future internet

MDPI

Article

Representing and Validating Cultural Heritage Knowledge Graphs in CIDOC-CRM Ontology

Ghazal Faraj [1,*] and András Micsik [2]

[1] Data Science and Engineering Department, Faculty of Informatics, ELTE University, Pázmány Péter stny. 1/C., 1117 Budapest, Hungary
[2] Institute for Computer Science and Control (SZTAKI), Eötvös Loránd Research Network (ELKH), Lágymányosi u. 11., 1111 Budapest, Hungary; micsik@sztaki.hu
* Correspondence: ghazal.faraj@gmail.com

Abstract: In order to unify access to multiple heterogeneous sources of cultural heritage data, many datasets were mapped to the CIDOC-CRM ontology. CIDOC-CRM provides a formal structure and definitions for most cultural heritage concepts and their relationships. The COURAGE project includes historic data concerning people, organizations, cultural heritage collections, and collection items covering the period between 1950 and 1990. Therefore, CIDOC-CRM seemed the optimal choice for describing COURAGE entities, improving knowledge sharing, and facilitating the COURAGE dataset unification with other datasets. This paper introduces the results of translating the COURAGE dataset to CIDOC-CRM semantically. This mapping was implemented automatically according to predefined mapping rules. Several SPARQL queries were applied to validate the migration process manually. In addition, multiple SHACL shapes were conducted to validate the data and mapping models.

Keywords: ontology mapping; CIDOC-CRM; SHACL

Citation: Faraj, G.; Micsik, A. Representing and Validating Cultural Heritage Knowledge Graphs in CIDOC-CRM Ontology. *Future Internet* **2021**, *13*, 277. https://doi.org/10.3390/fi13110277

Academic Editor: Eirini Eleni Tsiropoulou

Received: 7 October 2021
Accepted: 27 October 2021
Published: 29 October 2021

1. Introduction

The COURAGE (Cultural Opposition: Understanding the Cultural Heritage of Dissent in the Former Socialist Countries) project aims to highlight the variety of alternative cultural scenes that flourished in Eastern Europe under strict government regulations prior to 1989. It was established to investigate the methods of cultural opposition during the socialist era from 1950 to 1990 [1]. The COURAGE project has an online resource description framework (RDF) store represented by high-quality linked data on cultural heritage (see Figure 1). The collection entity, the main focus of the project, is one of the primary entities in the COURAGE registry. Each collection has one or more linked featured items. All the following entities are connected to one or more collection(s) via one or more role(s). Some of these entities are historic people, groups, and organizations that played significant roles in the collection's history. Furthermore, the store included events that were significant in the history of collections [2].

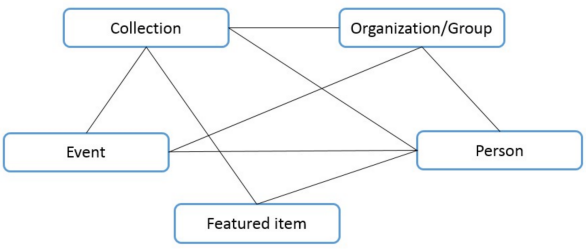

Figure 1. COURAGE registry.

Entity types and properties of COURAGE are organized into an ontology structure [2]. The base type in the schema is the historical item, which is used for describing entities from a historical viewpoint (see Figure 2). The main properties of this type are name, location, short description, and website. Subclasses of historical items focusing on content include collections, featured items, and publications. The second type at this level is the interview, which is considered a source of information. The third type is an event that has a start and end date and is linked to collections and agents. Agents are the most complex types in the schema; they can be people, groups, organizations, or networks. Agents may have various, time-limited roles in the lives of collections including creator, funder, operator, etc.

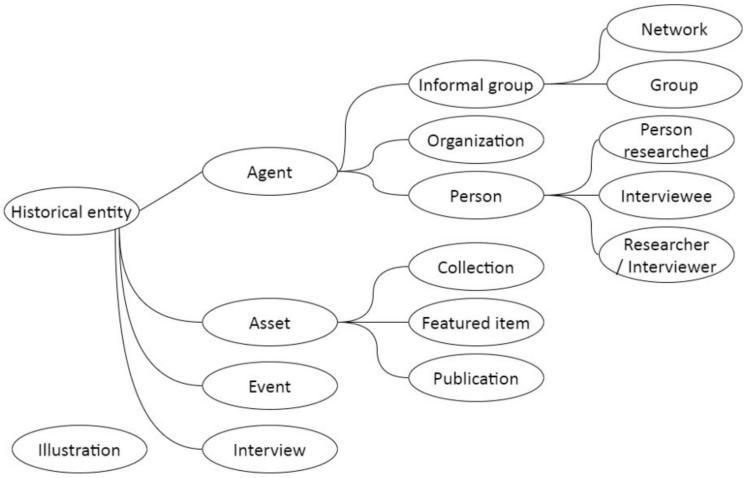

Figure 2. The main types of the COURAGE ontology.

CIDOC-CRM (CIDOC Conceptual Reference Model) is a formal ISO standard ontology that aims to integrate heterogeneous sources of cultural heritage information. This model can be used by cultural heritage institutions to describe their entities and improve knowledge sharing [3]. CIDOC-CRM provides definitions and a formal structure for the concepts and their relationships in cultural heritage documentation. CIDOC-CRM starts at the top level, with the classes of persistent and temporal entities. As a result, it can be regarded as a universal ontological model capable of describing people, objects, events, and activities alike [4]. It was developed for museums by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) to formalize the historical data in the cultural heritage field. CIDOC-CRM acts as a mediation between diverse cultural heritage concepts by providing the needed semantic glue to translate the heterogeneous datasets into cohesive sources (in Figure 3).

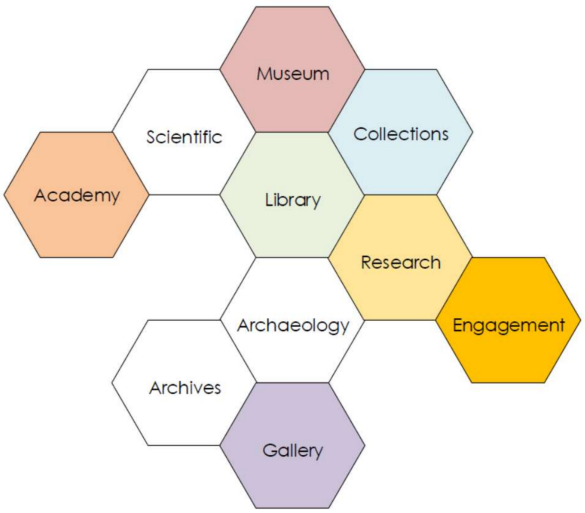

Figure 3. CIDOC-CRM is the semantic glue for cultural heritage datasets [5].

The CIDOC-CRM was selected for this research as it is the dominant schema in the cultural heritage field [6–8]. Using CIDOC-CRM also improves semantic integration and interoperability, as it is a standard and many datasets and research tools are using that representation. It already has many necessary concepts for archival descriptions, such as people, collections, events, and others [9].

This paper aims to prove the possibility of semantically mapping and aligning the COURAGE dataset, which represents an important part of recent European history that is hard to find in other datasets, to the CIDOC-CRM ontology [10]. By this mapping, the interoperability of this valuable data will be improved, and the COURAGE dataset may be linked and integrated with other populated CIDOC-CRM ontologies. Therefore, this paper introduces a mapping methodology and a validation technique to check the generated model. We semantically transferred the COURAGE dataset to the CIDOC-CRM ontology according to predefined mapping rules. These rules were created based on a consolidated set of principles and a console-based application. Several SPARQL queries were applied to check the mapped data and data model. However, the validation process was implemented not only through SPARQL queries to check the retrieved information through the Protégé and TopBraid Composer programs, but also via Shapes Constraint Language (SHACL) shapes. These shapes validated the generated RDF graph against a set of conditions. Thus, it is effective to ensure reliable links between various datasets without losing the specificity of the data's meaning. SHACL utilization supports checking the mapping model as well as the COURAGE data.

The remainder of this paper is divided into the following sections: Section 2 provides a brief resume of the current related work with CIDOC-CRM representation and the translation processes. Section 3 elucidates the alignment steps and the migration process. A discussion about the mapping rules is presented to ensure the final populated ontology's efficacy and coherence. Section 4 demonstrates the evaluation of the results of the migration process. A set of SPARQL queries and shapes were implemented to prove that the CIDOC-CRM representation was performed successfully. In Section 5, we discussed the migration process and its limitations and the final results that we achieved. We also demonstrated the challenges and faced problems. Finally, Section 6 concludes the whole research, as well as introduces further work and future evaluation.

2. Related Work

Many recent studies have worked on mapping their cultural heritage datasets to CIDOC-CRM. Several archival studies used ISAD (G) (General International Standard Archival Description) and ISAAR (CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons, and Families) in their mapping process to CIDOC-CRM.

The Portuguese National Archives were automatically migrated to the CIDOC-CRM standard [11]. The semantic migration process was applied according to both a predefined set of rules called "Mapping Description Rules" and both ISAD (G) and ISAAR recommendations. Other studies went with CIDOC-CRM after adding new classes and properties to its ontology to cover the source ontologies' entities, such as the level of description [12]. Another research [13] utilized CIDOC CRM and its extensions (CRMarchaeo, and CRMsci) to represent archaeological excavation activities and the archaeologists' observations. The proposed data model could serve as the foundation for developing an automated system for archaeological documentation and data integration. Another research introduced two CIDOC-CRM extension proposals [14]. The innovative concepts of the multiple interpretation data model (MIDM) were mapped into the CIDOC CRM. The extension to CIDOC-CRM is required due to the fact that some MIDM concepts do not correspond to current CRM entities and properties. The mapping manuscript migrations (MMM) system aims to model and publish three heterogeneous manuscript databases (Bibale, Schoenberg Database, and Medieval Manuscripts in Oxford Libraries) [15]. The MMM data model is based on FRRBoo and CIDOC CRM [16]. SILKNOW is a research project that presents European silk heritage from the 15th to the 19th centuries in a semantic manner [17]. The used data model in SILKNOW is based on CIDOC-CRM. The Corago repository has historical data related to opera during the 16th–20th century [18]. This repository was translated into the Corago Semantic Model according to the CIDOC-CRM and FRBRoo ontologies. On the other hand, some studies preferred mapping to the Europeana Data Model (EDM) over CIDOC-CRM. EDM is easier to use for museums [19], although there is Linked Art, which follows CIDOC-CRM [20]. Linked Art is a linked open data-based model that provides patterns to publish data on art. The authors of the paper [19] investigated RDFization, which transforms raw data into RDF format, from the museums' viewpoint.

The evaluation of the above mentioned mappings was limited to generating questions for researchers [11,15], evaluating a sample set of archival records [12], submitting a questionnaire to several categories of users [18], or checking inferred properties and SPARQL queries [14].

The introduced approach in this research shares common aspects and follows comparable steps to the papers [14,18] in aligning with the CIDOC-CRM and adding additional classes and properties. However, the other approaches had quite different representations of historical facts and missed using SHACL constraint checking and rule inferencing for model validation. Consequently, this paper maps valuable data collection to CIDOC-CRM in order to be more formalized and accessible. It also demonstrates the applied mapping process in detail due to COURAGE characteristics which are different from the previously mentioned knowledge graphs. Due to this mapping, a validation process has to be implemented to check the COURAGE data and the mapping model. The paper shows the significant role of SHACL in validating both the mapping model and the data model. In conclusion, this paper has a novel approach by mapping the COURAGE dataset to CIDOC-CRM and validating the mapping and data models with both SHACL and SPARQL queries.

3. Aligning COURAGE Entities to the CIDOC CRM Ontology

This section presents how the CIDOC CRM data was generated for the COURAGE dataset, starting with the process of mapping COURAGE ontology classes and properties to CRM classes and properties, and ending with the generation of triples.

3.1. Mapping of Classes

This section explains the preliminary investigations and decisions for the mapping. The first step was to decide upon which data needed to be converted into CIDOC CRM. We had to filter out private and personal data such as e-mail addresses or institutional budget numbers. Data for statistical purposes (e.g., approximated numbers and types of items in collections) and data very specific about collection operators (e.g., number of

online visitors, type of catalogue) were also omitted. Finally, we decided to include facts into the mapping process, which may be useful for historical research and fall into the scope of CIDOC CRM. The selected main classes include person, organization, collection, and featured item. Additionally, the auxiliary classes describing the roles of persons and organizations in the history of collections (e.g., founder, owner, etc.) should be mapped as well.

The second step was to select the target CRM classes for mapping. The selected COURAGE main classes had straightforward equivalents in CRM: E21 Person, E74 Group, E78 Collection, and E22 Man-Made Object. During the mapping of the rest of the classes, some extensions were necessary. In general, we tried to avoid the creation of new classes as much as possible. It occurred only if there was no equivalent class in CIDOC-CRM and it was a major concept for the context.

The first question was how to represent our specific type categories in CIDOC-CRM. As the E55 Type class is used to classify CIDOC-CRM instances, it was used extensively in our diagrams. For clarity, we decided to create different subtypes of E55 Type, namely E55.1 Profession, E55.2 Educational Background, E55.3 Operational Type, E55.4 Organisational Type, E55.5 Geographical Scope, E55.6 Topic, E55.7 Item Type, E55.8 Gender, E55.9 Name Type. These classes helped us to create each entity type once and use it later with other instances. This way, we could use the vocabularies developed in COURAGE in the CRM descriptions.

The second extension was initiated by the fact that certain roles had a single candidate class in CIDOC CRM: E87 Curation Activity. COURAGE differentiated between content creator, operator, collector, stakeholder, and supporter roles regarding collections. In order not to lose this important information, subclasses of E87 curation activity were created and used for the mapping, including E87a content creation activity, E87b operator activity, E87c collection activity, E87d stakeholder activity, E87e support activity.

3.2. Mapping of Properties to CIDOC-CRM

As the third step of the mapping process is the mapping of properties, the equivalent properties have been defined. These properties are listed in Table 1, Table 2, Table 3, and Table 4. Although mapping COURAGE properties to the CIDOC-CRM ontology was straightforward for some properties, in other cases it required analytical study to decide the best representation in CIDOC-CRM.

Table 1. COURAGE person entity representation in CIDOC-CRM.

Courage	CIDOC-CRM Entity	CIDOC-CRM Property
Given name	E82 Actor Appellation	P131 is identified by
Family name	E82 Actor Appellation	P131 is identified by
Label	E35 Title	P1 is identified by
Leader of	E85 Joining	P143 joined
Birthplace	E67 Birth	P7 took place at
Birthdate	E67 Birth	P98 brought into life
Profession	E55.1 Profession	P2 has type
Educational background	E55.2 Educational Background	P2 has type
Main member of	E74 Group	P107 has current/former member
Creator of	E12 Production	P14 carried out by
Main actor of	E5 Event	P11 had participant
Founder	E63 Beginning of Existence	P11.had participant
Owner roles	E8 Acquisition	P22 transferred title to
Creator roles	E87 Curation Activity	P14 carried out by
Operator roles	E87 Curation Activity	P14 carried out by
Collectorships	E87 Curation Activity	P14 carried out by
Stakeholder roles	E87 Curation Activity	P14 carried out by
Supporter roles	E87 Curation Activity	P14 carried out by

Table 2. COURAGE organization entity representation in CIDOC-CRM.

Courage	CIDOC-CRM Entity	CIDOC-CRM Property
hasOperationalType	E55.3 Operational Type	P2 has type
yearOfFunding	E66 Formation	P95 has formed
instType	E55.4 Organisational Type	P2 has type

The first decision happened regarding the expression of primitives, especially the time intervals of E52 time-span instances. In CIDOC CRM versions prior to 6 [21], there were P82a and P82b properties to denote the beginning and end of a time span. Contrary to paper [12], where they used the "P78_is_identified_by" property for time spans. We have chosen the use of P82a and P82b properties, so these two properties were added to the used version of the CIDOC CRM ontology. The benefit of applying these properties is that they can be used uniformly with both curation activities and birth and death events.

Table 3. COURAGE collection entity representation in CIDOC-CRM.

Courage	CIDOC-CRM Entity	CIDOC-CRM Property
Official name	E35 Title	P102 has title
contentLanguage	E56 Language	P2 has type
dateOfFounding	E63 Beginning of Existence	P4 has time-span
placeOfFounding	E63 Beginning of Existence	P7 took place at
collectionGeoScope	E55.5 Geographical Scope	P2 has type
hasTopic	E55.6 Topic	P2 has type
hasMasterpiece	E79 Part Addition	P110 augmented
Creators of content	E87a Content Creation Activity	P147 curated
Founders	E63 Beginning of Existence	P92 brought into existence
Operators	E87b Operator Activity	P147 curated
Collectors	E87c Collection Activity	P147 curated
Owners	E8 Acquisition	P24 transferred title of
Stakeholders	E87d Stakeholder Activity	P147 curated
Supporters	E87e Support Activity	P147 curated

Table 4. COURAGE featured item entity representation in CIDOC-CRM.

Courage	CIDOC-CRM Entity	CIDOC-CRM Property
masterpieceOf	E79 Part Addition	P111 added
masterpiece_type	E55.7 Item Type	P2 has type
hasItemTopic	E55.6 Topic	P2 has type

Entities of the COURAGE knowledge base have URIs ending in a unique and opaque identifier in the form of nxxxxx (where x stands for numbers). These URIs can be seen in the following figures, such as courage:n1014.

Another representation question was about names. The simplest and often used solution is to use the rdfs:label property to assign a name to an entity. The E41_Appellation class may be utilized instead of rdfs:label when there is a need to assign properties to the E41_Appellation [22]. As exemplified in Figure 4, E82_Actor_Appellation was used to represent the names of person and organization entities in order to prevent repeated names.

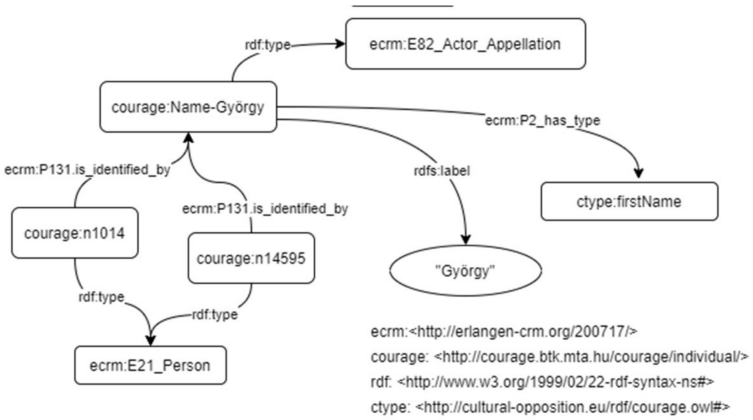

Figure 4. An example of a person first name representation in CIDOC-CRM.

CreatorOf is a property in COURAGE that connects people and organizations with their own featured items. It is represented through the E12 production class (see Figure 5), which includes all activities that create one or more new items.

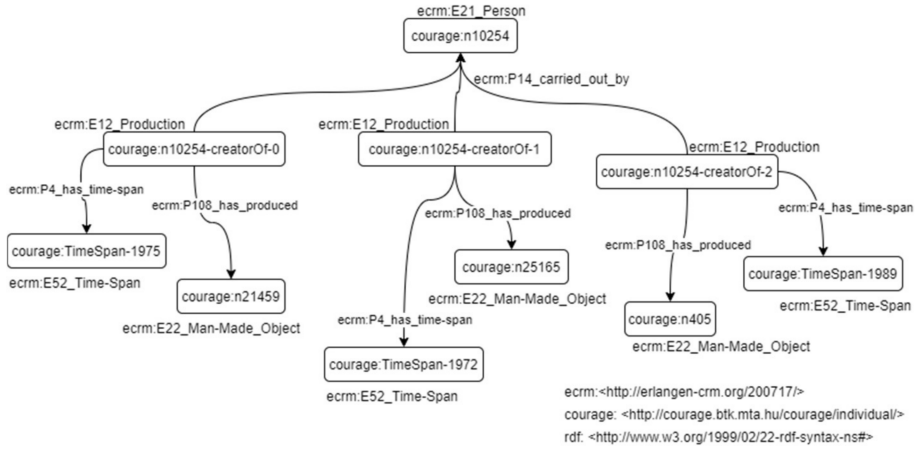

Figure 5. CreatorOf representation in CIDOC-CRM.

The owner role is represented by the E8 Acquisition class in contrast to other roles. Ownership of an item in COURAGE can be transferred to multiple people and organizations during that time (see Figure 6).

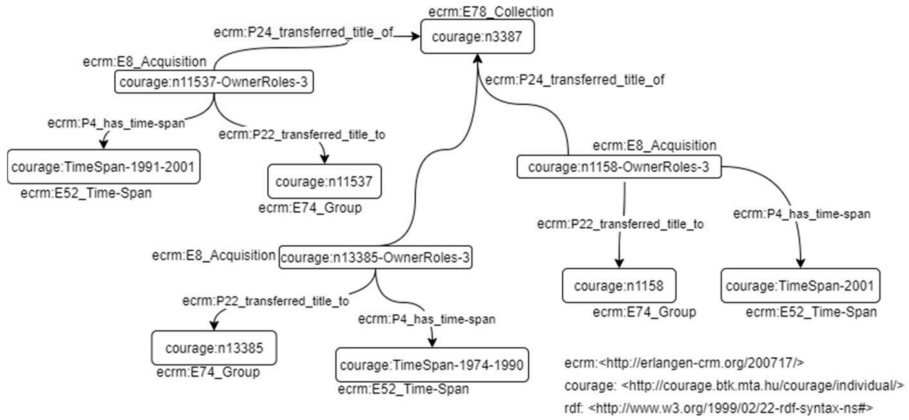

Figure 6. An example of owner role representation in CIDOC-CRM.

The founder role, which refers to the founder of a collection in COURAGE, was represented in CIDOC-CRM using the E66 formation. Then we found that the purpose of

the E66 formation is to form a group of people and its properties are inappropriate to our needs. Therefore, the E63 beginning of existence class was used instead (see Figure 7) as it includes all necessary connections involved in a foundation event.

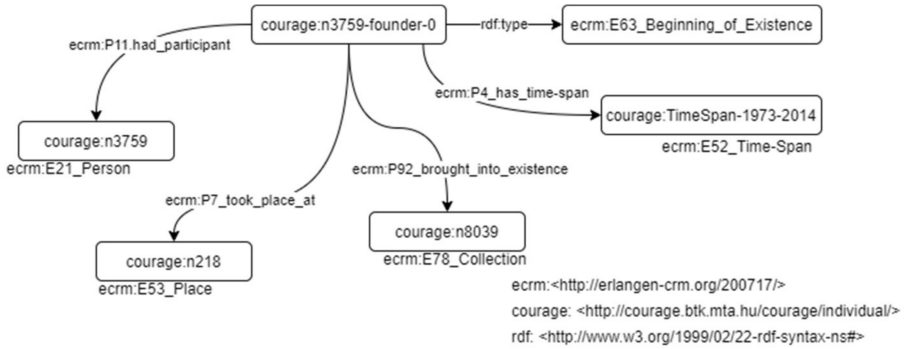

Figure 7. Founder role representation in CIDOC-CRM.

Roles in COURAGE (founder, owner, operator, collector, stakeholder, and supporter) link persons/organizations with collections/featured items. For example, a national library can have an operator role connected to several collections, and/or can be the owner of them. Figure 8 is an example of stakeholder role representation in CIDOC-CRM.

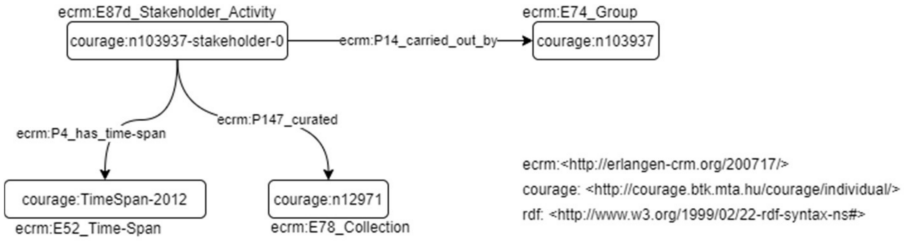

Figure 8. Stakeholder role representation in CIDOC-CRM.

Several properties in COURAGE, which are mentioned below in the tables, were mapped to the E55 Type class and linked to the entity via the P2_has_type property in CIDOC-CRM. For instance, the created type E55.4 organizational type was used to represent the types of organizations via the P2_has_type property. For each type, an instance of E55.4 organizational type was created once and linked to several organizations (in Figure 9).

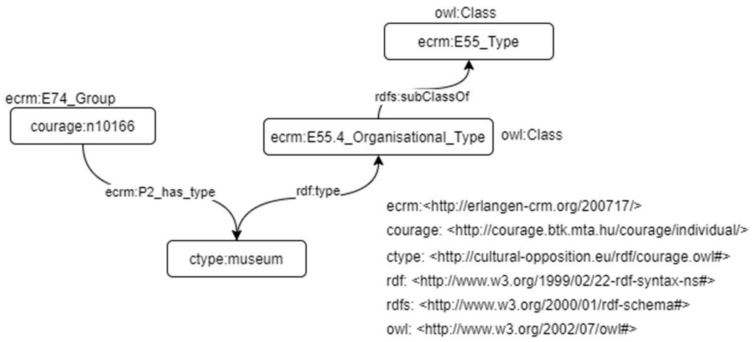

Figure 9. An example of an organizational type representation in CIDOC-CRM.

3.3. Mapping Tables

Mapping tables establish the basis for the automated migration process. These tables were developed in accordance with the previously mentioned principles.

People. The following steps were applied in order to translate a COURAGE person entity to the E21 Person entity in the CIDOC-CRM ontology. First, a list of essential

COURAGE properties with their equivalent classes and properties had to be determined (see Table 1)

Structural diagrams that clarify the mapping process were also created. For instance, Figure 10 demonstrates how the birthdate and birthplace properties in COURAGE were represented by the E67 Birth entity in CIDOC-CRM. In contrast to [12], we did not create an instance of E41_Appellation as there was no other property we needed to assign to it. Figure 10 exemplifies the representation mechanism utilized for all entity properties.

Organizations. The organization entity and the person entity in COURAGE share common properties. Organization entities were translated into the E74 Group entity type in CIDOC-CRM. Table 2 displays all the organization's particular properties.

Collections. A collection entity in COURAGE was mapped to the E78 collection type in CIDOC-CRM. The mapping rules for each collection's properties are as in Table 3.

Featured Items. A featured item entity (earlier named masterpiece) in COURAGE was represented by E22 man-made object, the most relevant class in CIDOC-CRM. The mapping rules for the featured item's properties are as in Table 4.

A creator role representation is shown as an example in Figure 11. The basis is E87 curation activity in CIDOC-CRM, here its subclass E87a content creation activity can be seen.

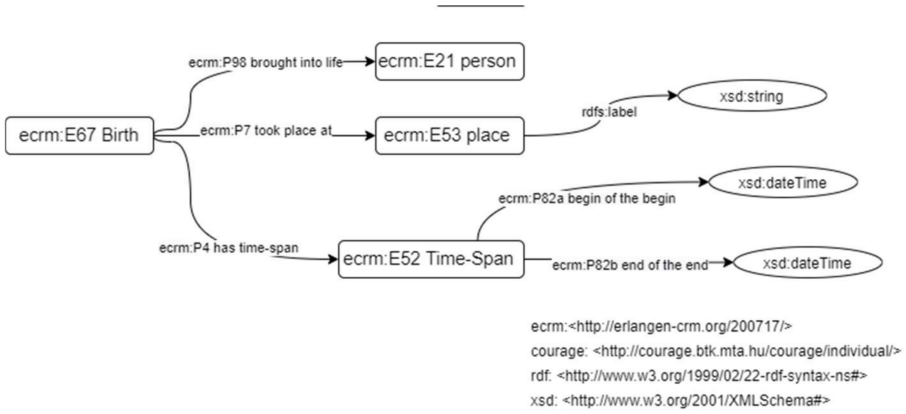

Figure 10. Birth event diagram in CIDOC-CRM ontology.

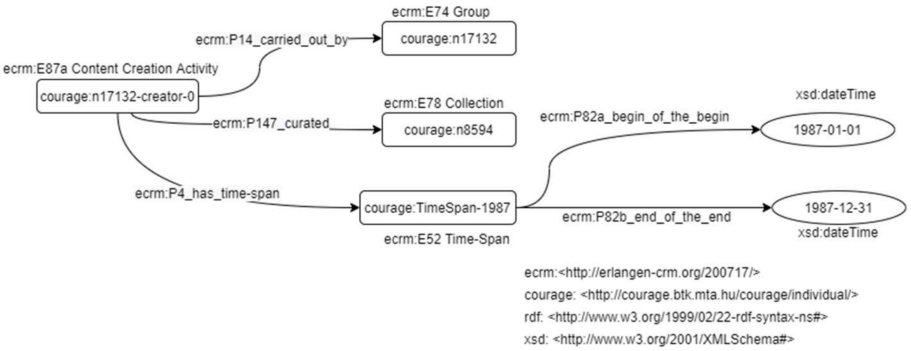

Figure 11. Creator role representation in CIDOC-CRM.

3.4. Generation of CIDOC CRM Facts

Based on the previously described mapping tables, a console application was built to generate an RDF file containing the selected COURAGE facts translated to CIDOC CRM.

Regarding fact generation, the first decision was to use the original COURAGE entities instead of creating their pairs for CIDOC-CRM representation. In this way, an entity has properties from both COURAGE and CIDOC-CRM ontologies and can be used in both 'ontological worlds' for inferencing or querying. This solution also makes validation and error detection easier.

As seen previously in the mapping tables and diagrams, a fact in COURAGE often requires the creation of several CRM facts, including new entities. These entities may be connected either with the described entity (e.g., collection names) or with a general 'entity pool' such as first names of persons (as appellation instances). In the first case, the entity id is based on the id of the described entity, with some discriminators appended to it. When the property may have several objects for the same subject, a number is also added to the URI. For example, if we have multiple names of an organization in multiple languages, we would use URIs like courage:n490770-1-Title, courage:n490770-2-Title for the new E35_Title instances, or just n49246-Birth for the new E67 Birth instance. In the second case, the new entities are identified by their type and content, for example, courage: TimeSpan-1956 denoting the year 1956.

First, we experimented with CONSTRUCT SPARQL queries to generate CIDOC CRM facts, which would have been a more interoperable solution. However, the creation of several new URIs within a CONSTRUCT query made this approach too difficultly, and a more manageable programmatic approach (using C#) was selected.

4. Analysis and Validation

For the purpose of validating and checking the generated knowledge graph, both SHACL and OWL-based techniques were selected. SHACL is a language for checking whether an RDF graph satisfies certain conditions given as "shapes" [23]. Using SHACL permits investigating string-matching patterns, value types, and other constraints. Moreover, the framework of SHACL supports high-level validation by expressing more complex conditions in the SPARQL query language. Validation of data and mapping were also performed based on OWL by running reasoner Pellet and SPARQL queries. This latter was performed via Protégé and TopBraid Composer to verify the consistency and correctness of mapped data manually based on inferred properties.

Protégé, a free open-source ontology editor, was used to validate the mapped data using Pellet. Pellet is a Java-based open-source OWL-DL reasoner. For instance, by running reasoner Pellet, an error was raised because the "P23 transferred title from" property was used to link E8 Acquisition to E78 Collection rather than E39 Actor.

After cleaning the generated RDF file from all syntax errors and bugs and running the reasoner Pellet, the inferred properties were displayed with a yellow background. In Figure 12, the inferred properties (p11i_participated_in, p12i_was_present_at . . .) of the person "n16143" are outputs of this reasoning based on the hierarchies.

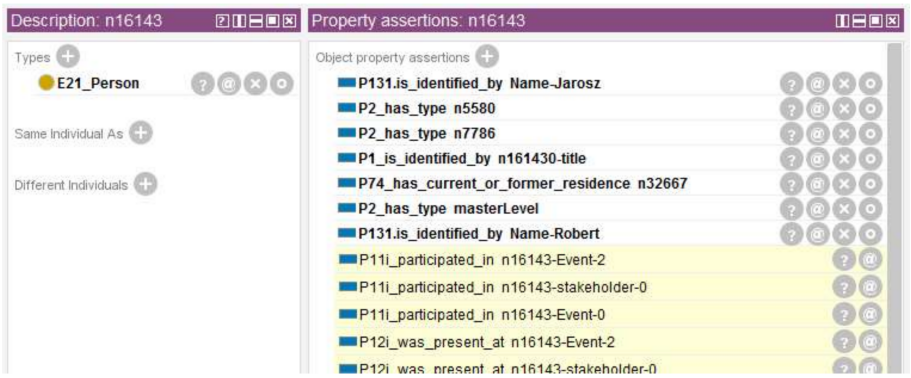

Figure 12. An example of inferred properties for a person n16143.

It was challenging to run SPARQL queries in Protégé due to performance issues. For instance, in Figure 13, we ran a SPARQL query to retrieve all collectors who collected a collection between 1988/01/01 and 2003/01/01.

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ecrm: <http://erlangen-crm.org/200717/>


SELECT DISTINCT  ?person ?hasCreationDate
WHERE {

      ?collectorRole  rdf:type ecrm:E87c_Collection_Activity.
      ?collectorRole  ecrm:P14_carried_out_by  ?person.
      ?collectorRole  ecrm:P4_has_time-span  ?tspan .
      ?tspan ecrm:P82a_begin_of_the_begin ?hasCreationDate.

      Filter(?hasCreationDate > "1988-01-01T00:00:00"^^xsd:dateTime  && ?hasCreationDate <  "2003-01-01T00:00:00"^^xsd:dateTime).
}
```

person	hasCreationDate
n10872	"1989-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n9448	"1999-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n11537	"1991-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n15031	"1999-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n14001	"2000-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n2674	"1994-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n25202	"1992-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n26048	"1989-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n42855	"1989-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n37719	"2000-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n24148	"1989-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n13653	"2000-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n1404	"2000-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
n39930	"1999-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>

Figure 13. SPARQL query about collectors between 1988 and 2003.

We chose TopBraid Composer over Protégé due to the fact that it permits generating SHACL shapes using a user-friendly interface. In Figure 14, for example, we executed the SHACL shape, which was expressed with more complex conditions in a SPARQL query, to check whether there were duplicated places with different URIs. The error messages with detailed information (focus node, value) are displayed in the SHACL Validation tab.

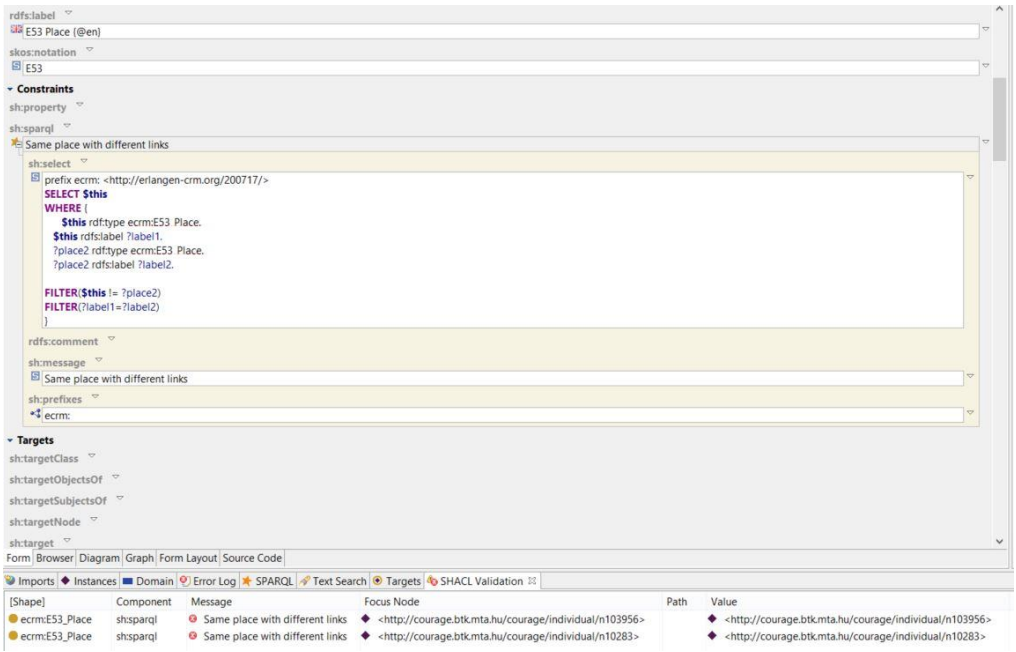

Figure 14. Validating same places with different URIs.

The overall validation can be classified into several categories:

1. Ontology-based validation: here the constraints defined using OWL can be validated, for example:
   - Type validation: checking whether the source data is mapped to the appropriate data type, and only to that. The validation of restrictions and cardinality constraints (if any) falls into this category as well;
   - Property validation includes the checking of the domain and range types and the cardinality of functional properties. For example, a painting is denied the birthdate property;

- Data validation: the format of data for data properties can be checked. For example, incorrect dates, incorrect language tags, or characters outside of UTF-8 may be detected.

2. Contextual validation is based on common sense and domain knowledge, and it includes constraints that are hard or impossible to represent in OWL or were simply forgotten during ontology implementation:
   - Temporal validation: the events should be in a logical order, e.g., birth, marriage, death. For example, events in a collection must follow its creation. Table 5 is an example of using SHACL to conduct temporal validation via TopBraid. The below shape inspects whether a person's birthdate occurred before the dates of all his/her activities;
   - Spatial validation: the locations must adhere to logical containment. For example, featured item I of collection C must be in the same location as C;
   - Uniqueness validation: there should be one URI and RDF representation for each entity. For example, we detected and merged multiple RDF individuals for cities and people;
3. Validation of mapping is used to check if properties and classes are translated in the right way.
   - For example, if the COURAGE P1 property should be translated into CRM P2, then for all facts using the P1 property, the existence of corresponding facts using the P2 property can be checked. In certain cases, the non-existence of other facts using P2 can also be checked. For example, if a collection is located in Budapest only according to COURAGE data, then in its CIDOC CRM translation the same collection should not have other locations either.

Table 5. Semantic validation using SHACL shape.

SHACL Validation
PersonShape rdf:type sh:NodeShape; rdfs:label " Validation of birthdate for E21 Person "@en; sh:sparql [sh:message "birthDate happened after createDate"; sh:prefixes <http://courage.btk.mta.hu/courage/individual>; sh:select """prefix ecrm: <http://erlangen-crm.org/200717/> SELECT DISTINCT $this ?birthDate ?createDate WHERE { $this ecrm:P98i_was_born ?birth. ?birth ecrm:P4_has_time-span ?birthIRI. ?birthIRI ecrm:P82a_begin_of_the_begin ?birthDate. $this ecrm:P14i_performed ?performed. ?performed ecrm:P4_has_time-span ?endIRI. ?endIRI ecrm:P82a_begin_of_the_begin ?createDate. FILTER (?createDate < ?birthDate). }""" ;] ;.

5. Discussion

The purpose of this work was to translate COURAGE into CIDOC-CRM. The COURAGE dataset is a valuable historical dataset created by historians with thorough quality control. At the beginning of the COURAGE project, the application of CIDOC CRM was not an option for several reasons; complete control was necessary over the ontology, as its structure was evolving in an agile way, and also the input of facts had to be supported with a historian-friendly user interface. There was no time to educate historians and explain to them the rationale behind the complexity of CIDOC CRM concepts. Therefore, the project decided to develop its own ontology for driving data input. Although simpler than CRM, the COURAGE ontology also embraced a temporal approach in the representation of human roles in the life of collections. As each role had a start and end year, these were easy to translate into CIDOC CRM timespans. Overall, we can claim that the transfer of the main relation graph (the essence of the COURAGE dataset) to CIDOC CRM was easy to

accomplish. Some new subclasses for curation task had to be created for the representation of the finer task typology in COURAGE.

However, the description of simple data properties in CIDOC CRM was found to be cumbersome. The description of collection topics, webpages, and answers to yes-no or numeric questions in the CRM schema becomes more complex than in a 'traditional' ontology.

Although CIDOC-CRM was designed to represent cultural heritage data [24], it was a serious mental task to find the right classes and properties for collection and artifact data. The cause may be that the naming of concepts we often found very abstract (for example, appellation or man-made feature is not easy to get familiar with or to explain to a layman).

The CIDOC-CRM version of our data has a more complex representation. In our example, the utilized COURAGE dataset includes approximately 31,500 triples, but it is represented in CIDOC-CRM by roughly 70,664 triples (21,426 instances). Moreover, the complexity of a SPARQL query in CIDOC-CRM is higher than its comparable query in COURAGE (in Supplementary Materials). For example, the queries in Figures 15 and 16 are retrieving the same data which is about selecting the creator of featured items belonging to a collection X (e.g., the Mimesis Collection (n26059)).

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX cn: <http://cultural-opposition.eu/rdf/courage.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?person ?GivenName ?FamilyName ?birthDate ?featured ?collection
WHERE {
            ?person rdf:type cn:Person.
            ?person cn:hasGivenName ?GivenName.
            ?person cn:hasFamilyName ?FamilyName.
            ?person cn:birthDate ?birthDate.
            ?person cn:creatorOf ?featured.
            ?featured cn:masterpieceOf ?collection.
            Filter( ?collection=<http://courage.btk.mta.hu/courage/individual/n26059>)
}
```

Figure 15. Sample query in COURAGE.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ecrm: <http://erlangen-crm.org/200717/>
PREFIX ctype: <http://cultural-opposition.eu/rdf/courage.owl#>

SELECT DISTINCT ?person ?GivenName ?FamilyName ?birthDate ?featured ?collection
WHERE {
    ?person rdf:type ecrm:E21_Person.
    ?person ecrm:P131.is_identified_by ?fname.
    ?fname ecrm:P2_has_type ctype:firstName.
    ?fname rdfs:label ?GivenName.
    ?person ecrm:P131.is_identified_by ?lname.
    ?lname ecrm:P2_has_type ctype:lastName.
    ?lname rdfs:label ?FamilyName.
    ?birth ecrm:P98_brought_into_life  ?person.
    ?birth  ecrm:P4_has_time-span  ?birthIRI.
    ?birthIRI  ecrm:P82a_begin_of_the_begin  ?birthDate.
    ?creator ecrm:P14_carried_out_by ?person.
    ?creator ecrm:P108_has_produced ?featured.
    ?part ecrm:P111_added ?featured.
    ?part ecrm:P110_augmented ?collection.
    Filter( ?collection=<http://courage.btk.mta.hu/courage/individual/n26059>)

}
```

Figure 16. Sample query in CIDOC-CRM.

The complexity of SPARQL queries demonstrated in the Figure 16 above may have the advantage of higher expressiveness, and the ability to formulate queries not possible otherwise. However, no expressivity problems were encountered with the original COURAGE ontology. On the other hand, it is useful if queries are easy to read and easy to compose. In the case of a simpler schema, even a scholar in the humanities may be able to create or customize SPARQL for its needs.

Overall, we expect that mapping to CIDOC CRM will improve the interoperability of this valuable data, and the COURAGE dataset may be linked and integrated with other

CIDOC-CRM datasets. Unfortunately, we have not found any public data using CIDOC CRM which could be relevant to connecting with our dataset.

6. Conclusions and Future Work

Providing a formal data structure and increasing the interlinks between different heterogeneous datasets results in the need to build a shared federated ontology. Through this ontology, we will be able to efficiently access a wide range of datasets on a large scale. The use of the CIDOC-CRM model is a guarantee that, on the one hand, there is already information available in the area of cultural patrimony that can be used to integrate and link with it. On the other hand, there are also many platforms available that can be used to explore the information migrated.

We presented the end-to-end mapping process by translating the COURAGE entities into CIDOC-CRM entities based on mapping rules. These rules established the basis for the automatic migration process. The evaluation of the results of the migration process was conducted manually through SPARQL queries and reasoners. Furthermore, SHACL was also used as a validation tool for ontology mapping. In contrast to OWL, SHACL is capable of a wide range of constraint validations, which comes in very useful when checking the results of a schema translation for datasets. Establishing a comprehensive validation model and generating SHACL shapes automatically is set as future work.

Supplementary Materials: The following files are available online at https://github.com/dsd-sztaki-hu/courage-crm, generated CIDOC-CRM data: MappedData.ttl, a sample of SPARQL queries and SHACL shapes: Sample of Validation Queries.txt. The SPARQL endpoint to test the mapped data is available here: http://cultural-opposition.eu:3030/dataset.html (accessed on 26 October 2021). The original COURAGE dataset and ontology are available at Zenodo. https://doi.org/10.5281/zenodo.3333540 (accessed on 26 October 2021).

Author Contributions: Supervision, A.M.; Writing–original draft, G.F.; Writing–review & editing, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not Applicable, the study does not report any data.

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

References

1. COURAGE Project Homepage. Available online: http://cultural-opposition.eu/ (accessed on 23 June 2021).
2. Csurgó, B.; Gárdos, J.; Kerényi, S.; Kovács, É.; Micsik, A. The Registry: Empirical and Epistemological Analyses. In *The Handbook of COURAGE. Institute of History, Research Centre for the Humanities*; Hungarian Academy of Sciences: Budapest, Hungary, 2018; pp. 27–49. ISBN 9789634161424.
3. CIDOC Conceptual Reference Model. 2019. Available online: https://en.wikipedia.org/wiki/CIDOC_Conceptual_Reference_Model (accessed on 23 June 2016).
4. CIDOC CRM Homepage. Available online: http://www.cidoc-crm.org/ (accessed on 20 June 2021).
5. Oldman, D.; CRM Labs. The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER. 2014. Available online: http://www.cidoc-crm.org/Resources/the-cidoc-conceptual-reference-model-cidoc-crm-primer (accessed on 1 July 2014).
6. Angelopoulou, A.; Tsinaraki, C.; Christodoulakis, S. Mapping MPEG-7 to CIDOC/CRM. In *Research and Advanced Technology for Digital Libraries. TPDL 2011*; Lecture Notes in Computer Science; Gradmann, S., Borri, F., Meghini, C., Schuldt, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6966. [CrossRef]
7. Meghini, C.; Doerr, M. A first-order logic expression of the CIDOC conceptual reference model. *Int. J. Metadata Semant. Ontol.* **2018**, *13*, 131–149. [CrossRef]
8. Doerr, M.; Bruseker, G.; Bekiari, C.; Emil Ore, C.; Velios, T.; Stead, S. ICOM/CIDOC Documentation Standards Group, Definition of the CIDOC Conceptual Reference Model, 7.0.1 edn, ICOM. 2020. Available online: http://www.cidoc-crm.org/Version/version-7.0.1 (accessed on 26 October 2021).
9. Sanfilippo, E.; Markhoff, B.; Pittet, P. Ontological Analysis and Modularization of CIDOC-CRM. In Proceedings of the Formal Ontology in Information Systems Conference, Bolzano, Italy, 14–17 September 2020; pp. 107–121. [CrossRef]
10. Erlangen CRM 200717. Available online: http://erlangen-crm.org/current-version (accessed on 15 June 2021).

11. Melo, D.; Pimenta Rodrigues, I.; Varagnolo, D. A Strategy for Archives Metadata Representation on CIDOC-CRM and Knowledge Discovery. Available online: http://www.semantic-web-journal.net/content/strategy-archives-metadata-representation-cidoc-crm-and-knowledge-discovery (accessed on 1 December 2020).
12. Koch, I.; Ribeiro, C.; Teixeira Lopes, C. ArchOnto, a CIDOC-CRM-Based Linked Data Model for the Portuguese Archives. In *Digital Libraries for Open Knowledge. TPDL 2020*; Lecture Notes in Computer Science; Hall, M., Merčun, T., Risse, T., Duchateau, F., Eds.; Springer: Cham, Switzerland, 2020; Volume 12246. [CrossRef]
13. Gergatsoulis, M.; Papaioannou, G.; Kalogeros, E.; Carter, R. Representing Archeological Excavations Using the CIDOC CRM Based Conceptual Models. In *Metadata and Semantic Research. MTSR 2020*; Communications in Computer and Information Science; Garoufallou, E., Ovalle-Perandones, M.A., Eds.; Springer: Cham, Switzerland, 2021; Volume 1355. [CrossRef]
14. Van Ruymbeke, M.; Hallot, P.; Nys, G.; Billen, R. *Implementation of Multiple Interpretation Data Model Concepts in CIDOC CRM and Compatible Models*; Technical University of Valencia: Valencia, Spain, 2018. [CrossRef]
15. Hyvonen, E.; Ikkala, E.; Koho, M.; Tuominen, J.; Burrows, T.; Ransom, L.; Wijsman, H. Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research. In *The Semantic Web*; Wiley: Hoboken, NJ, USA, 2021.
16. Bruseker, G.; Carboni, N.; Guillem, A. Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM. In *Heritage and Archaeology in the Digital Age. Quantitative Methods in the Humanities and Social Sciences*; Vincent, M., López-Menchero, B., V., Ioannides, M., Levy, T., Eds.; Springer: Cham, Switzerland, 2017. [CrossRef]
17. Schleider, T.; Troncy, R.; Gaitan, M.; Sebastian, J.; Mladenic, D.; Kastelic, A.; Besher Massri, M.; Leon, A.; Puren, M.; Vernus, P.; et al. The SILKNOW Knowledge Graph. *Semant. Web.* **2021**, 1–16. Available online: http://www.semantic-web-journal.net/content/silknow-knowledge-graph-0 (accessed on 26 October 2021).
18. Bonora, P.; Pompilio, A. Corago in LOD. The debut of an Opera repository into the Linked Data arena. *JLIS.it* **2021**, *12*, 54–72. [CrossRef]
19. Angelis, S.; Kotis, K. Generating and Exploiting semantically enriched, integrated, linked and open museum data. In Proceedings of the Special Track on Metadata & Semantics for Cultural Collections & Applications of the 14th Metadata and Semantics Research Conference, online, 30 November–4 December 2020; Springer: Berlin/Heidelberg, Germany, 2020.
20. Linked Art Homepage. Available online: https://linked.art/ (accessed on 20 October 2021).
21. Erlangen-crm/ecrm GitHub. Available online: https://github.com/erlangen-crm/ecrm/issues/5 (accessed on 14 June 2021).
22. Theodoridou, M.; Bruseker, G.; Daskalaki, M.; Doerr, M. Methodological tips for mappings to CIDOC CRM. Available online: http://www.cidoc-crm.org/Resources/methodological-tips-for-mappings-to-%0Bcidoc-crm (accessed on 1 August 2016).
23. Shapes Constraint Language (SHACL). Available online: https://www.w3.org/TR/shacl/ (accessed on 14 June 2021).
24. Kim, S.; Ahn, J.; Suh, J.; Kim, H.; Kim, J. Towards a semantic data infrastructure for heterogeneous Cultural Heritage data—Challenges of Korean Cultural Heritage Data Model (KCHDM). *Digit. Herit.* **2015**, *2*, 275–282.