heritage

MDPI

Technical Note

Using LIDO for Evolving Object Documentation into CIDOC CRM

Regine Stein [1,*] and Oguzhan Balandi [2,*]

[1] Georg-August-Universität Göttingen, Göttingen State and University Library, 37070 Göttingen, Germany
[2] Philipps-Universität Marburg, Deutsches Dokumentationszentrum für Kunstgeschichte-Bildarchiv Foto Marburg, 35032 Marburg, Germany
* Correspondence: regine.stein@sub.uni-goettingen.de (R.S.); oguzhan.balandi@fotomarburg.de (O.B.)

Received: 31 January 2019; Accepted: 8 March 2019; Published: 25 March 2019

Abstract: Over the last years, many projects and institutions have worked on transforming object documentation from several existing cataloguing systems into a CIDOC Conceptual Reference Model (CIDOC CRM) compliant graph representation, as expressed in RDF. There were also various attempts to provide a generally valid path for the transfer of data from Lightweight Information Describing Objects (LIDO), CIDOC's recommended XML Schema for metadata harvesting, into representations that are suitable for the Semantic Web. They all face the challenge that a detailed mapping, which fully exploits the CRM's expressiveness and requires semantic assumptions that may not always turn out to be valid. Broad mappings, on the other hand, fail to leverage the potential of Semantic Web technologies. In this paper, we propose a method for using LIDO combined with an associated terminology as a means of evolving existing object documentation into CRM-based RDF representations. By clearly distinguishing between controlled vocabulary and ontology, it is possible to transform object data relatively easily into a minimized, though efficient structure using the CIDOC CRM ontology. This structure will open up the whole world of Semantic Web technologies to be used for further semantic refinement and data quality analysis through exploiting the underlying controlled vocabularies

Keywords: object documentation; LIDO; CIDOC CRM; semantic web

1. Introduction

Over the last years, many projects and institutions have worked on transforming object documentation from existing cataloguing systems into a CIDOC Conceptual Reference Model (CIDOC CRM) [1] compliant graph representation expressed in RDF, including e.g., work at FORTH on the X3ML Toolkit [2], the ResearchSpace project [3], or the Linked Art initiative [4], which all address a wider group of museums or research organizations. There were also various attempts to provide a generally valid path for the transfer of data from Lightweight Information Describing Objects (LIDO) [5] into representations that were suitable for the Semantic Web, e.g., in the context of the German Digital Library [6] (http://www.ddb.de), the Europeana-related project AthenaPlus [7], or Foto Marburg's infrastructure for art historical photo collections [8]. However, it turns out that establishing a workflow from a file- or record-based system to a semantic graph is a major challenge for many institutions. This appears to be particularly true for many museums, although they are willing to share their museum collection data and to contribute to current efforts of releasing it as linked open data (LOD) [9] on the semantic web.

Being involved in the past years in the development of such transformations for Foto Marburg's object documentation as well as for LIDO based aggregations, we experienced that balancing dependencies and conditions that result from the data analysis in mapping is a highly complex task.

The issues that are encountered are manifold on the technical level: moving from a schema to a formal system, reconciling object types with target classes, the requirement of linking data, the handling of non-qualified data elements, and so on. Additionally, on an organizational level, the challenges are often difficult to address, due to a lack of common understanding and decision-making in an institution regarding the required resources from the different departments.

In this paper, we propose a step-by-step approach to move from a file system to a semantic graph with a higher level CIDOC CRM model, and building up afterwards on rule-based semantic enrichment, instead of making semantic assumptions from the start that may not always turn out to be valid. The goal is to get a consistent and robust semantic graph in a first step and to bypass the technical and organizational problems. While our approach is based on LIDO as an intermediate format, it can also be applied to any other metadata schema.

After giving a rough overview of the important features and differences between LIDO and CIDOC CRM in Section 2 of this paper, in Section 3 we develop the proposed method and discuss its benefits in Section 4. Finally, conclusions and indications for future work are given in Section 5.

2. Background: CIDOC CRM and LIDO

The CIDOC Conceptual Reference Model (CIDOC CRM) is an abstract, object-oriented model that creates a frame of reference for merging and unifying cultural information and it is valid, regardless of specific implementation. It defines a formal ontology as a logical framework for the formulation of valid statements regarding cultural objects. A key concept of CIDOC CRM is to link statements regarding objects to events in its history. This event-centric approach makes it possible to map the properties of an object with references to the actors involved, to location, and time more precisely. Thus, it supports the (automatic) uncovering of correlations between originally scattered information and it contributes to the contextualization of objects. The CIDOC CRM does not specify the content of the documentation of museum objects, but it defines rules for the logical linking of information. It has been developed by CIDOC since 1996, and it became the ISO standard for data modeling in the cultural heritage area (ISO 21127:2014) [1,10].

Figure 1 shows the CIDOC CRM's top level classes useful for integration:

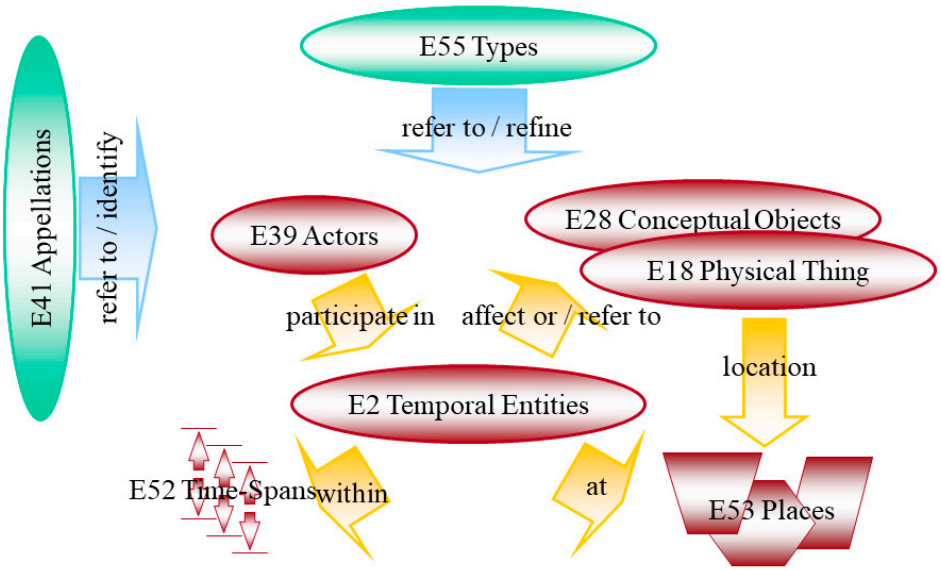

Figure 1. The CIDOC Conceptual Reference Model (CIDOC CRM) top level classes useful for integration. (Martin Doerr, CIDOC CRM Family. Harmonized Models for the Digital World: CIDOC CRM and Extensions, Nürnberg, 19.5.2015. http://www.cidoc-crm.org/sites/default/files/1_CRM_Family%28MD%29.ppt, p. 8).

LIDO is an XML harvesting schema, which builds on the CIDOC CRM and is intended for delivering metadata, for use in online services, typically in an organization's online collections database or in portals of aggregated resources, as well as exposing, sharing, and connecting data on the web. The strength of LIDO lies in its ability to support the full range of descriptive information regarding museum objects of different kinds, e.g., art, architecture, cultural history, history of technology, and natural history. An important feature of the LIDO schema is the systematic grouping of string-based, structured information about an entity, together with an (optional) identifier and a display element. This will allow for transferring string-based, non-qualified data into a semantic graph in a form, which is suitable for later automatic processing and normalization without minting URIs (Uniform Resource Identifiers) for each entity. LIDO integrates and extends its predecessor CDWA Lite [11] with elements of the CIDOC CRM and adopts, in particular the event-oriented approach [5,12]. For further background on evolving LIDO-based aggregations into linked data, we refer to [13].

The LIDO schema is currently being complemented by the LIDO Terminology, which is intended to further enhance the interoperability of LIDO data across different collections by introducing controlled type vocabularies for certain LIDO elements and attributes. The LIDO Terminology is committed to the Linked Open Data paradigm and it is available through an LOD service [14].

The LIDO schema defines complex types corresponding to the CIDOC CRM's top level classes (Table 1) and it reuses them wherever appropriate in the LIDO schema elements (Figure 2).

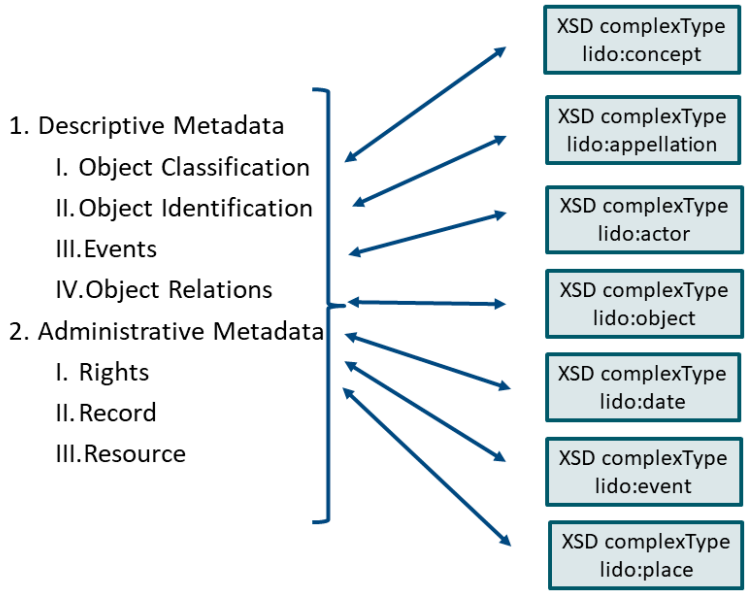

Figure 2. Overview of the Lightweight Information Describing Objects (LIDO) structure.

Table 1. CIDOC CRM classes versus complex types of the LIDO Schema.

CIDOC CRM Class	LIDO Schema—XSD Complex Type
■ E55 Type	■ lido:concept
■ E41 Appellation	■ lido:appellation
■ E39 Actor	■ lido:actor
■ E70 Thing	■ lido:object
■ E52 Time-Span	■ lido:date
■ E2 Temporal Entities → E5 Event	■ lido:event
■ E53 Place	■ lido:place

All of these complex types group—sometimes further refined—sub-elements together that identify and describe the entity. In addition, their type may be further qualified, e.g., the actor type or the event type, which is controlled by the LIDO Terminology and it may correspond to CRM classes, e.g., qualifying an actor as person (E21 Person) or an event as production (E12 Production).

In an overall comparison of CIDOC CRM and LIDO, the differences can be contrasted, as follows (Table 2):

Table 2. CIDOC CRM versus LIDO.

CIDOC CRM	LIDO
■ Formal Reference Model	■ Explicit format intended for harvesting
■ Ontology	■ Schema
■ Logic-based	■ Hierarchical structure
■ Graph-based data storing	■ XML-Document-based storing
■ Research-oriented	■ Documentation-oriented

3. Method

When starting a mapping of a schema to the CIDOC CRM ontology, it is the recommended approach to choose the most specific CRM class to which the data of a source element will map to: a seemingly easy task that, typically ends up with formulating and implementing—or rather trying to implement—many if-cases. The reason is that elements of the source schema tend to be multi-faceted and the semantics of an element differ depending on the actual data catalogued. For example, in a schema element 'production date', one will expect to find the date, referring to an instance of the CRM class E12 Production, but depending on the object type, e.g., for a print, it may actually be the creation date of the photograph and not the production date of the print. On the other hand, sticking with a broad mapping on the generic elements, in our example to E5 Event, fails to leverage the potential of Semantic Web technologies. To our experience, one has to choose, at some point, between either writing one single converter dealing with many if-cases or writing multiple converters, one for each collection or data set. Both of the solutions are not successful in practice due to:

- the complexity of the task,
- lack of resources and know-how in institutions,
- lack of respective decisions from policy makers, and
- discussions on the specification of the transformation (and possibly required extensions of the CRM) are delaying the process.

To further present our approach, first we need to distinguish between mapping and transformation.

3.1. Mapping Versus Transformation

A mapping is the transfer of the structure of a schema, without analyzing the actual data, while a transformation, on the other hand, typically includes semantic enrichment of the content. We illustrate the difference in the following figure (Figure 3), where, as throughout this paper, the cultural heritage object that is described in a LIDO document is referred to as the LIDO object.

In the mapping only the structure is analyzed: Without assuming anything about the nature of the LIDO object, ex:001 is an instance of the class crm:E70_Thing and the controlled vocabulary element qualifies its type. In the transformation, the LIDO object is an instance of a more specific class crm:E22_Man-Made_Object—the information is derived from the controlled vocabulary, where the concept "paintings (visual works)" is a narrower concept of "visual works (works)", for which the scope note explains they "occupy space [. . .] and are created, rather than naturally occurring".

In an application, the ontology one could choose to further refine the CRM and define e.g., a subclass ex:paintings_(visual_works) for crm:E22_Man-Made_Object. However, our approach is based on adopting the CRM as target ontology as it stands, while preserving the full information from the source data in the mapped data. Also note that such a subclass would be distinct from the AAT concept, which is an instance of E55 Type.

LIDO Example

```
{
 "lidoRecID": "001",
 "objectWorkType":{
                   "source": http://vocab.getty.edu/aat/300033618,
                   "term":   "paintings"
                  }
 }
```

Mapping to CRM

```
ex:001   a                 crm:E70_Thing,
         crm:P2_has_type   <http://vocab.getty.edu/aat/300033618>.
```

Transformation to CRM

```
ex:001   a                 crm:E22_Man-Made_Object,
         crm:P2_has_type   <http://vocab.getty.edu/aat/300033618>.
```

Figure 3. Mapping versus transformation.

In practice, a LIDO object may be many things: biological object, building, musical instrument, a cave painting, etc. Aiming at choosing the most specific CRM class requires in-depth analysis of the data, not only for the object types, but also usually for other elements, and it becomes more and more difficult to ensure the correctness of the instances against the ontology. Additionally, developing a general converter that any lido object converts to a specific CRM object becomes impossible. Therefore, we suggest that LIDO objects should be mapped to higher-level classes rather than specific classes of CIDOC CRM, and the actual data should be aligned with controlled vocabularies that can support further refinement afterwards.

3.2. Basic Mapping

As introduced in Section 2, LIDO consists of an XML schema and a complementing terminology, which provides controlled type vocabularies for some schema elements and attributes. The two parts are separately processed in the mapping. The basic idea is that the schema elements are mapped to high-level CIDOC CRM classes, while the elements from controlled vocabularies, including LIDO Terminology, are made available as E55 Type to specify the objects (see Figure 4). Based on our own data sets and use cases, we develop our converter for physical items as LIDO object, hence E70 Thing.

The main goal is to convert the LIDO objects in the first step into a lightweight CIDOC CRM compliant graph representation. Thus, the LIDO objects are initially provided in the CRM form. Building on this, the data can be further enriched, domain-specifically and semantically. Among others, Semantic Web technologies allow for rule-based semantic enrichment of the data. In the next Section, we show an example of how this works.

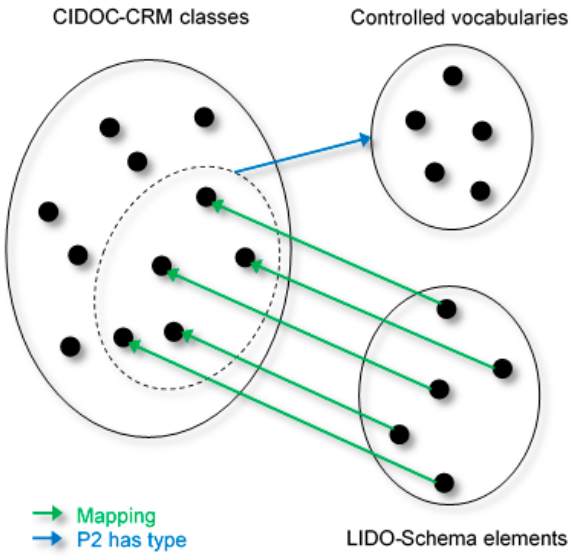

Figure 4. The basic mapping.

3.3. Rule-Based Transformation

An example of a LIDO record looks like Figure 5 (left side) after it has been mapped to the CRM form. It consists of high-level classes and E55 Type. Based on the types, rules are defined and executed to assign the objects to specific classes (Figure 5 right side).

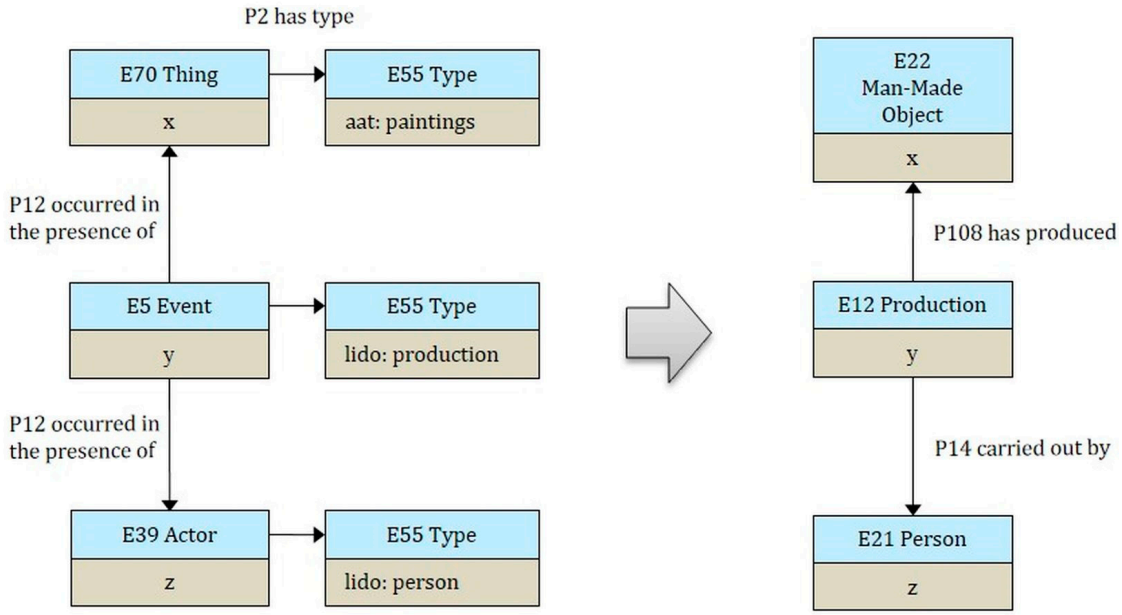

Figure 5. Rule-based transformation: visualization.

A rule consists of IF and THEN parts. In the IF part, the pattern is described, and in the THEN part, the operations are performed on the pattern. For example, if an entity ?x has type crm:E70_Thing and it is further qualified through the AAT concept aat:paintings (http://vocab.getty.edu/aat/300033618), then it has the type crm:E22_Man-Made_Object (Figure 6). If an entity ?y has type crm:E5_Event and is further qualified through the term lido:production (http://terminology.lido-schema.org/eventType/production) from the LIDO Terminology, then it has the type crm:E12_Production. If an entity ?z has type crm:E39_Actor and it is further qualified through the term lido:person (http://terminology.

lido-schema.org/actor_type/person) then it has the type crm:E21_Person. The following rule allows the transformation in Figure 5. This example shows the objects before and after the execution of rule-based reasoning.

```
[rule:
   (?x rdf:type crm:E70_Thing)
   (?x crm:P2_has_type aat:paintings)
   (?y rdf:type crm:E5_Event)
   (?y crm:P2_has_type lido:production)
   (?z rdf:type crm:E39_Actor)
   (?z crm:P2_has_type lido:person)
   (?y crm:P12_occurred_in_the_presence_of ?x)
   (?y crm:P12_occurred_in_the_presence_of ?z)
   ->
   (?x rdf:type crm:E22_Man-Made_Object)
   (?y rdf:type crm:E12_Production)
   (?z rdf:type crm:E21_Person)
   (?y crm:P108_has_produced ?x)
   (?y crm:P14_carried_out_by ?z)
]
```

IF

THEN

Figure 6. Rule-based transformation: in Code—Jena Rules.

Several refinements are derived from the type of information that is provided through controlled vocabularies: (1) RDF Type *E70 Thing* is refined to *E22 Man-Made Object*, (2) RDF Type *E5 Event* is refined to *E12 Production*, (3) Property *P12 occurred in the presence of* is refined to *P108 has produced*, (4) RDF Type *E39 Actor* is refined to *E21 Person*, (5) Property *P12 occurred in the presence of* is refined to *P14 carried out by*.

There are various rule engines and other technologies for semantically processing available data. The Apache Jena Rule Engine Notation illustrates this example [15].

4. Benefits

The practical implementation of such a converter and of a set of rules are currently underway, and the benefits of the approach are manifold:

- LIDO export features are widely available and implemented into a number of collection management systems, so using LIDO as an intermediate format can be a means to bypass the lack of technical know-how or resources within an institution and then provide data in a standard format for further processing e.g., in research projects.
- Curators, catalogers, and information specialists who actually create the data are used to work with controlled vocabularies, but they have difficulties to work with ontologies, such as CIDOC CRM. It is much easier to explain the benefits of working with controlled vocabularies and promote vocabulary mappings to them.
- LIDO, on the other hand, can handle string-based, structured (local) descriptions of entities, so the decision regarding which parts of the data are well consolidated for an LOD publication (e.g., only entities identified through published vocabularies) can be deferred and minting new URIs for entities that will not permanently be maintained can be avoided.

- Inconsistent states in CRM-based RDF representations are avoided, while the approach opens up for using Semantic Web technologies for further semantic refinement—which in turn may be used for data quality analyses and reveal inconsistencies in the source data.

In other words, the method allows for a "fast track" to the Semantic Web while retaining the potential of the CIDOC CRM.

5. Conclusions and Future Work

In this paper, we presented a method for using LIDO combined with an associated terminology as a means to evolve existing object documentation into CRM-based RDF representations. In an initial step, the approach has been validated in our local environment and the full implementation is subject to future work. This includes, in particular, the definition of rules that are based on the LIDO Terminology and e.g., the AAT, the evaluation and selection of a Semantic Web technologies-enabled toolset, and the provision of a mechanism for identifier handling, as a graph-based representation requires URIs for each entity addressed, which are not necessarily available from the source data.

We argue that by clearly distinguishing between controlled vocabulary and ontology, it is possible to relatively easily transform object data into a minimized, though efficient structure using the CRM ontology. This structure then opens up for the whole world of Semantic Web technologies to be used for further semantic refinement and data quality analysis through exploiting the underlying controlled vocabularies. The LIDO XML Schema, together with its recommended LIDO Terminology, provide useful features that help in bridging the gap between current object documentation and its representation in CIDOC CRM's encoding in RDFS.

Author Contributions: Conceptualization, R.S. and O.B.; methodology, R.S. and O.B.; software, O.B.; writing—original draft preparation, O.B.; writing—review and editing, R.S.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Crofts, N.; Doerr, M.; Gill, T.; Stead, S.; Stiff, M. (Eds.) Definition of the CIDOC Conceptual Reference Model—Version 5.0.4, 2011. Available online: http://www.cidoc-crm.org/get-last-official-release (corresponding to official ISO version of the standard) (accessed on 27 January 2019). Current Community Version: Available online: http://www.cidoc-crm.org/Version/version-6.2.3 (accessed on 27 January 2019).
2. X3ML Toolkit. Available online: https://www.ics.forth.gr/isl/index_main.php?l=e&c=721 (accessed on 3 March 2019).
3. Research Space. Available online: https://public.researchspace.org (accessed on 3 March 2019).
4. Linked Art. Available online: https://linked.art (accessed on 3 March 2019).
5. Coburn, E.; Light, R.; McKenna, G.; Stein, R.; Vitzthum, A. LIDO—Lightweight Information Describing Objects Version 1.0, 2010. Available online: http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf,http://www.lido-schema.org (accessed on 27 January 2019).
6. German Digital Library. Available online: https://www.deutsche-digitale-bibliothek.de/?lang=en (accessed on 3 March 2019).
7. Athena Plus. Available online: http://www.athenaplus.eu/ (accessed on 3 March 2019).
8. Infrastrukturen für ein Fotografisches Netzwerk. Available online: https://www.uni-marburg.de/de/fotomarburg/forschung/abgeschlossen/infrastrukturen-fuer-ein-fotografisches-netzwerk (accessed on 3 March 2019).
9. Berners-Lee, T. W3C Design Issues—A Roadmap to the Semantic Web—Linked Data. Available online: https://www.w3.org/DesignIssues/LinkedData.html (accessed on 27 January 2019).
10. Doerr, M. The CIDOC CRM—An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 24, 75, 2003. Available online: http://old.cidoc-crm.org/docs/ontological_approach.pdf (accessed on 27 January 2019).

11. CDWA Lite: XML Schema Content for Contributing Records via the OAI Harvesting Protocol. Available online: http://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.html (accessed on 27 January 2019).
12. McKenna, G.; Rohde-Enslin, S.; Stein, R. Lightweight Information Describing Objects (LIDO): The International Harvesting Standard for Museums, 2011. Available online: http://www.lido-schema.org/documents/LIDO-Booklet.pdf (accessed on 27 January 2019).
13. Simou, N.; Tsalapati, E.; Drosopoulos, N.; Stein, R. Evolving LIDO-based Aggregations into Linked Data. In Proceedings of the CIDOC 2012 Annual Conference "Enriching Cultural Heritage", Helsinki, Finland, 10–14 June 2012; Available online: http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/ConferencePapers/2012/simou.pdf (accessed on 27 January 2019).
14. LIDO Terminology. Available online: http://network.icom.museum/cidoc/working-groups/lido/lido-technical/terminology/ (accessed on 27 January 2019).
15. Apache Jena. Available online: https://jena.apache.org/ (accessed on 30 January 2019).