/picotype.xsd`. One of the most relevant encoding schemes introduced in the PICO AP is a Thesaurus specially conceived for the project itself, which comprehends hierarchically structured keywords indicating the topic of all the resources included into CulturaItalia (PICO Thesaurus 4.3). This ontology is also used to support the browsing into the Index of resources of Culturaitalia, therefore the assignment of a value taken from the PICO Thesaurus is mandatory for each metadata record. During the metadata generation, this assignment can be created for a whole repository or for a whole set, while in some other cases it was necessary to interpret a given value of the original database in order to create a mapping into the Thesaurus. The PICO Thesaurus is organized in four main categories: "Who" includes both people and corporate bodies; "What" comprehends tangible and intangible heritage, and all digital objects; "Where" covers Italian places (from regions to towns and villages) and "When" includes a list of chronological keywords associated to a sharp range of years. In order to be more compliant with international best practices, it seemed useful to adopt a SKOS format for the PICO Thesaurus. The SKOS format for Thesaurus PICO has also been designed to be extended and/or integrated with different thesauri pertaining to specific domains, managed by institutions that have a role in standardization, such as ICCD and ICCU, or to support multilingualism through the mapping between different national KOS.

3 Mapping between PICO Application Profile and CIDOC Conceptual Reference Model

PICO AP is a Dublin Core Application Profile. As already pointed out in the literature related to mapping between Dublin Core and CIDOC-CRM, for every value of the DC element "Type", specifying a type of a described resource, it must be specified a different mapping to a main entity of CIDOC CRM [3].

E.g.: IF DCMITipe = Image, THEN the described resource must be mapped as CIDOC-CRM entity = E38 Image. Consequently, each record encoded according to PICO AP will produce one main CIDOC-CRM corresponding entity, and the mapping of all the other PICO AP elements describing the resource will depend on the high-level mapping between the type of the resource and the corresponding CIDOC-CRM entity. Within PICO AP, dc:type element is mandatory and repeatable (occurrence: min 1, max unbounded) [4]. As a condition, it should always contain at least one value from DCMIType Vocabulary [5], (Col-

lection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text) or from PICOType Vocabulary [6] (CorporateBody, PhysicalPerson, Project). In the case that one PICO record contains more than one DCMIType and/or PICOType term, the mapping document (main DCMI/PICO Type term) specifies which must be considered the main term, according to which the mapping must be defined. Those simulations of complex mapping cases (between a PICO record containing more than one DCMI/PICO Type term and one CIDOC-CRM corresponding element) are described and are formulated on the basis of some real cases that can be found among CulturaItalia metadata resources (more than on a logic basis). Moreover, many of those cases are not real, and have been entered for completeness, just in case that in the future similar cases could occur. On the basis of the digital resources currently aggregated within CulturaItalia and of the PICO AP domain, the term DCMI Type = Physical Object is mapped to CIDOC-CRM entity = E22 Man Made Object (and not to E19 Physical Object). When DCMIType = Collection, the record generally contains many other DCMI/PICO Type terms. In all the cases, when "Collection" is present as a DCMIType, the PICO resource will always be mapped to CIDOC-CRM "E78 Collection" entity. On the basis of the mapping between the terms of DCMI and PICO Type Vocabularies, and CIDOC CRM entities, CulturaItalia resources encoded according to PICO AP can correspond to the following 12 CRM Entities:

E5 - - - - - - - Event
E22 - - - - - - Man-Made Object
E78 - - - - - - Collection
E28 - - - - - - Conceptual Object
E73 - - - - - - Information Object
E29 - - - - - - Design or Procedure
E33 - - - - - - Linguistic Object
E36 - - - - - - Visual Item
E38 - - - - - - Image
E39 - - - - - - Actor
E40 - - - - - - Legal Body
E21 - - - - - - Person

From this high-level mapping, based on the type of the described resource, derive different mappings between the various types of PICO AP resources and a corresponding CIDOC-CRM entity.

For a PICO resource with DCMIType= "PhysicalObject" (= crm:E22 Man Made Object), the PICO AP element $< pico : author >$ must be mapped as shown in figure 2.

For a PICO resource with DCMIType= "StillImage" (= crm:E38 Image), the same PICO AP element $< pico : author >$ will be mapped as shown in figure 3.

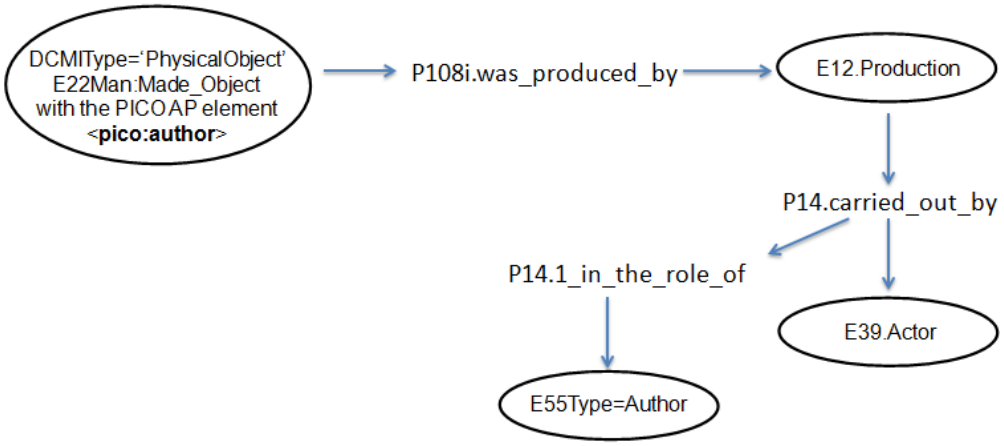

Fig. 2. Mapping of the PICO AP element $< pico : author >$ for a PICO resource with DCMIType= "PhysicalObject"

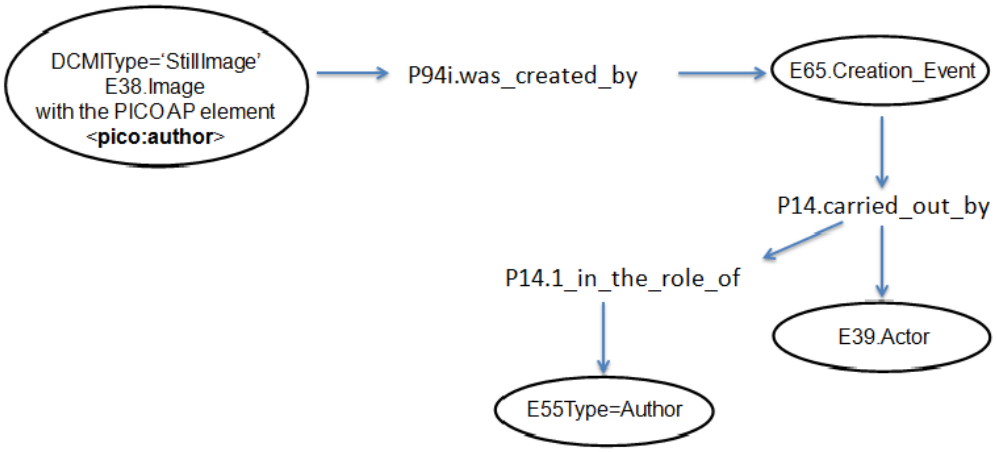

Fig. 3. Mapping of the PICO AP element $< pico : author >$ for a PICO resource with DCMIType= "StillImage"

As CRM entities are ordered in a poly-hierarchy, and as the properties associated to each class are inherited by the subclasses, it is possible to group the 12 identified entities (and, consequently, the mapping to be implemented) into 4 main groups:

1. E18 Physical Thing contains the mappings for: E22 Man Made Object and E78 Collection
2. E28 Conceptual Object contains the mappings for: E73 Information Object, E29 Design or Procedure, E33 Linguistic Object, E36 Visual Item, E38 Image
3. E39 Actor contains the mappings for: E40 Legal Body, E21 Person
4. E5 Event

The detailed mapping containing four mapping tables (one for each of the above-listed CRM main entities) is available on-line within the document: "Mapping between PICO Application profile and CIDOC Conceptual Reference Model" [7].

Figure 4 presents the main elements of the mapping related to E18 Physical Thing that contains the rules for E22 Man Made Object and E78 Collection.

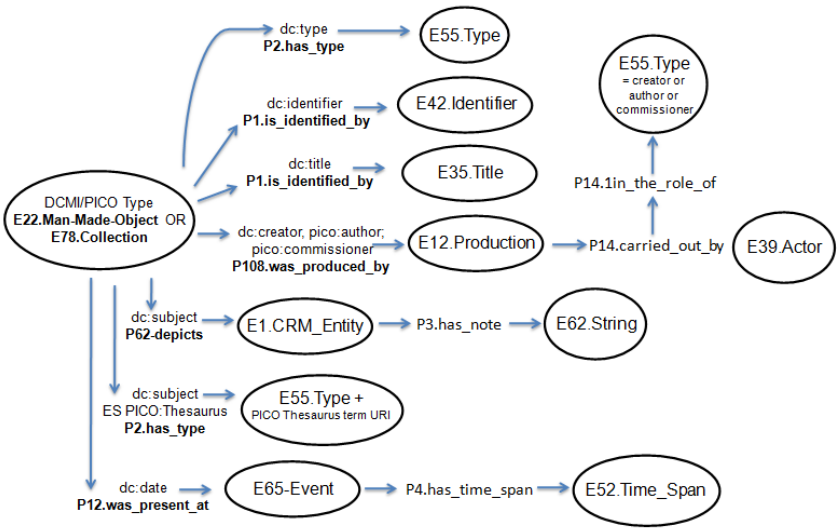

Fig. 4. Mapping of E18 Physical Thing

4 Dati.CulturaItalia

The pilot project dati.culturaitalia.it started in 2012 to build up a Linked Open Data (LOD) Service that will progressively make available open datasets from the web-portal CulturaItalia. A first release of this service is available on-line

since March 2013 [8] as a section of CulturaItalia dedicated to LOD. It presently makes available as LOD the Thesaurus PICO and metadata aggregated by the portal and licensed under CC0 1.0 - Universal Public Domain Dedication. These are data coming from: Accademia di Santa Cecelia, ArtPast Project, Digibess, ICCU, Internet Culturale, Michael Italia, Polo Museale Fiorentino, Regione Marche and Anagrafe delle Biblioteche Italiane. More datasets will be increasingly published as LOD, as soon as they will be delivered under CC0. CulturaItalia platform extracts the datasets, encoded in XML PICO format, that have been submitted by providers agreeing to take part of the pilot and to convert the PICO metadata into CIDOC [9] standard, according to the mapping document elaborated by M. E. Masci (Pisa, SNS). The mapping is implemented in an XML stylesheet and the result is an RDF/XML representation of each data provider's metadata. Then the CulturaItalia repository allows for the semantic enrichment with four types of reference resources (authority files):

- VIAF (Virtual International Authority File: www.viaf.org)
- GeoNames (www.geonames.org/)
- PICO Thesaurus in SKOS
- DCMI Type vocabulary

The SPARQL endpoint provides access to RDF metadata structured according to the CIDOC - Conceptual Reference Model in the implementation of Erlangen CRM/OWL. Data can be searched over three querying interfaces, corresponding to three sections of dati.culturaitalia.it:

- Text search: here it is possible to perform free text searches over all triples contained in dati.culturaitalia.it.
- SPARQL query: here you can try your hand at a SPARQL query. There are also some examples of queries.
- iSPARQL query: here there is an even more complex querying interface for advanced users.

Dati.culturaItalia.it exposes an OAI Provider that makes available XML or RDF metadata structured according to different schemas:

- oai-dc (xml): OAI-PMH schema adopted by Open Archives Initiative Protocol for Metadata Harvesting
- pico (xml): PICO Application Profile, the CulturaItalia Application Profile
- edm (rdf): Europeana Data Model, adopted by the portal Europeana EDM [10]
- cidoc (rdf): CIDOC - Conceptual Reference Model in the implementation of Erlangen CRM / OWL

5 Mapping between UNIMARC Bibliographic Format / SBN MARC and FRBRoo and next steps

ICCU is moving another step towards the Italian Linked Cultural Data Cloud by starting the mapping study of data from the OPAC SBN (On line Public

Access Catalog of National Library Service) in UNIMARC format to the class and the properties of FRBRoo, on the base of the model CIDOC CRM.
The collective catalogue of National Library Service provides access to 13.759.767 bibliographic records that contains:

- descriptions of documents acquired from SBN libraries starting from the '90s or since single libraries entered the SBN
- descriptions "book in hand" of documents of XVI - XX centuries
- descriptions obtained from catalogues on paper previous to 1990

In 2014 a working group formed by experts from ICCU and VAST-LAB (PIN) was established with the objective to analyze and test the publication of a subset of significant data in UNIMARC format as LOD according to the document *FRBR object-oriented definition and mapping to FRBR-ER (version 0.9)*. In particular, this activity focused on:

- analyzing and defining a basic methodology for creating Linked Open Data from bibliographic archives according to the international standards and cataloguing rules adopted by SBN;
- designing a schema with the conceptual description of how to relate SBN bibliographic information in a semantic way. FRBRoo, an harmonization between FRBR original conceptual model and CIDOC CRM, has been chosen as the reference intellectual guide for this activity
- selecting a first set of bibliographic records to be exported from OPAC SBN in UNIMARC format
- defining all the required namespaces and URI mechanisms to create meaningful identifiers for the converted UNIMARC entities

Activities performed by the working team lead to the definition of a mapping document describing the conceptual mapping between UNIMARC fields and FRBRoo entities and properties, with specific definition of mapping paths for every possible combination or special use cases of UNIMARC encoded information available in the SBN archive. The selected records subset was used for testing the conceptual coherence of the model in order to identify possible conflicts and to fix co-reference and cross-reference issues that might have arose.

Specific exporting scripts have been developed to encode the UNIMARC bibliographic information in a standard RDF format, to transform it in a machine-readable version using a formal language. Bibliographic information created in this way was afterward enriched with entities coming from VIAF, GeoNames, Linked Heritage, DBPedia, and other available online Linked Open Data resources.

A web tool has also been created to store semantic records and to query and retrieve relevant bibliographic data according with given semantic criteria. The tool is composed of various modules efficiently interacting with each other and based on open source technology. The modules include:

– an online triple store based on Sesame to accommodate the RDF triples created by the exporting framework and to manage the complex network of relationships defined by means of it;
– a set of responsive web interfaces based on Ajax/JQuery technologies and implementing the various features of semantic query and presentation of the relevant results. A basic faceted system for a more efficient browsing of the results was also implemented within the same interfaces.

The web tool also offers the possibility to download the full Linked Open Data network of bibliographic information in an RDF compatible format for local use. Further work on this topic will necessarily require a data cleaning phase for consolidating the legacy database in order to create a better representation of its content during the mapping and conversion process. Additional activities will concern the creation of a SPARQL end point for advanced semantic queries, the improvement of the web interface to allow connection of various libraries to the SBN index, to facilities retrieving and FRBRoo encoded triples representing entities of interest (work, expression, etc.) and the export of the same information in a standard Linked Open Data format for them to be used by other bibliographic tools and in other similar contexts. Data validation cycles to ensure the full compatibility of the formats with the fundamental principles of Linked Open Data and Semantic Web philosophy will be also performed, as well as multiple tests on the internal coherence of the newly created dataset.

References

1. http://www.culturaitalia.it/
2. http://www.iccu.sbn.it/opencms/opencms/it/
3. Main references for the present mapping: M. Doerr, Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM, Technical Report 274, ICS-FORTH, July 2000: http://www.cidoccrm.org/docs/dc_to_crm_mapping.pdf; K. Kakali, M. Doerr, C. Papatheodorou, T. Stasinopoulou, DC.type mapping to CIDOC/CRM, DELOS WP5-Task5.5, Department of Archives and Library Science / Ionian University, 26/01/2007: http://www.cidoc-crm.org/docs/WP5-T5_5-DC2CRMmapping-060728v0_2-final.doc ; I. Lourdi, C. Papatheodorou, M. Doerr, Semantic Integration of Collection Description. Combining CIDOC/CRM and Dublin Core Collections Application Profile, D-Lib magazine, July/August 2009, vol. 15 n. 7/8, ISSN: 1082-9873: http://www.dlib.org/dlib/july09/papatheodorou/07papatheodorou.html; M. Doerr, Updated graphical representation of the harmonized EDM-CRM-FRBRoo-DC-ORE models, September 2011: http://www.cidoc-crm.org/docs/EDM-DC-ORE-CRM-FRBR_Integration_ORE_fix.ppt
4. PICOAP/dc:type: http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/picoap/picoap1.0.xml#type
5. DCMI Type Vocabulary: http://dublincore.org/documents/dcmi-type-vocabulary/

6. PICO Type Vocabulary: http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/picoap/picoap1.0.xml#PICOType
7. M. E. Masci, Mapping between PICO Application Profile and CIDOC Conceptual reference Model version 1.0, 2013-01-24
8. See at http://dati.culturaitalia.it
9. See the document at http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/mapping/pico_cidoc/mapping_PICO_CIDOC-CRM_ITA-ENG.pdf
10. See the document at http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/mapping/pico_edm/Mapping-PICO-EDM-2.0.pdf