heritage

MDPI

Article

CIDOC-CRM and Machine Learning: A Survey and Future Research

Yannis Tzitzikas [1,2], Michalis Mountantonakis [1,2,*], Pavlos Fafalios [1] and Yannis Marketakis [1]

[1] Institute of Computer Science, FORTH-ICS, GR-700 13 Heraklion, Greece; tzitzik@ics.forth.gr (Y.T.); fafalios@ics.forth.gr (P.F.); marketak@ics.forth.gr (Y.M.)
[2] Computer Science Department, University of Crete, GR-700 13 Heraklion, Greece
* Correspondence:mountant@ics.forth.gr

Abstract: The CIDOC Conceptual Reference Model (CIDOC-CRM) is an ISO Standard ontology for the cultural domain that is used for enabling semantic interoperability between museums, libraries, archives and other cultural institutions. For leveraging CIDOC-CRM, several processes and tasks have to be carried out. It is therefore important to investigate to what extent we can automate these processes in order to facilitate interoperability. For this reason, in this paper, we describe the related tasks, and we survey recent works that apply machine learning (ML) techniques for reducing the costs related to CIDOC-CRM-based compliance and interoperability. In particular, we (a) analyze the main processes and tasks, (b) identify tasks where the recent advances of ML (including Deep Learning) would be beneficial, (c) identify cases where ML has been applied (and the results are successful/promising) and (d) suggest tasks that can benefit from applying ML. Finally, since the approaches that leverage both CIDOC-CRM data and ML are few in number, (e) we introduce our vision for the given topic, and (f) we provide a list of open CIDOC-CRM datasets that can be potentially used for ML tasks.

Keywords: cultural informatics; CIDOC-CRM; machine learning; semantic data management; digital humanities

Citation: Tzitzikas, Y.; Mountantonakis, M.; Fafalios, P.; Marketakis, Y. CIDOC-CRM and Machine Learning: A Survey and Future Research. *Heritage* **2022**, *5*, 1612–1636. https://doi.org/10.3390/heritage5030084

Academic Editor: Nicola Masini

Received: 3 June 2022
Accepted: 4 July 2022
Published: 7 July 2022

1. Introduction

The CIDOC Conceptual Reference Model (CIDOC-CRM) is an ISO standard ontology for the cultural domain that is used for enabling semantic interoperability between museums, libraries, archives and other cultural institutions. It can be considered a very successful case since it is used by dozens/hundreds of institutions and, currently, in several ongoing research projects (http://www.cidoc-crm.org/useCasesPage (accessed on 1 July 2022)). CIDOC-CRM ontology is maintained regularly by a special group, and the current (community) version is 7.1.1 (https://cidoc-crm.org/versions-of-the-cidoc-crm (accessed on 1 July 2022)).

However, CIDOC-CRM is only one artifact. For achieving interoperability in a particular context, several processes and tasks related to this ontology have to be carried out, including schema and instance mapping, data transformation, information extraction, querying and others. It is therefore important to support and automate as much as possible these processes for facilitating interoperability. For this reason, in this paper, we describe the related tasks and then we survey recent works that apply machine learning (ML) techniques for reducing the costs related to CIDOC-CRM-based compliance and interoperability.

Since CIDOC-CRM is an ontology, one could argue that any ontology-based approach using ML is related. However, in this paper, we focus only on CIDOC-CRM since we are mainly interested in techniques that can tackle the difficulties stemming from the distinctive characteristics of CIDOC-CRM, i.e., that it is an event-centric ontology with a plethora of classes and associations structured in specialization hierarchies. After all, there are other surveys for the general case, i.e., about machine learning and ontologies in general, such

as [1] (where a ML approach is adopted for ontology matching) and [2] (where knowledge graphs are used as tools for explainable machine learning), etc.

Note that ML may be used in earlier steps, e.g., for preparing the structured data, such as extracting tabular content from PDF documents [3]. In general, there are several applications of ML for cultural heritage; e.g., [4] describes an automatic method for chronological classification of ancient paintings. We do not include such works; we focus on applying ML for achieving semantic interoperability. Moreover, note that the application of ML is also aligned with the direction proposed in [5], which stresses the value of data analysis and knowledge discovery and the need for tools that automatically find interesting serendipitous patterns in the data and even solve problems, preferably with explicit explanations. We also observe related initiatives, such as the EuropeanaTech Challenge for Europeana AI/ML (https://pro.europeana.eu/post/europeanatech-challenge-for-europeana-ai-ml-datasets-announcing-the-winners (accessed on 1 July 2022)) where one of the winners proposed using CIDOC-CRM.

Given the aforementioned requirements and directions and the huge volume of cultural data, in this survey, we (a) analyze the main CIDOC-CRM processes and tasks, (b) identify tasks where the recent advances of ML (including Deep Learning) would be beneficial, (c) identify cases where ML has already been applied (and the results are successful/promising), and (d) suggest tasks that can be benefited by applying ML. Since there are only a few works that leverage ML over CIDOC-CRM data, we (e) present our vision for the given topic by providing examples and (f) provide a list of open datasets expressed through CIDOC-CRM model, which can be potentially used for ML tasks.

The rest of this paper is organized as follows: Section 2 describes the required background (about CIDOC-CRM and related surveys); Section 3 describes processes and tasks related to CIDOC-CRM; Section 4 surveys works that involve both CIDOC-CRM and ML and discusses the main points from this collection and analysis, while Section 5 provides visionary examples and available datasets. Finally, Section 6 concludes the paper.

2. Background and Context

In this section, we describe CIDOC-CRM (in Section 2.1) and present related surveys (in Section 2.2).

2.1. CIDOC-CRM Model

The CIDOC Conceptual Reference Model (CRM) (http://www.cidoc-crm.org/ (accessed on 1 July 2022)) is a high-level, event-centric ontology of human activity, things and events happening in spacetime, providing definitions and a formal structure for describing implicit and explicit concepts and relationships used in cultural heritage documentation [6]. It is the international standard (ISO 21127:2014) (https://www.iso.org/standard/57832.html (accessed on 1 July 2022)) for the controlled exchange of cultural heritage information intended to be used as a common language for domain experts and implementers to formulate requirements for information systems, providing a method to integrate cultural heritage information of different sources. CIDOC-CRM has been used in a plethora of projects and data management activities related to (mainly) cultural heritage, history and archaeology (http://www.cidoc-crm.org/useCasesPage (accessed on 1 July 2022)).

Its last release (7.1.1) consists of 81 classes and 160 unique properties, with the longest path in the subclass hierarchy being a length of nine. The highest-level distinction in the CIDOC-CRM is represented by the top-level concepts of *E77 Persistent Item* (equivalent to the philosophical notion of endurant), *E2 Temporal Entity* (equivalent to the philosophical notion of perdurant) and *E92 Spacetime Volume*, which describes the entities for which its substance has or is an identifiable, confined geometrical extent in the material world that may vary over time. Figure 1 depicts the high level properties and classes of CIDOC-CRM and how they are connected. To showcase the richness of the model, Figure 2 (left) shows a visualization of the entire model as produced by WebVOWL (http://vowl.visualdataweb.org/webvowl.html (accessed on 1 July 2022)), while the right figure zooms in one part of that visualization. An

RDFS implementation of CIDOC-CRM 7.1.1 is also available (https://gitlab.isl.ics.forth.gr/cidoc-crm/cidoc_crm_rdf (accessed on 1 July 2022)).

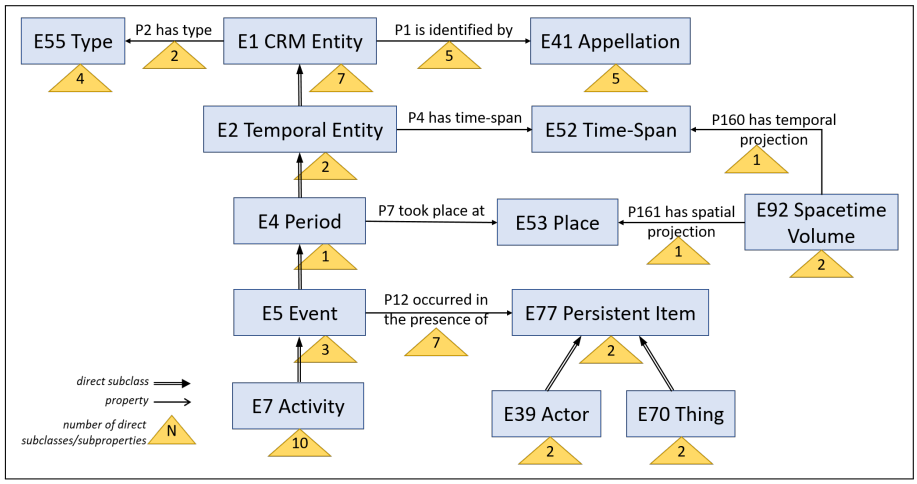

Figure 1. High level properties and classes of CIDOC-CRM.

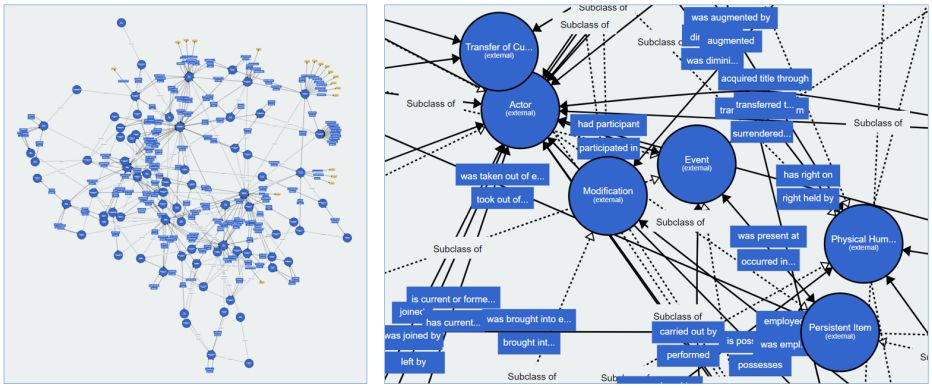

Figure 2. Visualization of the entire CIDOC-CRM (**left**) and of an excerpt (**right**).

2.2. Related Surveys

With respect to the related surveys, there are not many papers that attempt to survey this area. Thus far, we have found only a few papers that cover approaches exploiting either CIDOC-CRM data or machine learning algorithms for cultural heritage data.

The authors in [7] survey and classify 27 approaches that use CIDOC-CRM ontology, i.e., by merging, mapping or extending the mentioned model, for numerous tasks. Compared to our survey, the mentioned works do not focus on machine learning tasks over CIDOC-CRM data. In another survey [8], the authors study ML and CH literature to identify the theoretical changes that contribute to the algorithm and turn them into forms suitable for CH applications for the years 2015–2020. It lists several works that apply ML in the cultural domain for various tasks, including chronological classification of ancient paintings, prediction of painting's style, genre and artist, automatic annotation of visual contents in ancient manuscripts, classification of potteries and others. The major difference with our work is that they do not focus on specific ontologies (i.e., CIDOC-CRM). Moreover, the author in [9] provides a review about machine learning for archaeological data, by focusing on ML techniques that have been applied for geospatial, images, textual and numerical data, whereas it analyzes the advantages and limitations of those ML techniques for archaeological data. Similarly to the previous survey, the focus is not on a specific ontology.

Finally, there are available surveys covering topics that are related to cultural heritage and data mining; e.g., see the survey [10]; for issues related to the semantic integration of

linked data in general (i.e., including ontologies and data from any domain), at *large scales*, see the survey [11].

3. Using CIDOC-CRM: Processes and Tasks

The primary role of the CIDOC-CRM is to serve as a basis for the mediation of cultural heritage information and thereby provide the semantic 'glue' needed to transform disparate, localised information sources into a coherent and valuable global resource. Thus, its main use is for transforming one or more existing datasets to a CIDOC-CRM compliant dataset, i.e., to a rich semantic network of integrated information described through classes and properties of CIDOC-CRM. Nevertheless, there are also platforms, such as ResearchSpace [12], that allow documenting cultural heritage information, which is directly represented through CIDOC-CRM, i.e., thus creating a semantic network/knowledge base from the very beginning. Other systems, such as FAST CAT [13,14] and SYNTHESIS [15] for data transcription and documentation, include embedded processes that facilitate the construction of a CIDOC-CRM-compliant semantic network. Finally, there are processes where the main input is text; consequently, information extraction has to be performed [16].

These are elaborated in the subsections that follow: in particular, Section 3.1 describes the main use cases and processes, Section 3.2 describes the tasks of these processes, and Section 3.3 outlines the more time consuming tasks.

3.1. Use Case Scenarios and Processes

One main scenario starts with existing structured data (e.g., in relational databases) and the objective is to transform them into data expressed in CIDOC-CRM. A second scenario is where data have not been recorded but they have to be entered so the objective is to produce data expressed in CIDOC-CRM, but with less effort by humans. A third scenario is when we have data but they are unstructured, i.e., we have textual sources, and the objective is to produce data expressed in CIDOC-CRM by applying information extraction.

As regards structured data, we should also note that an attempt to formalize the processes and define a formal workflow has been carried out with the Synergy Reference Model (https://cidoc-crm.org/Resources/the-synergy-reference-model-of-data-provision-and-aggregation (accessed on 1 July 2022)). It is a reference model for a better practice of data provisioning and aggregation processes, primarily in the cultural heritage sector but also for e-science. It defines a consistent set of business processes, user roles, generic software components, and open interfaces that form a harmonious whole. The goal of Synergy Reference Model is the following: (a) describe the provision of data between providers and aggregators including associated data mapping components, (b) address the lack of functionality in current models, (c) incorporate the necessary knowledge and input needed from providers to create quality sustainable aggregations and (d) define a modular architecture that can be developed and optimized by different developers with minimal interdependencies.

Figure 3 shows the different scenarios and the main tasks involved in the creation of a CIDOC-CRM-compliant semantic network (a *knowledge graph*). The main tasks are described below.

3.2. Tasks

Here, we identify the main tasks that are required for supporting the aforementioned scenarios (presented in Figure 3):

1. **Modeling**. Sometimes, conceptual modeling extensions are required for tackling the requirements. Indicative papers that describe such extensions include [17] (for the provenance of digital objects), [18] (for archaeology), [19] (for geospatial extensions), [20] (for conservation processes), [13,21] (for maritime history) and others. Another related (complementary) modeling task is the one that aims at creating con-

trolled vocabularies/thesauri for the domain, which are used in conjunction with CIDOC-CRM.

2. **Transformation.** This task constructs ontological instances with respect to CIDOC-CRM (and/or any of its extensions) from the existing structured resources. Before actually transforming them, we have to perform the following: (a) define the schema mappings between the existing data and CIDOC-CRM and (b) specify the rules for generating URIs and values. The X3ML framework [22] provides a formal language called X3ML mapping definition language, as well as a set of tools for supporting the mapping definition process (e.g., 3M Editor (https://www.ics.forth.gr/isl/x3ml-toolkit (accessed on 1 July 2022))) and data transformation (e.g., X3ML engine (https://github.com/isl/x3ml/ (accessed on 1 July 2022))). We can identify the following subtasks:
    - Definition of mappings. This step provides the detailed guidelines demonstrating which parts from the original data will be used for constructing CIDOC-CRM based descriptions. They are called schema mappings and their role is to preserve and enhance the semantic descriptions of the existing data when they are transformed.
    - URIs and values generation rules specify how the identifiers and the values of the transformed data will be created. More specifically, they describe the syntax of URIs as a combination of constant values as well as values from the original data.
    - Data transformation uses the schema mappings and URI and value generation rules in order to create instances of the target model (i.e., CIDOC-CRM).

Note that, in some cases, the values that occur in an attribute of a source should become terms of a controlled vocabulary/thesaurus; therefore, the transformation could/should turn these values to references and into vocabulary terms.

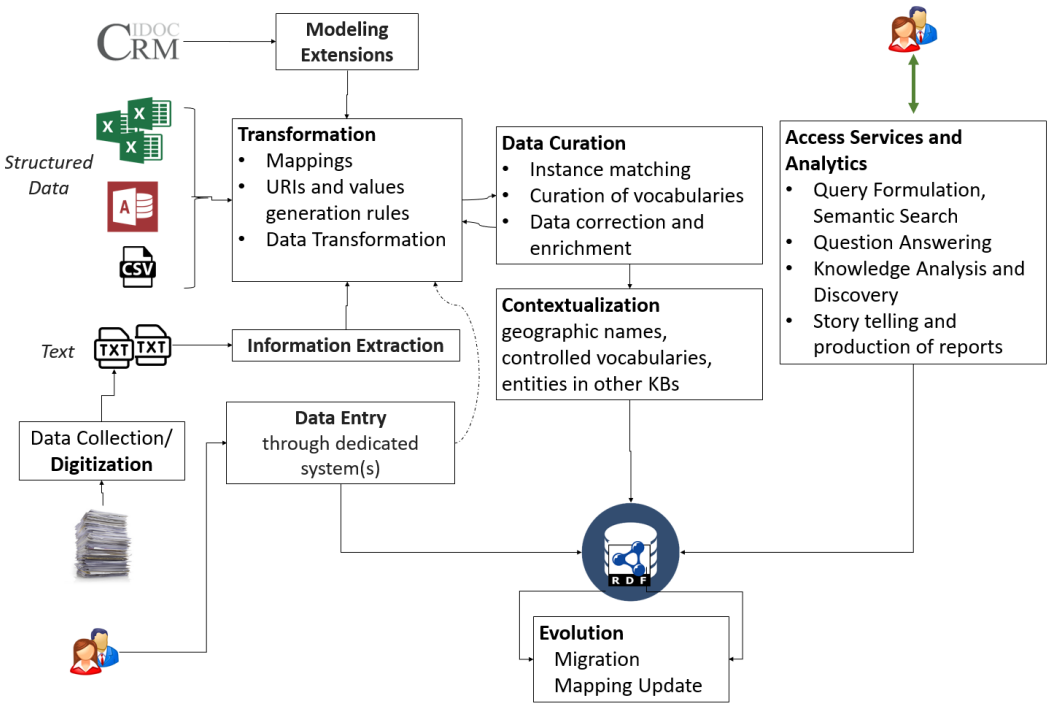

Figure 3. Different scenarios and involved tasks towards the creation of a CIDOC-CRM compliant semantic network.

3. **Data entry.** Data entry can be either manual (through some dedicated system) or automatic (e.g., through information extraction from texts). A method of entering

tabular data, which are subsequently transformed to CIDOC-CRM, is described in [13], while the *Synthesis* documentation system, offering embedded data transformation processes based on CIDOC-CRM, is described in [15]. Moreover, [23] describes a process for automating the creation of RDF triples from a repository containing life stories of common people.

4. **Data curation.** This refers to tasks that are required for connecting the transformed (or extracted) entities. This includes instance matching (entity matching and entity resolution), data linking, curation of vocabularies, data correction or enrichment, etc. For instance, [24] performs entity matching for finding existing person instances between the different datasets, which are described by using the CIDOC-CRM model, while [13] supports both automated (rule-based) and manual instance matching processes as well as manual vocabulary curation before data are transformed based on CIDOC-CRM. The authors of [25] evaluate the quality of mappings between COURAGE and CIDOC-CRM ontologies by using mainly SPARQL queries and reasoners. The author of [26] describes a process for relating paintings with events. Since CIDOC-CRM is event-centric, one important and challenging task is the instance matching over events.
   In general, the curation aims at improving the quality of the data, and this has several dimensions, including data validation, data completeness, quality of data interlinking and many others [11].
5. **Contextualization.** This refers to the linking the CIDOC-CRM based entities with external data (e.g., geographic names, controlled vocabularies and entities in a different knowledge base (e.g., DBpedia, wikidata, etc). For instance, [27] has linked events (described through CIDOC-CRM) to external sources, such as DBpedia. Moreover, [28] describes a tool for enriching the connections of a LOD (Linked Open Data [29]) dataset with the other LOD datasets, with emphasis on the cultural domain. It also quantifies the quality of data interlinking and detects possible errors for several datasets. In addition, [30] contains connectivity analytics of hundreds of RDF datasets, included five datasets expressed through CIDOC-CRM. For the latter datasets, a high connectivity was obtained with external RDF datasets, mainly of the cultural heritage and publications domain.
6. **Access Services and Analytics.** We can identify the following methods.
   - **Query Formulation** (from an information need to a SPARQL query over a CIDOC-CRM-based dataset). CIDOC-CRM is an adequate global schema, especially for integrating large amounts of cultural data, because of its rich schema of classes and properties and its event-based nature. However, this rich structure makes querying a complex procedure. A querying configuration that overcomes such problems is the fundamental categories and relationships [31]. These categories are bare classes covering the domain and the relationships are deductions from complex path expressions. Two well-known implementations of this querying configuration are ResearchSpace [12] and A-QuB [32].
   - **Question Answering** (and Dialog) over a CIDOC-CRM based dataset. The authors of [33] describe a QA system for the Cultural Heritage domain that gradually transforms input questions into queries that are executed on a CIDOC-compliant ontological knowledge base.
   - **Knowledge Analysis and Discovery**. This refers to analytics, knowledge discovery, and various mining tasks (as in [5]).
   - **Production of Reports.** This refers to the exploitation of the CIDOC-CRM-based knowledge base for producing presentations/narratives. For instance, [34] describes the process of producing narratives through the Narrative Ontology that has been implemented on top of CIDOC-CRM.
7. **Evolution.** We can identify the following subtasks.
   - **Migrate** an existing CIDOC-CRM based knowledge base to a newer version of the standard. If the new version is backwards compatible, then no issues can arise.

However, changes happen that require a few transformations for carrying out the migrations. The definition document of CIDOC-CRM provides migration instructions for all deprecated classes and properties (See the Appendix, pp. 229–232 at https://cidoc-crm.org/sites/default/files/cidoc_crm_v.7.1.1_0.pdf (accessed on 1 July 2022)). An example of such a migration instruction is the use of *E74 Group* for the deprecated class *E40 Legal Body*. Moreover, even if the new version is backwards compatible, the migration may cause loss of specificity that should be managed (see [35]).

- **Update a mapping** after a change in the source schema.
  The process can be complex, as described in [36] for marine data.

3.3. More Time Consuming Tasks

Even if this depends on the application context, from our experience on applying CIDOC-CRM, we have identified the following core tasks that are more crucial and time consuming:

- **Construction of mappings.** The cost is analogous to the number and size of sources, the attributes of each source that has to be mapped in addition to a cost at the end for entity matching over the aggregated transformed (in CIDOC-CRM) datasets and testing using queries. Moreover, it is a manual process and the overall quality of the implemented mappings relies on the experience of the person carrying them out and the good knowledge of both source schemata and the target ontology (e.g., CIDOC-CRM).
- **Data entry.** As mentioned earlier, data entry can be either manual (using some system), or automatic (e.g., using information extraction from texts). As regards manual data entry, its cost is analogous to the size of the data that have to be entered, which can be prohibitively expensive for a huge amount of data. Usually, the data are not entered in a graph-based format, but either through an appropriate system. Data can be entered in tabular form, as described in [13], and subsequently, they can be transformed to ontological ones; hence, the data entry cost includes the cost for mappings.
- **Instance matching.** If performed manually, this depends on the number of entities to be matched. If performed by custom rules, then this typically depends on the number of sources. If it is fully automatic, this depends on the effort required for checking (and approving/rejecting) matches. For example, the FAST CAT system [13] (in the context of maritime history) supports a multi-level instance matching process. A first process considers a set of source-specific rules for giving the same identity to a set of entity instances (e.g., all person instances in a specific source having the same firstname, lastname, father's name and birth date must be considered as the same person and obtain the same identity). Then, a second instance matching process allows historians (through a dedicated user interface) to manually indicate that two or more entity instances refer to the same real-world entity; thus, they must have the same identity, or a specific instance from a set of automatically matched instances is a different entity and, thus, must have a different identity.
- **Vocabulary curation.** Similarly to instance matching, vocabulary curation can be manual (using a user interface), automatic or semi-automatic. Here, the objective is to align equal or related terms by providing a *preferred* and a *broader* (if any) term for each distinct term in a vocabulary (e.g., for a vocabulary of professions: 'capitano' has preferred term 'captain' and has broader term 'sailor'). The FAST CAT system [13] offers a user interface for manual vocabulary curation in the context of maritime history. The cost of curating a singe vocabulary depends on the number of vocabulary terms (if manual) or the effort required for checking and correcting automatically aligned terms. For instance, in the case of the SealiT project (https://sealitproject.eu/ (accessed on 1 July 2022)), there are around 50 vocabularies that need curation, some of which contain thousands of terms (more in [13]).

- **Information extraction from texts.** The cost is analogous to the complexity of the information to be extracted, the effectiveness of the extraction method and the effort required for corrections. A related aspect of data entry from documents, is the classification of the documents according to various taxonomies, vocabularies and thesauri. For instance, [37] presents a method for document subject indexing based both on Topic Modeling and automated labeling processes, aiming to improve the performance of the indexing and the quality of the indexing terms assigned to a document.
- **Query formulation and browsing/exploitation in general.** This cost mainly affects the users of the integrated KG. Multiple access methods should be supported (as in [38] for a keyword search over DBpedia), including keyword search (offering both triple ranking and entity ranking), graph-based browsing, question answering, query templates, etc. There is also the cost for setting up such services. This may require defining competency queries (query templates in general), configuring assistive query building interfaces (such as A-QuB [32] or ResearchSpace [12]) for interactive query formulation, customizing pipelines for QA, etc.

4. Surveying Existing Methods

Here, we survey the existing works; at first, in Section 4.1, we describe the methodology that we have followed for finding approaches that exploit ML techniques for CIDOC-CRM data. In Section 4.2, we analyze the found works by mentioning the correspondence between each work and the tasks presented in Section 3, whereas Section 4.3 provides an analysis of the material found.

4.1. Methodology

Selection Strategy. For finding the related works, we used Google Scholar in the period of June 2021–May 2022 without any restrictions on the publication date. We used the following queries: (i) "CIDOC-CRM Machine Learning", (ii) "CIDOC-CRM Deep Learning", (iii) "CIDOC-CRM word embeddings" and (iv) "CIDOC-CRM neural networks". For each query, we performed a search on Google Scholar for related papers, i.e., we found approximately 1000 relevant papers for the given queries. For each paper, we manually checked its title, abstract and body by strictly keeping only the papers that use machine learning techniques over CIDOC-CRM data. In particular, we care about papers describing one or more processes that use a machine learning technique and the input or/and the output concerns data described using the CIDOC-CRM model.

Statistics. Tables 1 and 2 provide some statistics about the publication years and venues for the surveyed papers, respectively. As we can see, although we did not use such filter, the majority of works that we retrieved concern the last 5 years; i.e., most of them are after 2021 (see the last two rows of Table 1). On the contrary, concerning the publication venues, the most common one is the *Semantic Web* journal (see Table 2).

4.2. Analysis of the Surveyed Works

We categorize the surveyed approaches into several dimensions, including (a) the *Related Tasks* to Section 3.2, and each approach tries to solve using machine learning techniques; and (b) the *Subcategory* (of the related tasks) of each approach and (c) their *Domain*, e.g., archaeology. Moreover, we categorize the approaches according to some technical details, including (d) *Data Category*, i.e., whether the used data are texts, structured data, images, etc.; (e) the *Usage of CIDOC-CRM data*, i.e., whether CIDOC-CRM data are used as an input or/and as an output for the corresponding machine learning task; (f) the *Volume of Data* used in each approach; and finally, (g) the *machine learning tools/algorithms* that were applied for solving the required tasks.

Table 3 provides a synopsis of the categorized works for the dimensions (a)–(c), whereas Table 4 presents the values for dimensions (d)–(g). The works are presented in chronological order in the mentioned tables. Below, we provide more details for each of the surveyed works.

Table 1. Publication Years of Surveyed Works.

Year	Papers	Number per Year
2017	[39]	1
2018	[40,41]	2
2019	[24,42]	2
2021	[43–46]	4
2022	[47,48]	2

Table 2. Publication venues of the Surveyed Works in descending order with respect to the total number of publications.

Publication Venue (Conference, Journal)	Papers	Total Number
Semantic Web Journal	[24,45–47]	4
Information, MDPI	[44]	1
Data and Journal of Visual Languages and Computing	[41]	1
Journal of Information Science	[43]	1
ERCIM News	[39]	1
Proceedings of ECIR Conference	[42]	1
Knowledge Engineering	[48]	1
Proceedings of Digital Heritage International Congress	[40]	1

- TEXTCROWD [39,40] is a cloud based tool (developed within the framework of EOSCpilot project) for processing textual archaeological reports. The general objective is to aid the data entry process by building a system capable of reading excavation reports, recognising relevant archaeological entities and linking them to each other on linguistic bases. TEXTCROWD was initially trained on a set of vocabularies and a corpus of archaeological excavation reports. It offers POS tagging and Named Entity Recognition using two different ML tools, i.e., OpeNER and OpenNLP. TEXTCROWD is able to generate metadata encoding the knowledge extracted from the documents into CIDOC-CRM.
- An approach for aiding the Data Entry process is described in [41], which extracts the entities and relations from Chinese intangible cultural heritage texts and exploits CIDOC-CRM classes for describing the extracted entities (and their relations). In more detail, this paper focuses on knowledge extraction for the domain of intangible cultural heritage (ICH). The authors have created a training corpus and then applied deep learning through a Bidirectional Gated Recurrent Units (GRU) model with attention to extract entities and relations from ICH text data and for finding their corresponding CIDOC-CRM class.
- The authors in [42] proposed an approach for improving data entry by offering event detection and extraction over CIDOC-CRM data, based on directional Long Short-Term Memory (LSTM), which is a type of recurrent neural network. The target of this approach is to create narratives from the extracted data. It has been evaluated for hundreds of events by using a digital library containing narratives for the tasks of event detection and classification. Specifically, comparing to the baseline model,

they observed a small improvement, e.g., for the event detection task, the F1-Score was 0.73 versus 0.66 (baseline).

- WarSampo knowledge graph [24] contains data about the Second World War by focusing on Finnish military history by using CIDOC-CRM (and extensions). Concerning the related Machine Learning task, the resulted data, expressed through CIDOC-CRM, are given as input for performing entity matching (i.e., a subtask belonging to data curation) for thousands of persons between the different datasets by using probabilistic linkage techniques and logistic regression.
- The authors in [43] exploit a cultural knowledge base and machine learning techniques for offering advanced access services, i.e., recommendation of similar items. The Knowledge base is mainly constructed by using CIDOC-CRM ontology and other popular ontologies, such as SKOS, and word embeddings for computing personalised recommendations by taking into account the profile of a museum visitor. The data, which are expressed through CIDOC-CRM, and the details of the personal profile of the user are given as input for providing recommendations, based on embeddings produced using the word2vec model.
- The word embeddings are also used over CIDOC-CRM knowledge graphs in [44] for generating similarity recommendations and they demonstrate its functionality on the Sphaera Dataset, which was modeled according to the CIDOC-CRM data structure. The embeddings have been produced by using Relative Sentence Walk (RSW) and doc2vec model for hundreds of entities.
- In [45], the authors proposed an approach for improving the Data Entry process by performing text classification, extraction and representation for Portuguese National Archives records. The target is the extracted information (from text) to be represented by using CIDOC-CRM ontology and then is visualized by using a Query Ontology Interface. The tool has been evaluated by using 200 texts, and the classifiers have been built through models, such as N-Grams and TF-IDF, by using a decision tree.
- A method for improving data curation is described in [46] by predicting missing metadata in a given knowledge graph by using both image and text analysis. The data model is based on CIDOC-CRM model, and the missing metadata are predicted by using Deep Learning and Convolutional Neural Networks and multi-task learning over thousands of samples. Indicatively, they managed to predict, with an accuracy of over 92%, the class labels of previously unseen images.

Table 3. Overview of the surveyed works applying machine learning techniques over CIDOC-CRM—Dimensions (a)–(c).

Work	Related Task (to Section 3.2)	Subcategory	Domain
[39,40]	Data Entry	POS Tagging and Named Entity Recognition (NER)	Archaeology
[41]	Data Entry	NER and Relation Extraction	Cultural Heritage data of China
[42]	Data Entry, Access Services and Analytics	Information Extraction, Classification and Narratives	Digital Library with Narratives
[24]	Data Curation	Entity Matching	Finnish Military History
[43]	Access Services and Analytics	Context Personalisation and Recommendation	Personalised Cultural Heritage

Table 3. *Cont.*

Work	Related Task (to Section 3.2)	Subcategory	Domain
[44]	Access Services and Analytics.	Knowledge Graph Embeddings	Access services in general
[45]	Data Entry	NER and Relation Extraction	Portuguese National Archives Records
[46]	Data Curation	Predicting missing metadata (mainly about images)	Culture-related objects
[47]	Access Services and Analytics	Question Answering	Genealogical data
[48]	Access Services and Analytics	Document Clustering and Classification	French Historical Data (BNF)

- In another context, i.e., genealogical data, CIDOC-CRM data are used for offering Question Answering [47]. In particular, ti generates text passages from knowledge sub-graphs that contain genealogical data for creating questions and answers and for building a Question Answering system by exploiting deep neural network techniques with the Uncle-BERT model. The process has been trained and tested for millions of entities and events that belong to CIDOC-CRM classes, including Person, Birth and others. They managed to achieve a higher accuracy by using the Uncle-BERT model comparing to BERT; i.e., their F1-score was 0.81 versus 0.6 (BERT).
- Finally, the authors in [48] describe historical documents from France through an extension of CIDOC-CRM and exploit the K-Means algorithm for performing clustering by classifying whether a CIDOC-CRM-based document is either deteriorated or available. The experiments have been performed over 8000 documents, and they identified even 95% accuracy in the test set.

Table 4. Overview of the surveyed works applying machine learning techniques over CIDOC-CRM—Dimensions (d)–(f).

Work	Data Category	How CIDOC-CRM Data Are Used	Volume of Data	Machine Learning Tools/Algorithms
[39,40]	Texts	Output	30 large reports	OpeNER, OpenNLP
[41]	Texts	Output	∼1500 entities	Bidirectional (Gated Recurrent Units) GRU model with attention.
[42]	Texts	Input	∼600 events	Event extraction based on LSTM, a type of recurrent neural network.
[24]	Structured data	Input	94,676 entities	Probabilistic record linkage with a logistic regression based machine learning implementation
[43]	Structured data	Input	∼1,000,000 axioms	word2Vec, through pre-trained Google News corpus
[44]	Structured data	Input	359 entities (book editions)	Relative Sentence Walk (RSW) and doc2vec

Table 4. *Cont.*

Work	Data Category	How CIDOC-CRM Data Are Used	Volume of Data	Machine Learning Tools/Algorithms
[45]	Texts	Output	200 texts	A classifier using a N-Gram and a TF-IDF model for the sample data and a decision tree
[46]	Structured data and Images	Input	∼30,000 samples	Deep Learning and Convolutional Neural Networks (CNNs), and multi-task learning
[47]	Structured data	Input	1,847,224 entities	Deep Neural Networks with Uncle-BERT
[48]	Structured data	Input	8000 documents	K-means algorithm

4.3. Comparative Analysis of the Surveyed Works

First, we can observe, in Figure 4, that the most common tasks concern the exploitation of CIDOC-CRM data for offering more advanced access services and analytics (5 out of 10), especially through the creation of word embeddings and the process of Data Entry (4 out of 10 approaches), mainly for extracting information from texts, which can be expressed through CIDOC-CRM model. On the contrary, 2 out of 10 works provide solutions related to the *data curation* task, i.e., for entity matching and data enrichment (prediction of missing metadata). Moreover, CIDOC-CRM data were used both as input and as output, i.e., 70% of cases as input, as it is shown in Figure 5 for the described machine learning approaches. Finally, most approaches exploit deep learning models, e.g., word embeddings, for improving the desired task.

Concerning the major limitations, the number of approaches is quite low, although we observed an increase in the last two years (2021–2022). Moreover, we did not find approaches for numerous tasks of Section 3.2, including modeling, transformation, evolution and others, whereas most approaches have been evaluated with a small amount of data, i.e., only few approaches used and evaluated for millions of data. However, since we strongly believe that more CIDOC-CRM tasks (and even at large scale) can be assisted using machine learning, in Section 5, we introduce our vision and examples of how CIDOC-CRM could be highly benefited using ML techniques.

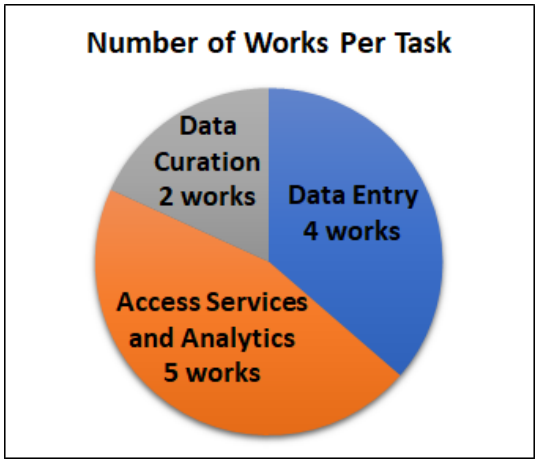

Figure 4. A chart showing the number of works per task.

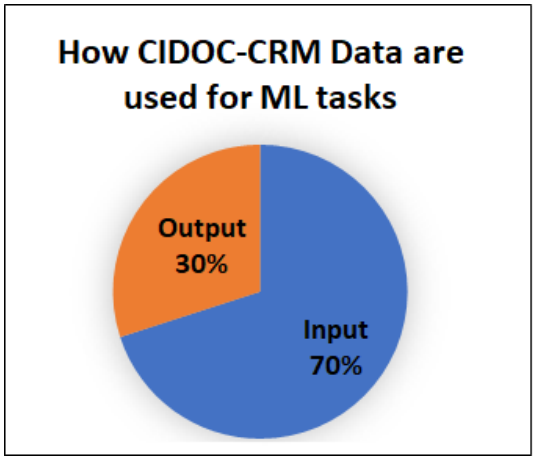

Figure 5. A chart showing how the CIDOC-CRM data are used in the presented works.

5. Vision, Examples and Datasets

Section 5.1 discusses a vision related to CIDOC-CRM and ML, Section 5.2 provides some concrete visionary examples and finally, Section 5.3 provides a list of open datasets expressed in CIDOC-CRM that could be exploited for training ML models.

5.1. Vision (CIDOC-CRM and ML)

Machine learning could facilitate and, thus, speedup various CIDOC-CRM-related tasks and processes. According to our opinion, it could greatly speed up tasks that are related to *unstructured data*. Below, we describe some indicative scenarios.

Data Entry. Since manual data entry can be very time consuming for large scale data; below, we identify tasks (categorized according to the input type) where ML could speed up this process.

- **Text Analysis**. For the automatic identification of actors and mainly *events* as the latter are very important for connecting the aggregated data, a recent survey [49] provides state-of-the-art deep learning methods for named entity recognition and relation extraction.
- **Images Understanding**. For automatically obtaining the descriptions of the form "This painting depicts 3 persons and one church". An interesting work for automatic art analysis is described in [50]. Note that there are several works that detect and identify objects from images: [51] describes the system YOLO that can be also used for detecting objects in moving images due to its fast response, [52] makes use of artificial neural networks to support the identification of objects in images and [53] proposes a fast and more accurate object detector. Finally, we could also mention Detector2 [54], which includes various object detection and segmentation algorithms and has been implemented by Facebook AI Research. Moreover, a recent large collection for visual QA for cultural heritage has been published [55]. It contains a list of question–answer pairs, where each of these pairs is associated to one image, which is derived from the ArCo Knowledge Graph of the Italian Cultural Heritage.
- **Music Classification**: For the automatic classification of music to genre, refer to [56].
- **Video Understanding**: For the automatic identification of persons, topics, etc. A 2008 survey for automatic video classification is [57], whereas a recent survey [58] lists several approaches that use deep learning methods for video understanding tasks, including the automatic generation of descriptions from videos.
- **Classification of cultural objects in general**: There are several works that apply ML for automatically classifying cultural objects. This includes, digital artwork classification [59], pottery types [60,61], ceramic artefacts [62], chronological classification of ancient paintings [4], prediction of painting style [63] and others. In CIDOC-CRM, the classification of a cultural object is represented using the classes `E22:Human-Made`

`Object` and `E28:Conceptual Object`, and the property `P2:has type`, which points to `E55:Type` and can be used for further classifying them. Consequently, all these methods can be leveraged for populating this relation.

Access and Exploitation. One important task is the translation of an information need expressed in *natural language* to a formal query over CIDOC-CRM or answered by a QA pipeline that returns a SPARQL query. Another task is *similarity-based browsing* for tackling the information overload, i.e., for being able to reveal the more important connections and for various kinds of recommendation services. An example is RDFSim [64] that offers similarity-based browsing over DBpedia based on knowledge graph embeddings.

Data Linking. With respect to data linking, ML could be used to link data from different sources. This concerns data enrichment and verification, e.g., [65], automatic labelling, [66], vocabulary curation, and others.

5.2. Examples of Machine Learning Tasks over CIDOC-CRM Data

Here, we describe five indicative examples of exploiting ML for CIDOC-CRM; some of them have already been tested with existing tools.

5.2.1. Example 1. Multi-Source Data Entry

The input data can be expressed in many formats including either unstructured data, such as texts in natural language, semi-structured data [67], such as CSV files, structured data, e.g., in RDF format, or even other files such as images, sounds and videos. Here, we provide an example of a typical input: a CSV file and an image. In particular, the indicative example is shown in Figure 6.

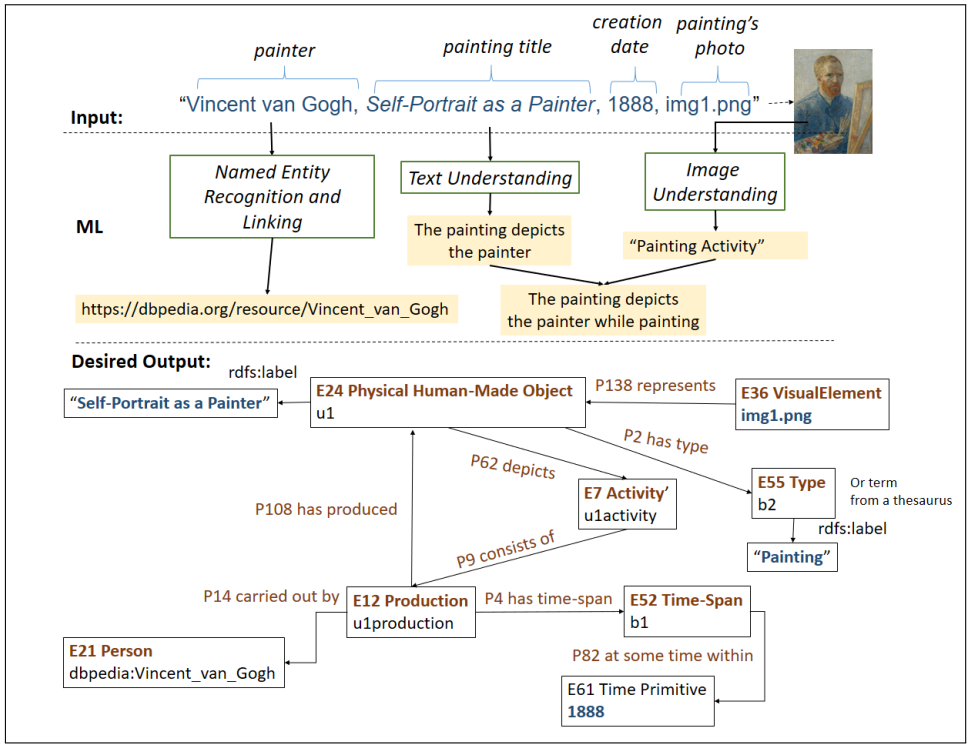

Figure 6. An example of documenting painting using CIDOC-CRM.

At the top, we can see a rather typical input: a photo of a painting and a short description of that painting that includes the painting's painter, title and creation date (As it is found in https://www.vangoghmuseum.nl/en/art-and-stories/stories/5-things-you-need-to-know-about-van-goghs-self-portraits (accessed on 1 July 2022)).

Notice that, with text understanding from the word "self-portrait" in the title, we could deduce that the painting "depicts the painter". By image understanding, we cab infer that the image refers to a "painting activity". In particular, Figure 7 depicts a screenshot of the real output of an existing tool (https://huggingface.co/spaces/sohaibcs1/Image-to-Text-Summary (accessed on 1 July 2022)) for that portrait. As we can see, a summary is provided, which mentions that the portrait shows a painting where a man has a brush in his hand.

By combining the above facts, we could conclude that "the painting depicts the painter while painting". Now at the bottom of Figure 6, we can see a possible modeling of this information according to CIDOC-CRM (We use an UML Object Diagrams-like notation: We denote individuals by boxes with two rows: the first has the class of the individual, the second the URI or blank node of the individual). Notice that the fact inferred from image understanding is modeled as an *activity* (connected with P62). The fact that this activity depicts the painter while painting enables us to state that painting activity contains the creation event of that painting (i.e., we connect them through P9). The latter enables answering questions of the form "paintings that depict their creation", "paintings that depict painting activities" and so on.

The following block shows the structured form (in XML) of the short narrative from Figure 6, which can be used as input, in order to describe the schema mappings for generating the corresponding CIDOC-CRM-related instances. The schema mappings have been described using X3ML.

```
<root>
  <painting>
    <creator>Vincent van Gogh</creator>
    <title>Self-Portrait as a Painter</title>
    <creation_date>1888</creation_date>
    <filename>img1.png</filename>
  </painting>
</root>
```

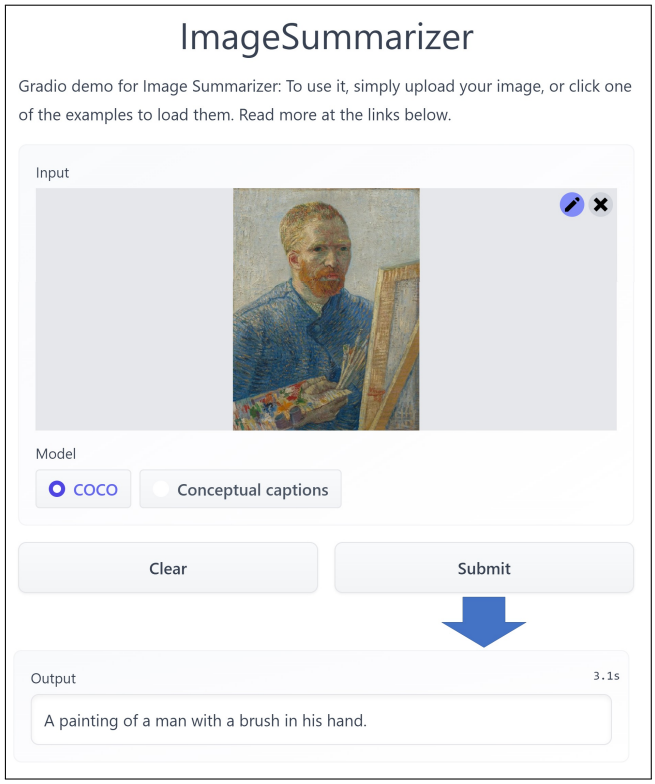

Figure 7. Example: Producing text from a given Image—machine learning model. Real Screenshot from the tool ImageSummarizer.

The XML input and the X3ML mappings can be used for producing the instances with respect to CIDOC-CRM. The details of the X3ML mappings are shown later in Section 5.2.3. The following block shows the output in Turtle format.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix crm-ex: <http://www.cidoc-crm.org/examples/> .
@prefix dbpedia: <https://dbpedia.org/resource/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

crm-ex:visual_item/E98F243A-B982-4142-83AE-88FE0BD058CE
        a crm:E36_Visual_Item ;
        rdfs:label ''Digital representation of painting Self-Portrait as a Painter'';
        crm:P1_is_identified_by  crm-ex:identifier/img1.png .

crm-ex:identifier/img1.png
        a crm:E42_Identifier ;
        rdfs:label  ''img1.png'' .

dbpedia:Vincent_van_Gogh
        a crm:E21_Person ;
        rdfs:label  ''Vincent van Gogh'' .

crm-ex:painting/F16671D1-7D69-354E-91E8-CDAD152A77B0
        a  crm:E22_Human-Made_Object ;
        rdfs:label
          ''Painting Self-Portrait as a Painter'' ;
        crm:P102_has_title
          crm-ex:title/F16671D1-7D69-354E-91E8-CDAD152A77B0 ;
        crm:P108i_was_produced_by
          crm-ex:production/748525A0-2E27-4274-A1DE-71F29ECB56DE ;
        crm:P138i_has_representation
          crm-ex:visual_item/E98F243A-B982-4142-83AE-88FE0BD058CE ;
        crm:P2_has_type
          crm-ex:object_type/painting .

crm-ex:object_type/painting
        a crm:E55_Type ;
        rdfs:label  ''Painting'' .

crm-ex:title/F16671D1-7D69-354E-91E8-CDAD152A77B0
        a crm:E35_Title ;
        rdfs:label  ''Self-Portrait as a Painter'' .

crm-ex:timespan/F61F62D8-A608-4B8F-AE02-F0607FBA2D93
        a crm:E52_Time-Span ;
        crm:P82_at_some_time_within  ''1888'' .

crm-ex:production/748525A0-2E27-4274-A1DE-71F29ECB56DE
        a  crm:E12_Production ;
        rdfs:label ''Creation of painting Self-Portrait as a Painter'' ;
        crm:P14_carried_out_by dbpedia:Vincent_van_Gogh ;
        crm:P4_has_time-span
          crm-ex:timespan/F61F62D8-A608-4B8F-AE02-F0607FBA2D93 .
```

5.2.2. Example 2. Text Analysis (with Emphasis on Event Detection)

Here, we provide an example of having a longer (in comparison to Example 1) text as input. In particular, consider the following text (derived from Wikipedia): *"Vincent van Gogh (30 March 1853–29 July 1890) was a Dutch post-Impressionist painter who posthumously became one of the most famous and influential figures in Western art history."* Text analysis can be used to extract information and express it according to CIDOC-CRM. For instance, from the above description, we can extract the birth and death date of the painter and his style. Indeed, by using huggingface (https://huggingface.co/ (accessed on 1 July 2022)), we can extract this information and then represent it using CIDOC-CRM, as shown in the following block. For creating the mentioned output, post-processing tasks are required,

including the creation of mappings (which can be quite complex and time consuming, as it is described in Example 3 in Section 5.2.3). However, a challenge is how to detect, extract and name various kinds of events. Related work that identifies events (and the constituents of events: actors, etc.), includes [68,69], whereas a pipeline for converting a text describing cultural data to RDF is described in [70].

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix crm-ex: <http://www.cidoc-crm.org/examples/> .
@prefix dbpedia: <https://dbpedia.org/resource/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

dbpedia:Vincent_van_Gogh
        a crm:E21_Person ;
        rdfs:label ‘‘Vincent van Gogh’’ ;
        crm:P2_has_type crm-ex:painter-post-impressionist .

crm-ex:painter-post-impressionist
       a crm:E55_Type ;
       rdfs:label ‘‘Post-Impressionist Painter’’ .

crm-ex:birthOfVvG
        a crm:E67_Birth ;
        rdfs:label ‘‘Birth of Vincent van Gogh’’ ;
        crm:P98_brought_into_life dbpedia:Vincent_van_Gogh ;
        crm:P7_took_place_at dbpedia:Netherlands ;
        crm:P4_has_timespan crm-ex:birthTimeSpan .

dbpedia:Netherlands
        a crm:E53_Place ;
        rdfs:label ‘‘Netherlands’’ .

crm-ex:birthTimeSpan
        a crm:E52_Time-Span ;
        crm:P82_at_some_time_within ‘‘1863’’^^xsd:gYear .

crm-ex:deathOfVvG
        a crm:E69_Death ;
        rdfs:label ‘‘Death of Vincent van Gogh’’ ;
        crm:P100_was_death_of dbpedia:Vincent_van_Gogh ;
        crm:P4_has_timespan crm-ex:deathTimeSpan .

crm-ex:deathTimeSpan
        a crm:E52_Time-Span ;
        crm:P82_at_some_time_within ‘‘1890"^^xsd:gYear .
```

5.2.3. Example 3. Mapping Process

The mapping process can include both mappings of ontologies and instances among different data sources. This process aims at identifying which parts from the input will be mapped to particular classes and properties of CIDOC-CRM. This is a rather complex task that requires good knowledge of the involved schemata. In the sequel, we show the X3ML mappings (specifically their representation in XML) that are used for transforming the input from Example 1 (of Section 5.2.1). It is only one part of the overall schema mappings (The complete version of the schema mappings and the XML input can be found at https://github.com/isl/CIDOC-CRM-datasets (CIDOC-CRM and Machine Learning/Example Dataset/) (accessed on 1 July 2022)), showing how the painting and its title from the XML input are mapped to the corresponding classes and properties (`crm:E22_Human-Made_Object` and `crm:E35_Title`, respectively). Apart from the definition of the mappings, we have also described how the URIs will be generated (i.e., using `instance_generator` and `value_generator` definitions).

```
<mapping>
  <domain>
    <source_node>/root/painting</source_node>
    <target_node>
      <entity>
        <type>crm:E22_Human-Made_Object</type>
        <instance_generator name="LocalTermURI">
          <arg name=‘‘hierarchy’’ type=‘‘constant’’>panting</arg>
          <arg name=‘‘term’’ type=‘‘xpath’’>title/text()</arg>
        </instance_generator>
        <label_generator name=’’CompositeLabel’’>
          <arg name=‘‘label_part1’’ type=‘‘constant’’>Painting</arg>
          <arg name=‘‘label_part2’’ type=‘‘xpath’’>title/text()</arg>
        </label_generator>
          <additional>
            <relationship>crm:P2_has_type</relationship>
              <entity>
                <type>crm:E55_Type</type>
                  <instance_generator name=‘‘LocalTermURI’’>
                    <arg name=‘‘hierarchy’’ type=‘‘constant’’>object_type</arg>
                    <arg name=‘‘term’’ type=‘‘constant’’>painting</arg>
                  </instance_generator>
                  <label_generator name=‘‘SimpleLabel’’>
                    <arg name=‘‘label’’ type=‘‘constant’’>Painting</arg>
                  </label_generator>
                </entity>
          </additional>
        </entity>
      </target_node>
  </domain>
  <link>
    <path>
      <source_relation>
        <relation>title</relation>
      </source_relation>
      <target_relation>
        <relationship>crm:P102_has_title</relationship>
      </target_relation>
    </path>
    <range>
      <source_node>title</source_node>
      <target_node>
        <entity>
          <type>crm:E35_Title</type>
          <instance_generator name=‘‘LocalTermURI’’>
            <arg name=‘‘hierarchy’’ type=‘‘constant’’>title</arg>
            <arg name=‘‘term’’ type=‘‘xpath’’>text()</arg>
          </instance_generator>
          <label_generator name=‘‘SimpleLabel’’>
            <arg name=‘‘label’’ type=‘‘xpath’’>text()</arg>
          </label_generator>
        </entity>
      </target_node>
    </range>
  </link>
</mapping>
```

Although such tasks can be performed through instance and schema matching tools that are based on predefined rules [11], there is a trend for machine learning-based algorithms based on embeddings for solving such tasks, e.g., [71]. A possible approach could be to exploit past manually created mappings as training data for learning mapping patterns using machine learning algorithms and then suggesting future data mapping rules. Mapping such as the one that is provided above could also be used to propose such mappings. More specifically, it could train a model to construct a mapping that creates instances of `crm:E22_Human-Made_Object` linked with a proper type (`crm:P2_has_type` ->

`crm:E55_Type`) for elements in the input with name Painting. Moreover, in case that the input is given in a structured format, e.g., XML, neural networks solutions can be applied for transforming the data to RDF, e.g., [72].

5.2.4. Example 4. Question Answering

The user formulates a question in natural language, and the objective is to either produce a SPARQL query or directly provide an answer, as in a classical QA task. This can complement the other access methods (SPARQL, Fundamental Categories, Plain and similarity-based browsing). Consider for instance, a CIDOC-CRM-compliant knowledge graph describing data about popular painters and that the graph includes the triples of the desired output in the example of Figure 6. The natural language question *What is the creation date of Van Gogh's painting 'Self-Portrait as a Painter'* can be transformed to the following SPARQL query that provides the correct answer.

```
SELECT ?creationDate
WHERE {
  ?production crm:P14_carried_out_by dbpedia:Vincent_van_Gogh ;
              crm:P108_has_produced ?painting ;
              crm:P4_has_time-span/crm:P82_at_some_time_within ?creationDate .
  ?painting rdfs:label ?label
      FILTER (REGEX (STR(?label), ''self-portrait as a painter'', ''i'')) }
```

To the best of our knowledge, no system can produce that query since it is not trivial. In particular, it requires techniques of natural language Processing, such as Named Entity Recognition, e.g., for detecting the entity Van Gogh and finding its DBpedia link, and relation extraction, e.g., for finding the relevant CIDOC-CRM properties that can be used for answering the query.

However note that transforming to SPARQL is not a panacea, since, for answering some questions, one has to extract the answers from the literals. Therefore, instead of transforming a natural language question to am SPARQL query, one can apply various QA pipelines (see [73] for a survey), e.g., an alternative method is to transform the resulting CIDOC-CRM graph to natural text and to use existing BERT-based models for answering the question. Returning to our example, the approach presented in [74] (that relies on keyword search, SPARQL and pre-trained neural networks) can answer the above question (over the DBpedia dataset). In general, the more complex the question is the harder it is to answer it, especially in cases where one has to exploit various deductions from the knowledge graph. For this purpose, an interesting direction that is worth researching is to build a QA pipeline over the query model of the fundamental categories and relationships [31] (described earlier in Section 3.2).

5.2.5. Example 5. Knowledge Graph Embeddings in Data Access and Exploration

In many cases, the users desire to browse similar things, e.g., similar painters or paintings, and usually similarity methods are required that proceed beyond manually designed similarity functions.. For example, by searching for related entities to the Painter El Greco, we expect to find his paintings, similar paintings from other painters or similar painters. It would also be desirable to discover as a similar entity the Italian mathematician Fransesco Barozzi, since these two persons that lived in the 16th century were of primary importance for the Kingdom of Candia (the official name of Crete from 1205 to 1667). For making it feasible to discover such similarities and also to reveal the most important connections for any Actor or Event, knowledge graph embeddings and similarity-based models can be used, such as Word2Vec, Doc2Vec, BERT and others. An example of similarity-based browsing over DBpedia using the tool RDFSim is described in [64] and a real screenshot of that tool is provided in Figure 8, where we can see relevant entities relative to Pablo Picasso, such as his famous painting Guernica; the french poet Guillaume Apollinaire, who was a friend of Picasso; painters, such as Renoir; and the novel "*In Search of Lost Time*", which was written from Proust at the period when Picasso was living in Paris.

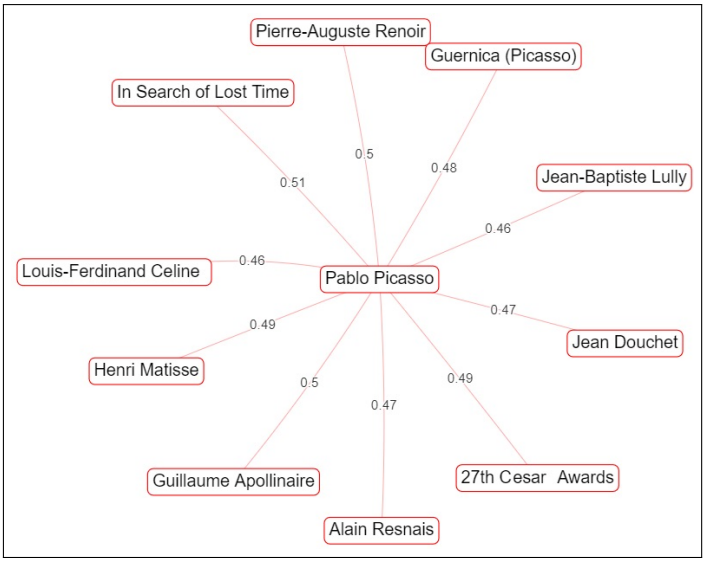

Figure 8. Example: A screenshot with relevant entities to Pablo Picasso by using RDFsim [64].

5.3. Available Datasets Expressed through the CIDOC-CRM Model

For applying Machine Learning to CIDOC-CRM data, it would be useful to test and evaluate machine learning methods over real datasets that have been created through that model. For this reason, in Table 5, we provide a list of 18 datasets that have been expressed using the CIDOC-CRM model, and their data can be derived online through an API (e.g., SPARQL service) or a data dump. Most datasets were found through the official webpage of CIDOC-CRM and particularly via the *CRM Community Activity Documentation* (https://www.cidoc-crm.org/useCasesPage (accessed on 1 July 2022)), whereas some additional datasets were found by searching through Google Dataset Search (https://datasetsearch.research.google.com/ (accessed on 1 July 2022)).

Table 5 shows only the datasets that are accessible online (i.e., all links of Table 5 accessed on 1 July 2022), since there are several datasets that were published in the past and their corresponding websites are currently down. For instance, some of these datasets were mentioned in [30], and although they were quite connected with several other RDF datasets, they are not accessible anymore. Moreover, some datasets are not publicly available, e.g., due to privacy reasons. As we can see, most datasets offer a SPARQL endpoint or another API for accessing/querying the data, whereas in many cases, a data dump is provided. Moreover, most datasets offer millions of triples and they cover various disciplines of cultural heritage data, including music, historical data, museums and others. Finally, we should also mention attempts to develop crowd-sourcing approaches that are CIDOC-CRM compatible, such as [75].

Given the increasing number of available datasets expressed through CIDOC-CRM, it would be interesting to create a centralized service for storing and for finding fast all the available datasets expressed by CIDOC-CRM, or at least a metadata-based service providing rich metadata for each dataset. For the time being, for making it feasible to i) keep track of any updates on the existing datasets and ii) add new datasets in the future, we have created a github repository (https://github.com/isl/CIDOC-CRM-datasets, accessed on 1 July 2022), which contains the up-to-date list of available CIDOC-CRM datasets. This would be also beneficial for the preservation and maintenance of datasets, since many of them are not available after a period of time [76]. Finally, it would be quite helpful to create open CIDOC-CRM datasets that could be used for training and comparative evaluation, e.g., either subsets of real datasets or/and synthetic ones for covering complex cases.

Table 5. Online datasets expressed through CIDOC-CRM (alphabetical order).

ID	Dataset	Link	Domain	Number of Triples	SPARQL End-point/API	Data Dump
1	Archaeology Data Service	http://data.archaeologydataservice.ac.uk (accessed on 1 June 2022)	Heritage Data of United Kingdom	1,559,912	✓	
2	Auckland Museum	https://api.aucklandmuseum.com/ (accessed on 1 June 2022)	Auckland Museum, New Zealand	>10,000,000	✓	
3	Beni Culturali	https://dati.cultura.gov.it/linked-open-data/ (accessed on 1 June 2022)	Cultural Institutions in Italy	755,702,389	✓	✓
4	Corago LOD	https://zenodo.org/record/3377586 (accessed on 1 June 2022)	Italian Opera, 1600 to 1900	22,399,698		✓
5	Cultura Italia	https://dati.culturaitalia.it/ (accessed on 1 June 2022)	Italian Painting, Painters, Sounds and Videos	41,901,551	✓	
6	Doremus	https://data.doremus.org/ (accessed on 1 June 2022)	World Classical Music	91,093,377	✓	
7	Foundation Zeri	http://data.fondazionezeri.unibo.it/ (accessed on 1 June 2022)	Photography and Italian Painters	11,827,416	✓	✓
8	Joconde Database	https://zenodo.org/record/3986498 (accessed on 1 June 2022)	French cultural heritage	∼11,000,000		✓
9	Kerameikos	http://kerameikos.org (accessed on 1 June 2022)	Ceramics of Ancient Greece	289,590	✓	✓
10	Nomisma.org	https://nomisma.org/ (accessed on 1 June 2022)	Numismatic concepts	9,933,870	✓	✓
11	OEBL	https://zenodo.org/record/3873203 (accessed on 1 June 2022)	Austrian Biographical Dictionary	∼600,000		✓
12	OpenArchaeo	http://openarchaeo.huma-num.fr/explorateur/home (accessed on 1 June 2022)	Platform for archaeological data	1,548,827	✓	
13	Persons and Names of the Middle Kingdom	https://pnm.uni-mainz.de/ (accessed on 1 June 2022)	Egyptian Middle Kingdom Persons and Names	1,490,284	✓	
14	RePIM	https://zenodo.org/record/5692109 (accessed on 1 June 2022)	Italian Music, 1500–1700	4,427,647		✓
15	SeaLiT Knowledge Graphs	https://zenodo.org/record/6460841 (accessed on 1 June 2022)	Maritime History, 1850s–1920s	∼18,500,000		✓
16	Smithsonian Museum	https://triplydb.com/smithsonian/-/overview (accessed on 1 June 2022)	Smithsonian American Art Museum	2,802,768	✓	✓
17	WarSampoo	https://seco.cs.aalto.fi/projects/sotasampo/en/ (accessed on 1 June 2022)	Finnish World War II	14,322,426	✓	✓
18	WW1LOD	https://www.ldf.fi/dataset/ww1lod/ (accessed on 1 June 2022)	Finnish World War I	47,616	✓	✓

How to Apply ML. In the context of scientific documentation, where precision is important, ML-based tasks can be used to assist the humans and not to replace them. Therefore, for applying effectively ML in a professional context, implementing workflow systems that allow the users to inspect the output of such tools and to approve, improve or reject the outputs is required to [77,78].

6. Concluding Remarks

For reducing the effort that is required for achieving semantic interoperability through CIDOC-CRM across museums, libraries, archives and other cultural institutions, in this paper, we investigated how recent advances of machine learning can be beneficial. At first, we identified the processes and tasks related to CIDOC-CRM-based data management and discussed their cost. Then, we surveyed the literature on applying ML and CIDOC-CRM. We identified 10 in number of works, where most of them focus on providing access services over CIDOC-CRM data and for improving data entry. Subsequently, we described a vision and provided concrete examples that showcase how ML could reduce the effort of various CIDOC-CRM processes. To facilitate the investigation of ML techniques, we collected a list

of 18 CIDOC-CRM datasets that are available through APIs (e.g., SPARQL endpoints) or data dumps, which can be used for training purposes, and we have created an open github repository (https://github.com/isl/CIDOC-CRM-datasets (accessed on 1 July 2022)) for updating the list in the future. Finally, we have identified several directions that are worth exploring for research and experimentation.

Author Contributions: Conceptualization, Y.T., M.M., P.F. and Y.M.; methodology, Y.T., M.M, P.F. and Y.M.; formal analysis, Y.T., M.M., P.F. and Y.M.; investigation, Y.T., M.M., P.F. and Y.M.; writing—original draft preparation, Y.T., M.M., P.F. and Y.M.; writing—review and editing, Y.T., M.M., P.F. and Y.M.; supervision, Y.T.; project administration, Y.T; funding acquisition, Y.T. and P.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work has received funding (a) from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 890861 (Project ReKnow), and (b) from the European Union's Horizon 2020 coordination and support action 4CH (Grant agreement No 101004468).

Data Availability Statement: Not applicable.

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

References

1. Doan, A.; Madhavan, J.; Domingos, P.; Halevy, A. Ontology matching: A machine learning approach. In *Handbook on Ontologies*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 385–403.
2. Tiddi, I.; Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **2022**, *302*, 103627. [CrossRef]
3. Corrêa, A.S.; Zander, P.O. Unleashing tabular content to open data: A survey on pdf table extraction methods and tools. In Proceedings of the 18th Annual International Conference on Digital Government Research, Staten Island, NY, USA, 7–9 June 2017; pp. 54–63.
4. Chen, L.; Chen, J.; Zou, Q.; Huang, K.; Li, Q. Multi-view feature combination for ancient paintings chronological classification. *J. Comput. Cult. Herit. (JOCCH)* **2017**, *10*, 1–15. [CrossRef]
5. Hyvönen, E. Using the Semantic Web in Digital Humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semant. Web* **2020**, *11*, 187–193. [CrossRef]
6. Doerr, M. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Mag.* **2003**, *24*, 75.
7. Moraitou, E.; Aliprantis, J.; Christodoulou, Y.; Teneketzis, A.; Caridakis, G. Semantic Bridging of Cultural Heritage Disciplines and Tasks. *Heritage* **2019**, *2*, 611–630. [CrossRef]
8. Fiorucci, M.; Khoroshiltseva, M.; Pontil, M.; Traviglia, A.; Del Bue, A.; James, S. Machine learning for cultural heritage: A survey. *Pattern Recognit. Lett.* **2020**, *133*, 102–108. [CrossRef]
9. Bickler, S.H. Machine Learning Arrives in Archaeology. *Adv. Archaeol. Pract.* **2021**, *9*, 186–191. [CrossRef]
10. Rapti, A.; Tsolis, D.; Sioutas, S.; Tsakalidis, A. A survey: Mining linked cultural heritage data. In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS), Rhodes Island, Greece, 25–28 September 2015; pp. 1–6.
11. Mountantonakis, M.; Tzitzikas, Y. Large-scale semantic integration of linked data: A survey. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–40. [CrossRef]
12. Oldman, D.; Tanase, D. Reshaping the Knowledge Graph by connecting researchers, data and practices in ResearchSpace. In Proceedings of the International Semantic Web Conference, Monterey, CA, USA, 8–12 October 2018; pp. 325–340.
13. Fafalios, P.; Petrakis, K.; Samaritakis, G.; Doerr, K.; Kritsotaki, A.; Tzitzikas, Y.; Doerr, M. FAST CAT: Collaborative Data Entry and Curation for Semantic Interoperability in Digital Humanities. *J. Comput. Cult. Herit. (JOCCH)* **2021**, *14*, 1–20. [CrossRef]
14. Petrakis, K.; Samaritakis, G.; Kalesios, T.; i Domingo, E.G.; Delis, A.; Tzitzikas, Y.; Doerr, M.; Fafalios, P. Digitizing, Curating and Visualizing Archival Sources of Maritime History: The case of ship logbooks of the nineteenth and twentieth centuries. *Drassana Rev. Del Mus. Marít.* **2020**, *28*, 60–87. [CrossRef]
15. Fafalios, P.; Konsolaki, K.; Charami, L.; Petrakis, K.; Paterakis, M.; Angelakis, D.; Tzitzikas, Y.; Bekiari, C.; Doerr, M. Towards Semantic Interoperability in Historical Research: Documenting Research Data and Knowledge with Synthesis. In Proceedings of the International Semantic Web Conference, Virtual Event, 24–28 October 2021; pp. 682–698.
16. Varagnolo, D.; Melo, D.; Rodrigues, I.P. A Tool to Explore the Population of a CIDOC-CRM Ontology. *Procedia Comput. Sci.* **2021**, *192*, 158–167. [CrossRef]
17. Theodoridou, M.; Tzitzikas, Y.; Doerr, M.; Marketakis, Y.; Melessanakis, V. Modeling and querying provenance by extending CIDOC CRM. *Distrib. Parallel Databases* **2010**, *27*, 169–210. [CrossRef]
18. Niccolucci, F. Documenting archaeological science with CIDOC CRM. *Int. J. Digit. Libr.* **2017**, *18*, 223–231. [CrossRef]

19. Hiebel, G.; Doerr, M.; Eide, Ø. CRMgeo: A spatiotemporal extension of CIDOC-CRM. *Int. J. Digit. Libr.* **2017**, *18*, 271–279. [CrossRef]
20. Vassilakaki, E.; Zervos, S.; Giannakopoulos, G. CIDOC-CRM extensions for conservation processes: A methodological approach. In Proceedings of the AIP Conference Proceedings, Virtual, 23–29 September 2015; Volume 1644, pp. 185–192.
21. Kritsotaki, A.; Fafalios, P.; Doerr, M. SeaLiT Ontology—An Extension of CIDOC-CRM for the Modelling of Maritime History Information, 2022. Available online https://doi.org/10.5281/zenodo.5964240 (accessed on 1 July 2022).
22. Marketakis, Y.; Minadakis, N.; Kondylakis, H.; Konsolaki, K.; Samaritakis, G.; Theodoridou, M.; Flouris, G.; Doerr, M. X3ML mapping framework for information integration in cultural heritage and beyond. *Int. J. Digit. Libr.* **2017**, *18*, 301–319. [CrossRef]
23. Araújo, C.; Martini, R.G.; Henriques, P.R.; Almeida, J.J. Annotated documents and expanded CIDOC-CRM ontology in the automatic construction of a virtual museum. In *Developments and Advances in Intelligent Systems and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 91–110.
24. Koho, M.; Ikkala, E.; Leskinen, P.; Tamper, M.; Tuominen, J.; Hyvönen, E. WarSampo knowledge graph: Finland in the second world war as linked open data. *Semant. Web* **2019**, 1–14. [CrossRef]
25. Faraj, G.; Micsik, A. Representing and Validating Cultural Heritage Knowledge Graphs in CIDOC-CRM Ontology. *Future Internet* **2021**, *13*, 277. [CrossRef]
26. Capuano, N.; Gaeta, A.; Guarino, G.; Miranda, S.; Tomasiello, S. Enhancing augmented reality with cognitive and knowledge perspectives: A case study in museum exhibitions. *Behav. Inf. Technol.* **2016**, *35*, 968–979. [CrossRef]
27. Mäkelä, E.; Törnroos, J.; Lindquist, T.; Hyvönen, E. WW1LOD: An application of CIDOC-CRM to World War 1 linked data. *Int. J. Digit. Libr.* **2017**, *18*, 333–343. [CrossRef]
28. Mountantonakis, M.; Tzitzikas, Y. How your Cultural Dataset is Connected to the Rest Linked Open Data? In Proceedings of the TMM-CH2021 (Transdisciplinary Multispectral Modelling and Cooperation for the Preservation of Cultural Heritage), Communications in Computer and Information Science, Athens, Greece, 12–15 December 2021; Springer Press: Athens, Greece, 2021.
29. Heath, T.; Bizer, C. Linked data: Evolving the web into a global data space. In *Synthesis Lectures on the Semantic Web: Theory and Technology*; Morgan & Claypool: San Rafael, CA, USA, 2011; pp. 1–136.
30. Mountantonakis, M.; Tzitzikas, Y. LODsyndesis: Global scale knowledge services. *Heritage* **2018**, *1*, 335–348. [CrossRef]
31. Tzompanaki, K.; Doerr, M. *Fundamental Categories and Relationships for Intuitive Querying CIDOC-CRM based Repositories*; ICS-FORTH Technical Report; Institute of Computer Science: Heraklion, Greece, 2012; p. TR-429.
32. Kritsotakis, V.; Roussakis, Y.; Patkos, T.; Theodoridou, M. Assistive Query Building for Semantic Data. In Proceedings of the SEMANTICS Posters&Demos, Vienna, Austria, 10–13 September 2018.
33. Cuteri, B.; Reale, K.; Ricca, F. A logic-based question answering system for cultural heritage. In *Proceedings of the European Conference on Logics in Artificial Intelligence*, Rende, Italy, 7–11 May 2019; pp. 526–541.
34. Meghini, C.; Bartalesi, V.; Metilli, D. Representing narratives in digital libraries: The narrative ontology. *Semant. Web* **2021**, *12*, 241–264. [CrossRef]
35. Tzitzikas, Y.; Kampouraki, M.; Analyti, A. Curating the specificity of ontological descriptions under ontology evolution. *J. Data Semant.* **2014**, *3*, 75–106. [CrossRef]
36. Marketakis, Y.; Tzitzikas, Y.; Gentile, A.; Niekerk, B.V.; Taconet, M. On the Evolution of Semantic Warehouses: The Case of Global Record of Stocks and Fisheries. In Proceedings of the Research Conference on Metadata and Semantics Research, Madrid, Spain, 2–4 December 2020; pp. 269–281.
37. Sfakakis, M.; Papachristopoulos, L.; Zoutsou, K.; Tsakonas, G.; Papatheodorou, C. Automated Subject Indexing of Domain Specific Collections Using Word Embeddings and General Purpose Thesauri. In Proceedings of the Research Conference on Metadata and Semantics Research, Rome, Italy, 28–31 October 2019; pp. 103–114.
38. Nikas, C.; Kadilierakis, G.; Fafalios, P.; Tzitzikas, Y. Keyword Search over RDF: Is a Single Perspective Enough? *Big Data Cogn. Comput.* **2020**, *4*, 22. [CrossRef]
39. Felicetti, A. Teaching archaeology to machines: Extracting semantic knowledge from free text excavation reports. *ERCIM News* **2017**, *111*, 9–10.
40. Felicetti, A.; Williams, D.; Galluccio, I.; Tudhope, D.; Niccolucci, F. NLP tools for knowledge extraction from Italian archaeological free text. In Proceedings of the 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018), San Francisco, CA, USA, 26–30 October 2018; pp. 1–8.
41. Dou, J.; Qin, J.; Jin, Z.; Li, Z. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage. *J. Vis. Lang. Comput.* **2018**, *48*, 19–28. [CrossRef]
42. Metilli, D.; Bartalesi, V.; Meghini, C. Steps Towards a System to Extract Formal Narratives from Text. In Proceedings of the Text2Story@ ECIR, Cologne, Germany, 14 April 2019; pp. 53–61.
43. Dahroug, A.; Vlachidis, A.; Liapis, A.; Bikakis, A.; Lopez-Nores, M.; Sacco, O.; Pazos-Arias, J.J. Using dates as contextual information for personalised cultural heritage experiences. *J. Inf. Sci.* **2021**, *47*, 82–100. [CrossRef]
44. El-Hajj, H.; Valleriani, M. CIDOC2VEC: Extracting Information from Atomized CIDOC-CRM Humanities Knowledge Graphs. *Information* **2021**, *12*, 503. [CrossRef]
45. Melo, D.; Rodrigues, I.P.; Varagnolo, D. A strategy for archives metadata representation on CIDOC-CRM and knowledge discovery. *Semant. Web* **2021**, *1*, 1–32. [CrossRef]

46. Schleider, T.; Troncy, R.; Gaitán, 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*, 1–16.
47. Suissa, O.; Zhitomirsky-Geffet, M.; Elmalech, A. Question answering with deep neural networks for semi-structured heterogeneous genealogical knowledge graphs. *Semant. Web*, **2022**, Preprint, 1–29. [CrossRef]
48. Zreik, A.; Kedad, Z. Matching and analysing conservation–restoration trajectories. *Data Knowl. Eng.* **2022**, *139*, 102015. [CrossRef]
49. Nasar, Z.; Jaffry, S.W.; Malik, M.K. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–39. [CrossRef]
50. Garcia, N.; Renoust, B.; Nakashima, Y. Context-aware embeddings for automatic art analysis. In Proceedings of the International Conference on Multimedia Retrieval, Ottawa, ON, Canada, 10–13 June 2019; pp. 25–33.
51. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
52. Matuszewski, J.; Rajkowski, A. The use of machine learning algorithms for image recognition. In *Proceedings of the Radioelectronic Systems Conference 2019*; International Society for Optics and Photonics: Bellingham, WA, USA, 2020; Volume 11442, p. 1144218.
53. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
54. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 1 July 2022).
55. Asprino, L.; Bulla, L.; Marinucci, L.; Mongiovì, M.; Presutti, V. A Large Visual Question Answering Dataset for Cultural Heritage. In Proceedings of the International Conference on Machine Learning, Optimization, and Data Science, Grasmere, UK, 4–8 October 2021; pp. 193–197.
56. Lau, D.S.; Ajoodha, R. Music Genre Classification: A Comparative Study Between Deep Learning and Traditional Machine Learning Approaches. In Proceedings of the 6th International Congress on Information and Communication Technology, Tallinn, Estonia, 4–6 May 2022; pp. 239–247.
57. Brezeale, D.; Cook, D.J. Automatic video classification: A survey of the literature. *IEEE Trans. Syst. Man, Cybern. Part C (Appl. Rev.)* **2008**, *38*, 416–430. [CrossRef]
58. Aafaq, N.; Mian, A.; Liu, W.; Gilani, S.Z.; Shah, M. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–37. [CrossRef]
59. Sabatelli, M.; Kestemont, M.; Daelemans, W.; Geurts, P. Deep transfer learning for art classification problems. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
60. Charalambous, E.; Dikomitou-Eliadou, M.; Milis, G.M.; Mitsis, G.; Eliades, D.G. An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relationships. *J. Archaeol. Sci. Rep.* **2016**, *7*, 465–471. [CrossRef]
61. Rivero, D.G.; Núñez, J.M.J.; Taylor, R. Bell Beaker and the evolution of resource management strategies in the southwest of the Iberian Peninsula. *J. Archaeol. Sci.* **2016**, *72*, 10–24. [CrossRef]
62. López-García, P.; Argote-Espino, D.; Fačevicová, K. Statistical processing of compositional data. The case of ceramic samples from the archaeological site of Xalasco, Tlaxcala, Mexico. *J. Archaeol. Sci. Rep.* **2018**, *19*, 100–114. [CrossRef]
63. Wilber, M.J.; Fang, C.; Jin, H.; Hertzmann, A.; Collomosse, J.; Belongie, S. Bam! The behance artistic media dataset for recognition beyond photography. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1202–1211.
64. Chatzakis, M.; Mountantonakis, M.; Tzitzikas, Y. RDFsim: Similarity-Based Browsing over DBpedia Using Embeddings. *Information* **2021**, *12*, 440. [CrossRef]
65. Piché, D.; Zouaq, A.; Gagnon, M.; Font, L. Masked Language Model Entity Matching for Cultural Heritage Data. In Proceedings of the International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage Co-Located with the Bolzano Summer of Knowledge 2021 (BOSK 2021), Virtual Event, 20–21 September 2021; Volume 2949.
66. Alokaili, A.; Aletras, N.; Stevenson, M. Automatic generation of topic labels. In Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 25–30 July 2020; pp. 1965–1968.
67. Ryen, V.; Soylu, A.; Roman, D. Building Semantic Knowledge Graphs from (Semi-) Structured Data: A Review. *Future Internet* **2022**, *14*, 129. [CrossRef]
68. Scholz, M. A Mapping of CIDOC CRM Events to German Wordnet for Event Detection in Texts. In Proceedings of the CRMEX@ TPDL, Valletta, Malta, 22–26 September 2013; pp. 1–10.
69. Wettlaufer, J.; Johnson, C.; Scholz, M.; Fichtner, M.; Thotempudi, S.G. Semantic Blumenbach: Exploration of text–object relationships with semantic web technology in the history of science. *Digit. Scholarsh. Humanit.* **2015**, *30*, i187–i198. [CrossRef]
70. Byrne, K. Putting hybrid cultural data on the semantic web. Available online: https://jodi-ojs-tdl.tdl.org/jodi/index.php/jodi/article/view/700 (accessed on 1 June 2022).
71. Ayala, D.; Hernández, I.; Ruiz, D.; Rahm, E. Leapme: Learning-based property matching with embeddings. *Data Knowl. Eng.* **2022**, *137*, 101943. [CrossRef]
72. Song, J.; Lin, Z. Neural Machine Translating from XML to RDF. In Proceedings of the 2021 6th International Conference on Mathematics and Artificial Intelligence, Chengdu, China, 19–21 March 2021; pp. 130–136.

73. Dimitrakis, E.; Sgontzos, K.; Tzitzikas, Y. A survey on question answering systems over linked data and documents. *J. Intell. Inf. Syst.* **2020**, *55*, 233–259. [CrossRef]
74. Nikas, C.; Fafalios, P.; Tzitzikas, Y. Open Domain Question Answering over Knowledge Graphs using Keyword Search, Answer Type Prediction, SPARQL and Pre-trained Neural Models. In Proceedings of the International Semantic Web Conference, Hangzhou, China, 24–28 October 2021; pp. 235–251.
75. Kesäniemi, J.; Koho, M.; Ikkala, E.; Hyvönen, E. Using Wikibase for Managing Cultural Heritage Linked Open Data Based on CIDOC CRM. In Proceedings of the 6th Conference, DHNB 2022: Digital Humanities in the Nordic and Baltic Countries, Uppsala, Sweden, 5–18 March 2022.
76. Debattista, J.; Attard, J.; Brennan, R.; O'Sullivan, D. Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud. In Proceedings of the Companion Proceedings of the 2019 World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 850–858.
77. Xin, D.; Ma, L.; Liu, J.; Macke, S.; Song, S.; Parameswaran, A. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In Proceedings of the 2nd Workshop on Data Management for End-to-End Machine Learning, Houston, TX, USA, 15 June 2018; pp. 1–4.
78. Akata, Z.; Balliet, D.; De Rijke, M.; Dignum, F.; Dignum, V.; Eiben, G.; Fokkens, A.; Grossi, D.; Hindriks, K.; Hoos, H.; et al. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* **2020**, *53*, 18–28. [CrossRef]