

## Article

# Extraction of Event-Related Information from Text for the Representation of Cultural Heritage

Emmanouil Ntafotis <sup>1</sup>, Emmanouil Zidianakis <sup>1,\*</sup>, Nikolaos Partarakis <sup>1</sup> and Constantine Stephanidis <sup>1,2</sup><sup>1</sup> Foundation of Research and Technology, N. Plastira 100, Vassilika Vouton, 700 13 Heraklion, Greece<sup>2</sup> Department of Computer Science, University of Crete, Voutes Campus, 700 13 Heraklion, Greece

\* Correspondence: zidian@ics.forth.gr

**Abstract:** In knowledge representation systems for Cultural Heritage (CH) there is a vast amount of curated textual information for CH objects and sites. However, the large-scale study of the accumulated knowledge is difficult as long as it is provided in the form of free text. By extracting the most significant pieces of information from textual descriptions of CH objects and sites and compiling them in a single comprehensive knowledge graph, conforming to a standard would facilitate its exploitation from multiple perspectives including study, presentation and narratives. The method proposed by this research work was to employ Natural Language Processing, and reinforcement learning for semantic knowledge extraction, and a knowledge representation standard of the CH domain for the knowledge graph thus making the extracted knowledge directly compatible with linked open data platforms and CH representation systems.

**Keywords:** natural language processing; NLP; museum; Python; SpaCy; invisible museum; CIDOC-CRM; NLP4CH



**Citation:** Ntafotis, E.; Zidianakis, E.; Partarakis, N.; Stephanidis, C. Extraction of Event-Related Information from Text for the Representation of Cultural Heritage. *Heritage* **2022**, *5*, 3374–3396. <https://doi.org/10.3390/heritage5040173>

Academic Editor: Francesco Soldovieri

Received: 3 October 2022

Accepted: 7 November 2022

Published: 9 November 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Today there are many platforms that bring the expressive power of semantic web technologies to the hands of knowledge curators in the domain of CH. Starting from the need to represent artefacts and their metadata such platforms have matured into knowledge representation systems that can support the representation of various dimensions of tangible and intangible CH [1–4].

With the development of the “Invisible Museum” platform [1] and its acceptance and support by the target userbase, there arose a significant realization: Among the numerous exhibits created and displayed on the platform, there was a wealth of knowledge and information accompanying each of them, in the form of a free-text narration, and there was a lot to be gained from mass-studying the contents of the various narrations. However, as free text is difficult to automatically study and process, especially in a greater scale, a more specialized tool needed to be developed, in order to effectively provide a way for valuable information to be separated, extracted and stored for future use. To that end, a solution was conceived, for text analysis through the use of Natural Language Processing. While an accurate and complete definition of the term might be difficult to agree upon [5], for the purposes of this project let be the following rough definition:

*Natural Language Processing is the term referring to the computational methods developed with the goal of automatizing the process of text analysis and processing, with human levels of accuracy and efficiency.*

It is worth noting that all systems concerning natural language processing, examine given inputs in one or more of the several different levels that have been proposed [5,6]. More specifically:

- Phonology: The level concerning the speech sounds of the spoken word.

- Morphology: The level centered around words being broken down to morphemes to extract meaning.
- Lexical: The level at which the meaning of each word is interpreted, partially with the use of part-of-speech tagging.
- Syntactic: At this level, sentences are deconstructed and examined for their grammatical structure.
- Semantic: The level at which the meaning of a sentence is determined by first deducing the relationships between the different words and terms.
- Discourse: In contrast to previous levels, this one centers around analyzing more complex and extensive pieces of text, and interpreting their meaning by examining sentences and similar components.
- Pragmatic: This level concerns the inference of meaning that has not been strictly encoded in the text in any way by analyzing the surrounding context, and usually requires significantly more world knowledge to function.

The Natural Language Processor for Cultural Heritage Texts (hereafter referred to as “NLP4CH”), mainly covers the Lexical and Semantic levels.

## 2. Related Work

### 2.1. Invisible Museum

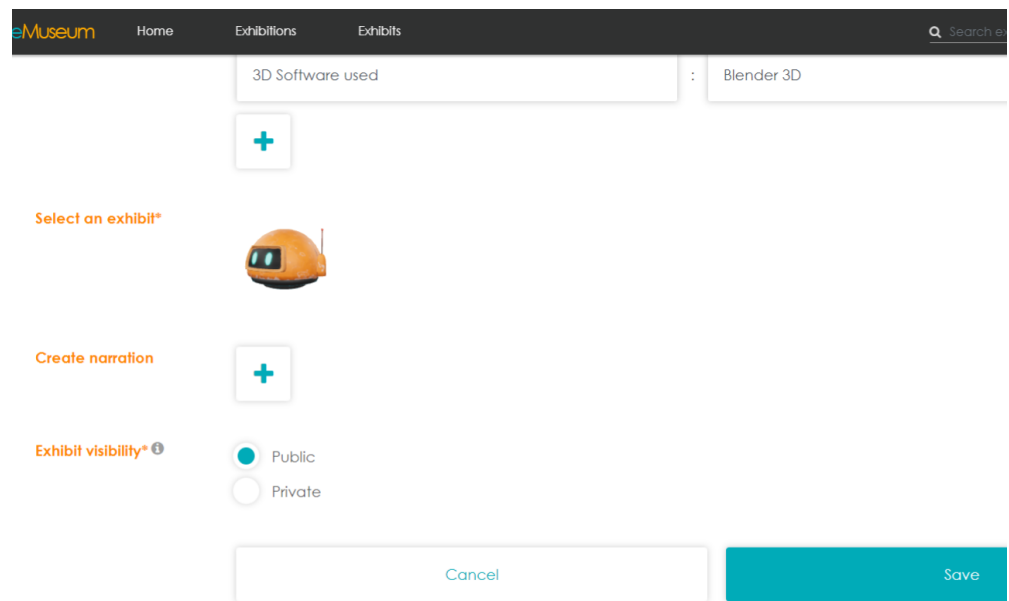
The work presented in this manuscript was developed for the purposes of the “Invisible Museum” (IM) platform [1], in order to offer a more robust experience to its users. The Invisible Museum itself is a web-based platform centered around the concept of bringing together the ever-increasing capabilities of VR and AR technologies with the cultural significance of physical museums. The central feature of the IM platform is the user-centered design of the human experience regarding museums, from touring exhibitions in digitally-designed museums, to creating one’s own museum space filled with exhibits of their own. Within the bounds of the platform, users can upload their own 3D rendered exhibits, embellishing them with narrations and pictures that enrich the history and character of a given exhibit, as presented in Figure 1a,b.

The screenshot shows the 'CREATE AN EXHIBIT' form on the Invisible Museum platform. The form is structured as follows:

- Exhibit language\*:** Two buttons are visible: 'English' and 'Ελληνικά'.
- Exhibit title\*:** A text input field containing 'Robot Assistant'.
- Short description:** A text input field containing 'This robot assistant 3D model was created for the research work done on Cave Alistrati VR Exhibition.'
- Full description:** A larger text input field containing 'This robot assistant 3D model was created for the research work done on Cave Alistrati VR Exhibition. The robot's design was mainly inspired by May's robot from Overwatch.'
- Exhibit categories\*:** Two tags are present: '#INDUSTRIAL DESIGN' and '#MINERALOGY'.

(a)

Figure 1. Cont.



(b)

**Figure 1.** (a) The user creates a virtual exhibit... (b) ... and adds additional info and content to embellish their creation.

The users are then able to design and decorate the digital space of their museums, managing every aspect of the design, from the floor layout to the lighting and including their own and others' exhibits in their creations. An example of this functionality, is presented in Figures 2 and 3.

After its creation, users of the IM platform are able to visit the museum space, according to tours created by the owner or browsing freely at their own pace, admiring the various exhibits and acquiring a more in-depth perspective with the aid of the narration accompanying them.

## 2.2. Semantic Representation of Cultural Heritage

Due to the vast amounts of knowledge contained in text corpora, the codification and summation of information regarding objects of cultural heritage, is a goal that has been approached many times in the past. One notable example would be the work of Regine Stein and Erin Coburn [7] in the form of the XML schema CDWA Lite [8], and its improvement, the museumdat [9] schema. Both of these schemas were created for the purpose of efficient content analysis of cultural heritage objects by including specific properties for each object, and the schemas were designed to be CIDOC-CRM [10] compliant, another key similarity with the work presented in this paper. In fact, CIDOC-CRM compliance seems to be a common basis for many of the approaches to the semantic representation of cultural information, and even beyond. As outlined in the work of Guenther Goerz and Martin Scholz [11], in the interest of avoiding misunderstandings, conflicting terminologies, and other inter- and intra-disciplinary differences, the CIDOC-CRM provides a stable framework to store, receive, and parse large amounts of semantically enriched information.

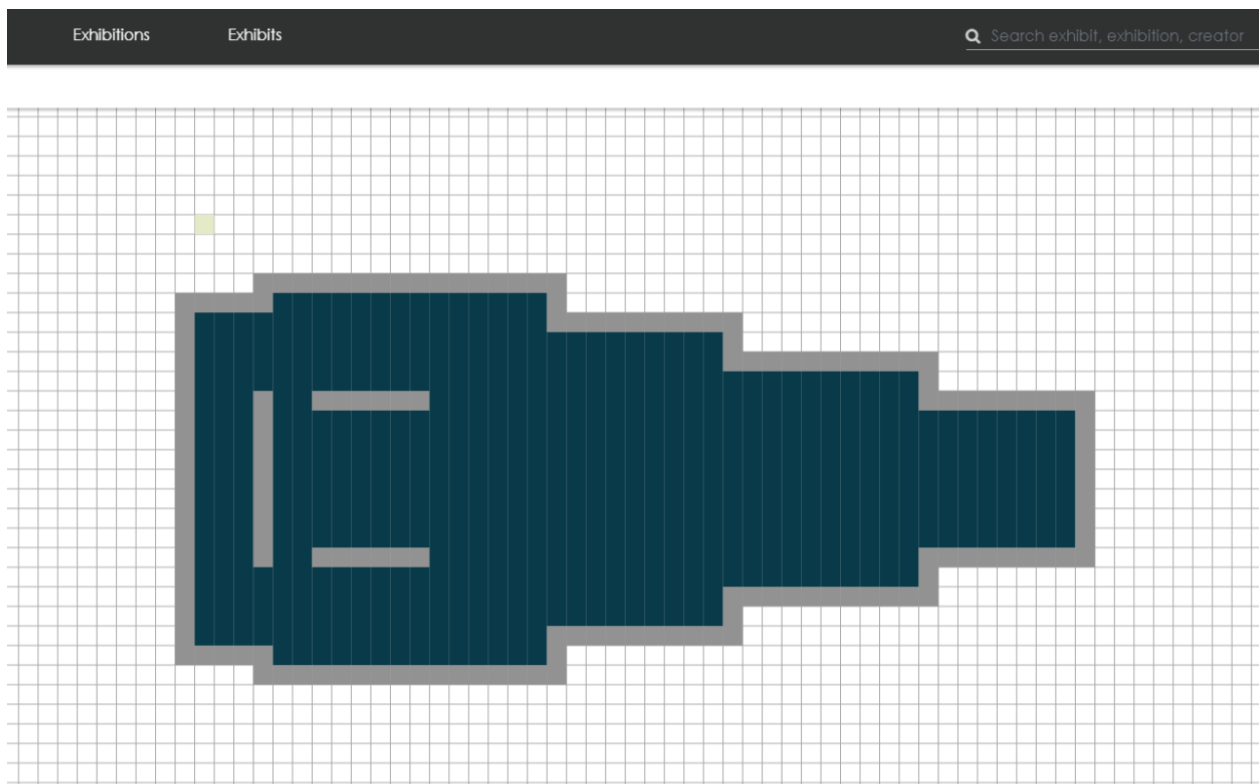


Figure 2. A user-created museum layout.

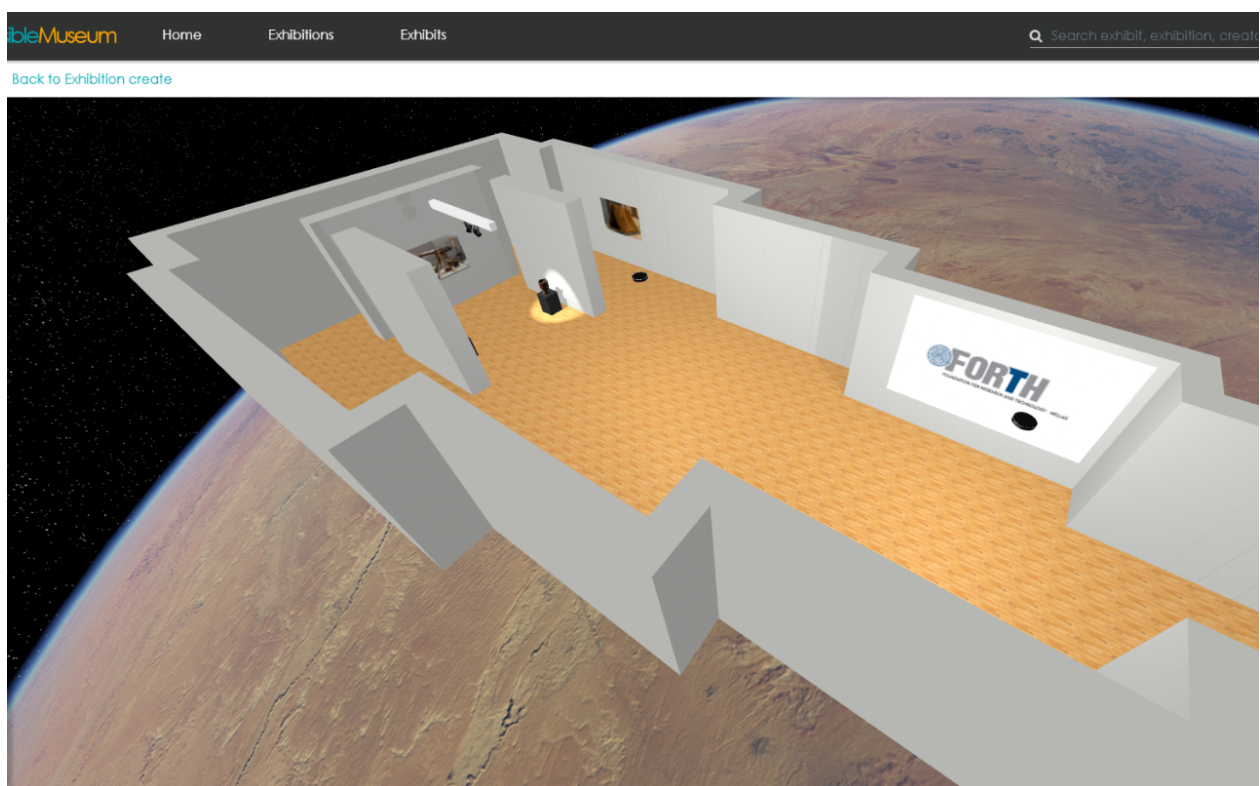


Figure 3. The museum space of Figure 2, with added decorations and lighting.

### 2.3. Natural Language Processing and Application in Cultural Heritage Research

During the past two decades numerous exploited the semantic representation of CH entities using CIDOC-CRM, meta-models of CIDOC-CRM or other domain ontologies. Popular examples are the British Museum's ResearchSpace [12], the Finish CultureSampo [13], DigiCULT [14], CASPAR [15], CrossCult [16] and many others.

While an ever-growing interest in the digitisation of cultural heritage information has provided the incentive for extensive research in the application of NLP techniques, there exist many challenges to be overcome for any such project to achieve any degree of demonstrable success. Such difficulties (e.g., disparities between older and newer forms of a language), as well as more general guidelines and common practices (e.g., part-of-speech tagging as a pre-processing step) in the field, are expertly summarized by Caroline Sporleder [17]. One major success of NLP application to cultural heritage, is the Wisski virtual research environment [11,18]. Wisski provides a robust tool for the extraction and presentation of semantically enhanced information, as well as its storage according to multiple different schemas (CIDOC-CRM included), making it a uniquely complete application that covers all the basic benchmarks that were previously discussed.

### 2.4. Existing Narrative Authoring Systems

The online authoring of narratives on Cultural Heritage is a research subject that has gained the attention of many research works. One of the first approaches was the Narrative Building and Visualising Tool (NBVT), a software that allows users to construct and visualize narratives through a Web interface [19,20]. In the same context DanteSources, was a Web application that allowed free access to the knowledge about Dante Alighieri's primary sources, i.e., the works of other authors that Dante cites in their texts [21–23]. Later approaches extended these works by providing a well defined methodology [24] and the appropriate tools to support socio-historic narratives and traditional craft processes as forms of Intangible Cultural Heritage (ICH) [3,4]. In the same context, another dimension studied is recipes as a form of ICH strongly connected to the place, community, and identity expressed as a collective narrative [2]. Finally, approaches have studied the possibility of integrating the conceptualisation of narratives to existing systems for CH representation such as Europeana [25].

All of the above approaches require a lot of manual labour both in terms of social science and humanities research and in terms of authoring CH resources used to construct and model narratives and narrations.

In our work, we attempt to simplify this process by integrating a semi-automated mechanism for extraction of knowledge from texts and reinforcement learning to enhance our model on the detection of entities in CH texts.

### 2.5. Contribution of This Research Work

The research presented in this paper serves a dual purpose. The first goal of the NLP4CH is to provide a robust narration enhancing tool to the existing Invisible Museum platform, in order to further enrich the overall experience of its users, offering them the chance to link knowledge they possess with relevant information offered by other people. The second and more important role of this work is to attempt to provide a valuable asset for researchers, museum curators and other parties interested in the study of cultural heritage on a larger scale. As it stands, attempting to manually extract any measure of knowledge from large corpora of text requires a lot of time and effort to be expended by experts of a specific field. As will be made clear during the course of this paper, the NLP4CH has shown promising signs as an adaptable, user-moderated tool, that could in the long run provide a significant alternative solution to that problem, and assist the scientific community by partially removing the need of such manual effort.

According to the work of Metilli, Bartalesi and Meghini [26] main requirements to support meaningful social and historic information for texts are the ability to:

- detect and classify Events
- identify named entities
- identify entities that act as arguments of the events
- extract temporal entities
- extract relations
- link entities and events

With the exception of the last item of entity and event linking, all other requirements are currently being met by the NLP4CH system. Entity linking is also planned to be covered in the future, but due to issues such as terms referring to the same entity (e.g., 'WWII', 'World War II', and 'Second World War'), it was deemed best that such a feature be considered after a complete version of the system was ready. The details of each feature are explained in the following sections but briefly summarised, NLP4CH detects and classifies events by identifying their structural components as separate entities. Events such as births are detected by specific words, e.g., the word 'born', and the arguments of the events are different entities, including temporal and named, which are detected by a model trained through machine learning. The various entities are then examined according to grammatical, contextual and syntactic rules and if they fit the required patterns, they are then linked, forming an event.

### 3. Proposed Methodology

In this work, we propose a reinforcement learning methodology for extracting semantic and lexical information from plain texts of cultural content and significance. The main objective is to support curators and CH experts in the enhancement of authored text-based knowledge with semantic information both during and after text processing. More specifically, this is accomplished through two distinct loops. The first loop regards the real-time authoring of narrations where the system provides recommendations from existing resources that could be linked to enhance the narration text. The second loop regards the post-authoring phase where the entire text is analyzed and semantic constructs are identified within the text. In both cases, the user remains in the loop to validate system recommendations. User decisions on the provided recommendations are used to reinforce the recommendation system for future iterations.

The full pipeline of the NLP4CH, consists of three major components: the processor itself, the knowledge graph of stored knowledge along with the schema that outlines the information detected and stored, and the user participation aspect. The end goal is a steady feedback loop between the system and the users, in which the system is constantly improved in its processing capacity and enriched with more information, and the users benefit from the system in order to enhance and connect their own narrations. The full pipeline is explained in broad strokes in Figure 4.



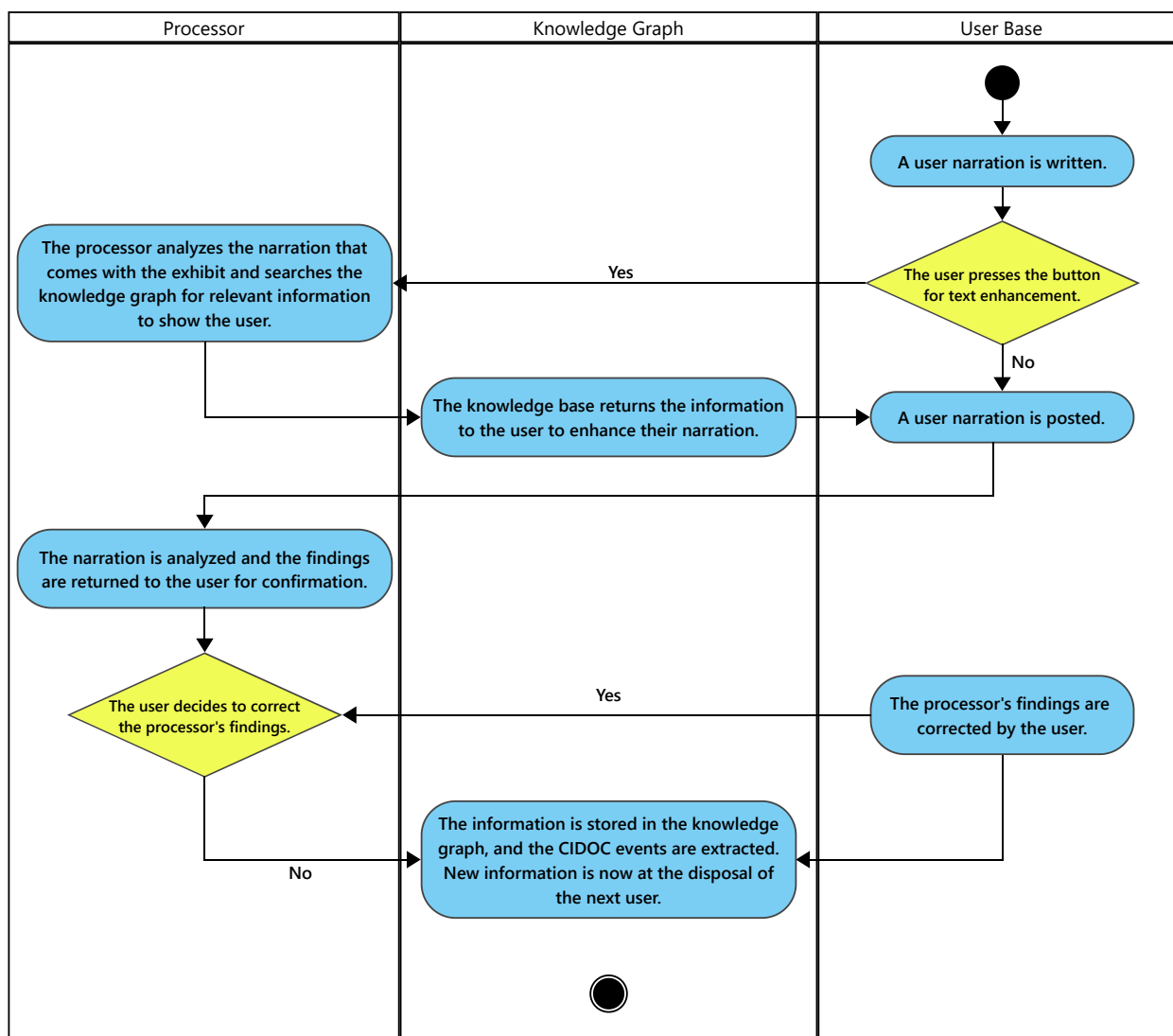


Figure 4. Full diagram of the NLP4CH loop.

#### 4. Implementation

##### 4.1. Extracting Knowledge from Text

The first major decision for the project, was the selection of the main tools that would extract the desired information. Because of the complexity of the task, a high-level language was deemed best, so as to avoid needless effort that would go into menial tasks, such as separating a given text into individual words. Python’s abundance of complimentary, open-source libraries made it a natural choice.

As Python’s basic functions were not sufficient for the scale of the project, there was a need for more appropriate tools provided by additional libraries. Many alternatives were considered (e.g., the NLTK library), but eventually the SpaCy library [27] was selected for a variety of reasons [28–30]:

- SpaCy provides pre-determined functions for individual tasks that would be needed, such as separating a free text into words, or grouping together multiple words when they refer to a single term.
- While SpaCy offers a pre-trained model for the identification of Named Entities, it also enables the training of custom models that best suit the specific context of use.

The role SpaCy filled in the project was to provide a model that would recognise the separate entities of each document, so that further rules could build upon these recognitions to extract CIDOC-CRM events. CIDOC-CRM events will be explained in the following sec-

tion, but a comprehensive example would be a “birth event”, e.g., the birth of Domenikos Theotokopoulos. This particular event can be summed up as the aforementioned “birth event”, but a typical SpaCy model is not capable of recognising it by itself, especially in the cases where the information is interrupted by irrelevant free text. The method that was followed instead was to use a SpaCy model to recognise the separate terms that would represent a birth event, and then compile that information into events using rough syntax rules (see “Extracting CIDOC-CRM events” section). Returning to the aforementioned example, the model would not be trained to identify the birth of Domenikos Theotokopoulos, but rather it would identify the name “Domenikos Theotokopoulos” as a person, the phrase “1 October 1541” as a date, and potentially the verb “born” as a verb that signifies birth.

In the following section, the exact form of CIDOC-CRM events will be discussed, so that *modus operandi* of the NLP4CH is made clearer.

#### 4.2. CIDOC-CRM

The second important characteristic that had to be defined early in the design phase of the NLP4CH was the standard according to which the information extracted would be stored in the knowledge graph. The most convenient choice would have been to define a custom standard to work with, according to the needs and expectations of the project. In the end, however, a more universal approach was selected in the form of a more widely-used standard, namely the CIDOC Conceptual Reference Model [31].

The CIDOC Conceptual Reference Model is a standard used worldwide for the storage and codification of cultural information, providing a formal structure for common relationships and attributes regarded in cultural heritage documentation, such as the births of important figures and the constructions of monuments. Factors that determined the choice of the CIDOC-CRM model include:

- Its worldwide use in the context of historical documentation, and adopting it ensures a more convenient way for experts to study the data extracted.
- It continues to be updated constantly (as of the publication date of this paper), ensuring the relevance and validity of the structure it provides.
- Has a very wide scope of definitions, relationships and characteristics, enabling the selection of the appropriate subset that would be relevant to the project’s goals.
- Has a wide compatibility with existing CH data sources through the definition of several knowledge mapping approaches from existing datasets to CIDOC-CRM (e.g., [32–35]).
- Its structure is event-centric, which makes it a very appropriate choice for the representation and description of the various events

The CIDOC-CRM structure centers around the ontological definition of various events, objects, locations and people as “entities”. There is a specific hierarchy between entity types, with their subclasses further down the hierarchy, inheriting all attributes and properties of their superclasses and including more specific aspects of the entities they refer to. Each entity has a specific code attached to its name, which determines (a) the content type it codifies, (b) its position within the CIDOC-CRM hierarchy, and (c) the properties that characterise the entity in question.

For a more specific example, according to the CIDOC-CRM model version 7.1.2 [31], entity type “E39 Actor” refers to “people, either individually or in groups, who have the potential to perform intentional actions of kinds for which someone may be held responsible.”. It has some properties, such as “P75 possesses”, which refers to an entity “E30 Right”, to denote a right this actor entity has. In addition, it inherits all the properties of its superclass “E77 Persistent Item”, and the superclass of E77 in turn. In addition, E39 has two subclasses: “E21 Person” and “E74 Group” referring to individual people or multiple people acting as a single distinguished group. These classes inherit all the properties of their E39 superclass, and add some unique properties of their own. It is worth noting that each property is tied to a specific entity type that it accepts as a value (as previously mentioned, “P75 possesses” has a “E30 Right” value).



While the CIDOC-CRM's plethora of entities and properties is one of its biggest strengths, it is also too vast a territory to be immediately covered by this project. Scalability is important, and the eventual goal of the project is to envelope as much of the standard as possible, but to provide the community with a starting point, a subset of the CIDOC-CRM was chosen. This working subset includes the entities as presented below:

E1	CRM Entity
E2	- Temporal Entity
E4	- - Period
E5	- - - Event
E7	- - - - Activity
E11	- - - - - Modification
E12	- - - - - - Production
E13	- - - - - - Attribute Assignment
E15	- - - - - - Identifier Assignment
E65	- - - - - - Creation
E83	- - - - - - - Type Creation
E63	- - - - - Beginning of Existence
<i>E65</i>	- - - - - <i>Creation</i>
<i>E12</i>	- - - - - <i>Production</i>
E67	- - - - - Birth
E64	- - - - - End of Existence
E6	- - - - - Destruction
E69	- - - - - Death
E52	- Time-Span
E53	- Place
E54	- Dimension
E77	- Persistent Item
E70	- - Thing
E72	- - - Legal Object
E18	- - - - Physical Thing
E24	- - - - - Physical Human-Made Thing
E22	- - - - - - Human-Made Object
E90	- - - - - Symbolic Object
E41	- - - - - Appellation
E73	- - - - - Information Object
E31	- - - - - - Document
E32	- - - - - - - Authority Document
E33	- - - - - - Linguistic Object
E36	- - - - - - Visual Item
E71	- - - - Human-Made Thing
<i>E24</i>	- - - - - <i>Physical Human-Made Thing</i>
E28	- - - - - Conceptual Object
<i>E90</i>	- - - - - <i>Symbolic Object</i>
E89	- - - - - Propositional Object
<i>E73</i>	- - - - - <i>Information Object</i>
E30	- - - - - Right
E55	- - - - - Type
E39	- - Actor
E21	- - - Person
E74	- - - Group

At the time of publishing, the NLP4CH API supports the location of births and deaths of people, and creation and destruction of monuments, with attribute assignment and modification being works in progress. The focus of the following segment, will be to clarify the inner workings of the NLP4CH pipeline, between the identification of the terms that the API finds, and the storage of the CIDOC-CRM events to the knowledge graph.

#### 4.3. Extracting CIDOC-CRM Events

As previously explained, locating complex events solely with the use of the SpaCy model proved to be somewhat troublesome. In addition to CIDOC-CRM events containing multiple properties that are found in non-consecutive order within a free text, it is more practical to separately identify the building blocks of each event, as some of these basic elements contribute information to multiple entities simultaneously. For example, let the sentence “Richard I (8 September 1157–6 April 1199) was crowned King of England in 1189”. In this sentence, one could distinguish 3 separate events: The birth of Richard I (“E67 Birth”), their death (“E69 Death”), and their being crowned King of England (“E7 Activity”). The SpaCy library does not allow for a single piece of text to be labeled with more than one label, nor does it allow for interruptions in a single entity. Thus, it is not realistically possible to always use a single SpaCy label to denote a CIDOC-CRM event. From now on, in this section, these CIDOC-CRM events will be referred to as “complex entities”. Instead, SpaCy models are being used to note more basic concepts, such as a person or a word that would signify an entity. Those entities that are annotated by the SpaCy model are going to be referred to as “simple entities”.

The first step is to tokenize the text as SpaCy does, which returns a list of all the separated elements of the free text, including words, numbers and punctuation marks. Afterwards, the tokenized text is processed by the trained model, in order to label all the simple entities it can locate. In its current version, the model is being trained to identify the following simple entities:

- **PERSON**: a person, real or fictional (e.g., Winston Churchill).
- **MONUMENT**: a physical monument, either constructed or natural (e.g., the Parthenon).
- **DATE**: a date or time period of any format (e.g., 1 January 2000, 4th century B.C.).
- **NORP**: a nationality, a religious or a political group (e.g., Japanese).
- **EVENT**: an “instantaneous” change of state (e.g., the birth of Cleopatra, the Battle of Stalingrad, World War II).
- **GPE**: a geopolitical entity (e.g., Germany, Athens, California).
- **SOE**: “Start of existence”, meaning a verb or phrase that signifies the birth, construction, or production of a person or object (e.g., the phrase “was born” in the phrase “Alexander was born in 1986.”).
- **EOE**: “End of existence”, meaning a verb or phrase that signifies the death, destruction, or dismantling of a person or object (e.g., the verb “Jonathan died during the Second World War”)

Returning to the example above, the model would mark the text with following labels:

PERSON	DATE	DATE		GPE		DATE
Richard I	(8 September 1157–6 April 1199)		was crowned King of	England	in	1189

To make it more clear, let us examine another sentence, with the appropriate annotation:

PERSON		SOE		GPE		DATE		EOE		GPE		DATE
Frida Kahlo,	who was	born	in	Mexico	on	6 July 1907,	died	in	Mexico	on		13 July 1954

It is worth noting that already, the “PERSON” label denotes an “E21 Person” complex entity and the “DATE” label denotes an “E52 Time-Span” complex entity, however, without properties or a relationship to another entity, they have no significant value, and thus are not stored in the knowledge graph. Now that the text is properly labeled, a more syntactical approach is in order. In this stage, the text is examined separately for each kind of complex entity. Each processing is based on common patterns encountered in the context of the complex entity type that is examined. For an example pertaining to the sentence examined before, when locating potential birth or death events (in other words, “E67 Birth” or “E69 Death” complex entities), the two common syntaxes encountered above are as follows:

“<PERSON> (<DATE OF BIRTH> - <DATE OF DEATH>)...”

“<PERSON> <SOE> <PLACE OF BIRTH> <DATE OF BIRTH>, <EOE>  
<PLACE OF DEATH> <DATE OF DEATH>...”

Translating the syntaxes above in terms of labeling, the terms <DATE OF BIRTH> and <DATE OF DEATH> would be labeled with the more generic “DATE” label, and the terms PLACE OF BIRTH and PLACE OF DEATH would be accordingly labeled with GPE (at least at this stage, as it makes the model more accurate and easier to train). As such the program examines the text, and finding a syntax that fits the complex entity type being examined (i.e., the birth event), creates a *Python Dictionary* to represent the complex entity. The dictionary created follows the CIDOC-CRM hierarchy presented in the previous section, by nesting each class as a “child” of its superclass, starting from “E1 CRM Entity”. Each dictionary has a very specific format and contains the following “key:value” pairs:

- “**etype**”: The identifier of the specific entity type according to the CIDOC-CRM standard (e.g., “E1” for entities of type “E1 CRM Entity”).
- “**final\_entity**”: The identifier of the innermost “child” entity type (see below). It is identical in value format to the “etype” key, and is used to make the retrieval and parsing of information easier, serving as a beforehand piece of information on the entity this dictionary represents.
- “**narration\_step**”: As the narration the API receives is separated in steps, this number represents the specific step this entity was found in. It is worth noting that an assumption was made, that no single entity would be scattered in multiple steps.
- “**Pxxx**”: A dictionary can contain any number of properties. These fields represent the various properties of each entity type. The values are always strings, even if the entity represented would normally be more complicated, containing properties of its own.
- “**child**”: The key used to maintain the CIDOC-CRM hierarchy. The value of this key is a dictionary representing their subclass in the hierarchy, until the final entity is reached. The child in turn contains its own “Pxxx” fields and an “etype” field. The “final\_entity” and “narration\_step” keys do not occur after the outermost dictionary.

Figure 5 is an indicative example of the birth of Richard I, as it would be extracted from the first given sentence:

```

{
  "child": {
    "P115": null,
    "P116": null,
    "P117": "8 September 1157",
    "P120": null,
    "P4": null,
    "child": {
      "P7": null,
      "child": {
        "P11": null,
        "P12": null,
        "child": {
          "child": {
            "P98": "Richard I",
            "child": null,
            "etype": "E67"
          },
          "etype": "E63"
        },
        "etype": "E5"
      },
      "etype": "E4"
    },
    "etype": "E2"
  },
  "etype": "E1",
  "final_entity": "E67",
  "narration_step": 1
}

```

**Figure 5.** The event “Birth of Richard I”.

As the figure indicates, all entities will have an outermost “E1” entity and then an appropriate number of “child” dictionaries nested within, until the final entity is reached. In the final entity, the “child” key will have a null value, and its “etype” value will be the same as the outermost “final\_entity” value.

The same process is followed for other complex entity types, and as all the complex entities are identified, they are saved as a list, which is afterwards saved in the knowledge graph for future use.

#### 4.4. User Correction & Training

What was described above was the full pipeline of the NLP4CH, from the input of an exhibit narration, to the extraction and storage of the CIDOC-CRM entities that stem from the text. However, natural language processing is an inexact technique and often prone to failure or mistake. To mitigate the inherent inaccuracy of the project, support for user correction was implemented and added to the pipeline. Before expanding on that however, one ought to be familiar with the supervised learning [36] capabilities for model training provided by SpaCy.

SpaCy offers pre-trained models ready to be used for text labeling, but such models are for more general use, and do not include sufficient support for more narrow and specific fields of study, while also complicating the process somewhat, by potentially locating information irrelevant to the subject of this project. Because of this, a decision was made to train a new model, using SpaCy’s built-in support of supervised learning. Custom training allows for inclusion of only a specifically selected set of labels to be included (referenced in Section 4.3 *Extracting CIDOC-CRM events*), as well as the training of the model in contexts encountered in the specific fields of interest.

In order to do that, the training program that produces new models must be provided with “pre-annotated” text, meaning any piece of free text, along with a list of all the labels that should ideally be identified by a perfectly trained model. A sample of pre-annotated text is provided in Figure 6.

```
('Nikos Kazantzakis, possibly the world's best-known  
Modern Greek writer, was born in Heraklion in 1883.',  
{'entities': [(0, 17, 'PERSON'), (58, 63, 'NORP'), (72,  
80, 'SOE'), (97, 101, 'DATE')]})
```

**Figure 6.** A sample pre-annotated piece of text for the training of the SpaCy model. The annotation consists of the starting and ending character of a single annotation, along with the appropriate label.

The training program's function is to iterate over the training data, and then produce a predictive annotation for the given texts according to the weight values of the model. Following that, the annotation predicted by the model is compared to the reference texts, and the weight values are adjusted according to the loss gradient produced, which in turn is determined by how accurate the prediction was (loss is further explained in Section 5). Afterwards, a new prediction is produced by the model-in-training, and the loop begins again, continuing for a number of cycles determined by the user each time the trainer is run. When determining the number of iterations, one should keep in mind that the number that yields the best models differs according to the training set. Specifically, too few repetitions could produce an insufficiently trained model, while too many would result in a model that is very well trained but only in the specific contexts represented in the training set, with practically no training for other cases. One additional aspect to consider, especially when hand-picking the texts that would comprise the training sets for the experimental iterations of the NLP4CH was the coverage of different contexts, syntaxes and phrasings. A set of very similar sets of texts, would result in the same overspecialization to specific patterns that was mentioned before, while texts with wildly different styles (e.g., a poem, an excerpt from formal report, and a short story full of slang terms) would make it difficult for a model to correctly annotate a piece of text belonging to any single one of those categories (even more so if it belonged to none of them). Finally, it is worth noting, after each iteration, the program shuffles the training texts in order to avoid biases that could result from a specific, concrete order of input.

Using a collection of pieces of labeled text like the one above, preferably with different syntaxes, the model is trained on identifying all the named entities encountered in its entire list of training material. A simple model could be considered functional with tens of pieces of training text, while more sophisticated models, with more labels to consider, a wider array of expected syntaxes, and a greater demand for accuracy, would need significantly more training material. One problem that arises with this prospect, is the custom labeling of large quantities of text; annotation of the training material would need to be done by hand to ensure its validity, but that would be relatively inefficient and time-consuming. Still, the NLP4CH model needed a way to constantly improve, especially since one of the main focuses during its development was scalability. In order to accomplish this, the solution implemented was a form of *user correction* of the NLP4CH's outputs, and their recirculation as input.

The ideal solution would allow the users to correct the CIDOC-CRM events themselves in order for them to be as accurate as possible. However, while the stored information would be valid, such a method would not facilitate the improvement of the NLP4CH system, as the machine learning aspect of it in the context of the SpaCy library could only happen at the entity recognition stage. In addition, that would put a significant burden on the part of the users, detracting from the intended user experience of the platform. Thus, the corrections mediated by the user improve the entity recogniser of the NLP4CH system, but leave the syntactic rules that produce the CIDOC-CRM events unaffected. Please note that the method of user correction proposed here is still in the experimental stage and is under examination both for its effectiveness and its ease of use by the user.

According to the current methodology, after the user has completed their composing of an exhibit narration, they will be prompted by the system to examine and correct the entities that have been marked by the NLP4CH. Declining is of course an option, and it will complete the narration, while accepting will take the user to a screen where the narration they have written will have the entities detected highlighted and color-coded to convey to the user the findings of the entity recogniser. Not all entities will be displayed, as some more obscure or technical ones, such as those with the 'NORP', 'GPE', 'SOE' and 'EOE' labels might confuse the user. Instead, 'NORP' will be shown to the user as "Religious, political or national affiliation" and 'GPE' will be shown as "Location", in order to be more comprehensive. 'SOE' and 'EOE' entities will be skipped altogether as they are both easier to detect and potentially more difficult for the user to define, because of the disconnect of the terms from natural everyday language. In this correction screen, users can mark the text differently, removing and adding entities within the constraints of the labels included, and once they have completed their corrections, the CIDOC-CRM events generated by the narration will result from the application of syntactic rules on the now correctly annotated text. The Invisible Museum platform has no requirements for signing up, so the userbase consists of a wide spectrum of people. However, for the purposes of the NLP4CH, the intended user is anyone with at least primary education, who is proficient in English at a B1 level.

In addition, every user-corrected narration will be stored in a separate training database along with its annotations (as presented in Figure 6). Every week (interval prone to change according to experimentation results), the training program will use the entirety of the training database as described earlier in this section, to train a new model and replace the old one. This means that the ever-increasing amount of training material available, in addition to the widening scope of its content as more users sign up to the IM platform, will ensure that each model produced will be more accurate and fine-tuned to the needs and wants of the userbase. Specifically, two separate models, designated as "model A" and "model B" will be stored on the server, as well as a backup one designated "model stable". Each weekly training will replace A or B, alternating between the two so as to always retain a previous version to fall back to in case any problems arise with a new version. In essence, after week 1 the training program will produce a model to replace model A and it will be the one in use until after week 2, when a new model replaces model B and is put to use instead, and so on. The stable model will be replaced once every 5 weeks with the same model that is to be put to use by the system, and its function is purely for security purposes in the case of a more severe fault in models A and B.

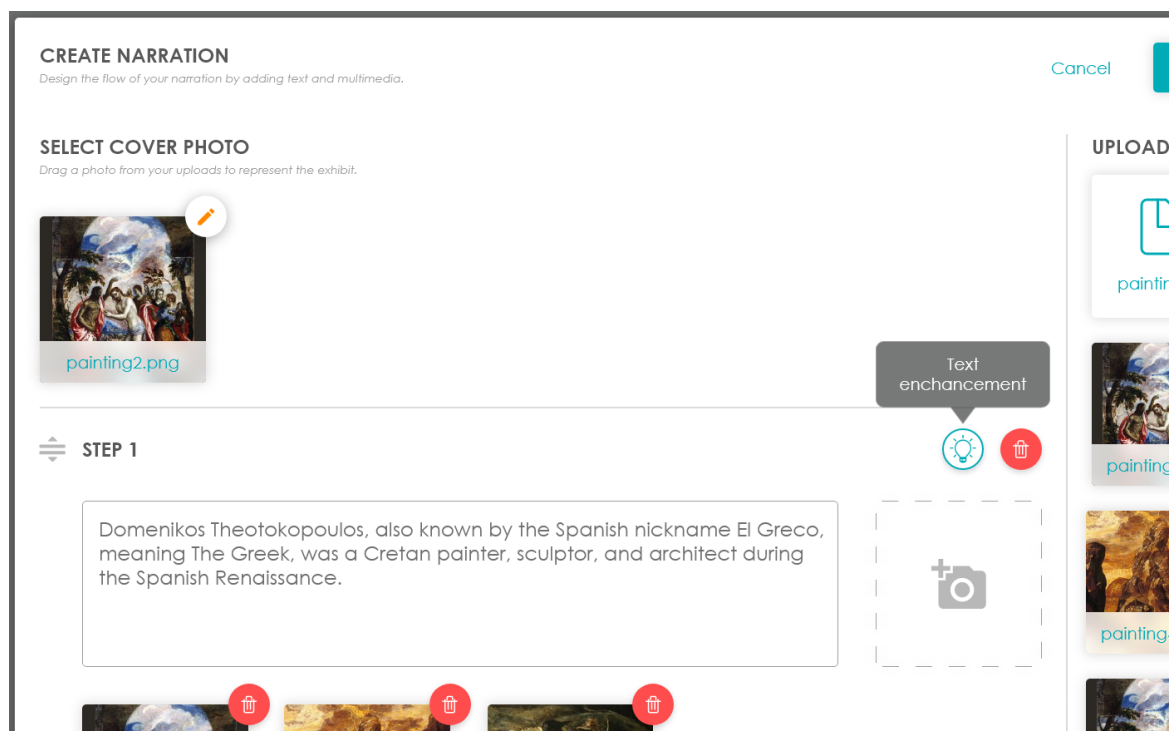
On the scale it is currently operating, the NLP4CH with all its functionality and the training program has been manageable enough that it runs on the same server as the rest of the Invisible Museum platform without problem, requiring no additional hardware or resources. The training only needs manual operation to begin, and even that will soon be updated, so that the retraining happens automatically at the aforementioned time interval with no need for administrative intervention. In addition, the training of a new model takes less than a minute, but even if its performance takes a hit, the system will still be operational as there are always two models available (A and B as explained above). The method described would admittedly present problems if the time it takes for a new model to be trained ever reached the time interval between model productions (one week for now), but as it stands, we have no indication of that being a realistic possibility.

#### 4.5. Text Enhancement

Now that it has been made clear how the users will contribute to the improvement of the system, let us examine how the NLP4CH will achieve its end goal of enhancing the experience of the users of the Invisible Museum platform.



The main focus of this part of the NLP4CH cycle is to facilitate easy access to information regarding a user's narratio and present it in a comprehensive and non-intrusive way. To that end, the text enhancement function is entirely optional. Specifically, the user can press the appropriate button when they have written a narration step they would like to enhance, as presented in Figure 7.



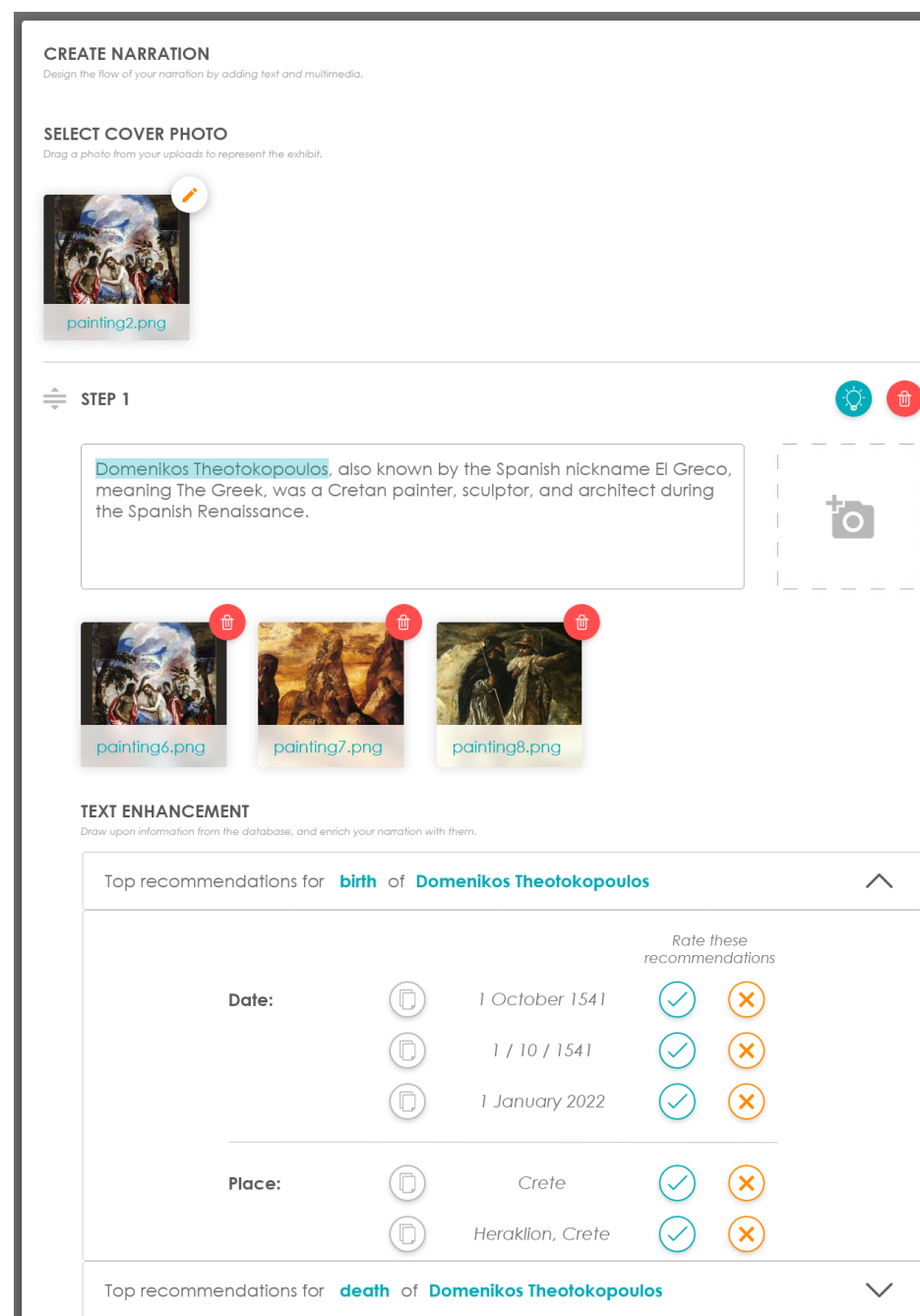
**Figure 7.** The user has the option to enhance their narration through the NLP4CH, with the help of the designated button.

The text is then internally annotated by the NLP4CH (as described in Section 4.3), and a list is compiled of all the 'MONUMENT' and 'PERSON' entities detected. Named entities such as specific people or named objects are significantly more likely to be the focus of a segment, and appear recurrently among different narrations, while providing additional information about other entities, e.g., specific dates or entire countries, would be too vague and counter-productive. A user writing "Domenikos Theotokopoulos was born in 1541." is far more likely to enhance their narration with more information about Domenikos Theotokopoulos, rather than further information about the year 1541. After the list is compiled, the knowledge graph will be searched for CIDOC-CRM events concerning these entities, and the results are returned to the user categorized by CIDOC-CRM event. This helps account for multiple different instances of the same "type" of information across multiple different narrations (e.g., two narrations where the death of Domenikos Theotokopoulos is detected, but the date or format is possibly different).

Once a user expands on the information offered by the system, they can access all the details of the events they have received. In the example of Figure 8, expanding on the recommendations for the "Birth" of "Domenikos Theotokopoulos", reveals stored information about the date and place of the event, as they have been detected in other narrations. As previously mentioned, placing minimal strain on the user's experience was of paramount importance, and thus while the system would be able to offer knowledge that would prove useful to the text, it shouldn't be as a forced intrusion to the user's text. To that end, the user is free to copy any piece of information they prefer, in order to include it to their own narration or ignore the provided recommendations altogether.

**CREATE NARRATION**  
Design the flow of your narration by adding text and multimedia.

**SELECT COVER PHOTO**  
Drag a photo from your uploads to represent the exhibit.



**STEP 1**

Domenikos Theotokopoulos, also known by the Spanish nickname El Greco, meaning The Greek, was a Cretan painter, sculptor, and architect during the Spanish Renaissance.

**TEXT ENHANCEMENT**  
Draw upon information from the database, and enrich your narration with them.

Top recommendations for **birth** of **Domenikos Theotokopoulos**

		Rate these recommendations	
<b>Date:</b>	<input type="checkbox"/>	1 October 1541	<input checked="" type="checkbox"/> <input type="checkbox"/>
	<input type="checkbox"/>	1 / 10 / 1541	<input checked="" type="checkbox"/> <input type="checkbox"/>
	<input type="checkbox"/>	1 January 2022	<input checked="" type="checkbox"/> <input type="checkbox"/>
<b>Place:</b>	<input type="checkbox"/>	Crete	<input checked="" type="checkbox"/> <input type="checkbox"/>
	<input type="checkbox"/>	Heraklion, Crete	<input checked="" type="checkbox"/> <input type="checkbox"/>

Top recommendations for **death** of **Domenikos Theotokopoulos**

**Figure 8.** The user is offered valuable information regarding their narration’s content.

Lastly, two concerns that had to be addressed were the potential volume of information provided to the user and the margin of misinformation that would be taken into account. Specifically, it is entirely within reason for a piece of information such as a birth event of a historic figure to have many instances of the same information, even ones formatted slightly differently. In addition, information detected is not guaranteed to be accurate by any measure, as the content of a narration is entirely up to the user, and monitoring its validity is not a measure that would be feasible (or uncontroversially beneficial) for the platform.

Attempting to tackle both of these problems simultaneously, one proposed solution was a form of user-based regulation of the knowledge provided by the NLP4CH. According to this methodology, for one specific piece of information, up to three (number still under consideration) available instances are shown to the user, who can then “rate” these recom-

recommendations by approving or rejecting them, regardless of whether they use the information in their own text. Each instance of information has a separate score, which is increased or decreased by user approval or rejection, and it is that score that determines the order the recommendations will appear in. Thus, inaccurate information will be less likely to be propagated between users after it has been disapproved by multiple users, while commonly accepted information is ensured to be among the first recommendations. This form of crowd-dependant regulation is not error-proof, and complete validity of information is still not ensured, but more concrete solutions are still being explored at this experimental stage, and similar forms of open regulation and information exchange has been proven to be sufficiently reliable in the past, as is the case with Wikipedia.

## 5. Preliminary Evaluation

In order to assess the effectiveness of the NLP4CH pipeline, there had to be measurable results to compare and draw conclusions from. The first parameter to be examined, is the effectiveness of retraining the model of SpaCy Named Entity Recogniser. To that end, the first metric to turn to, is the loss of the model being trained, as one common goal in the field of machine learning is the minimization of loss. As there are no similar products with which to compare the NLP4CH, the alternative solution was to compare different versions of models to examine how they comparatively improve as their training expands.

The initial corpus chosen for the purpose of producing measurable results consists of 40 short pieces of text, each hand-annotated and conforming as much as possible to the specifications detailed in Section 4.4 to ensure that the models are properly trained. Each text consists ranges from a single sentence to a small paragraph, which is the size of an average narration step expected by the user. All the texts chosen were written in English, as the IM platform currently supports Greek and English as its system languages, and because having all available data in a more widely-known language makes it more accessible to any other parties interested in the research presented. Finally, the texts chosen, were all written in mostly plain language, of a simple informative style, similar to what one might encounter on Wikipedia or similar sources, while it must acknowledged that the freedom offered to creators by the IM platform means that the style of any given text could range from a poem, to a single line stating the ISBN number of a book, in making the choice of training texts it was decided that the first experimental models produced should be closer to the ones typically expected.

Three sets of five models each were trained for the purposes of this use case; the first set was trained with 10 samples of the aforementioned pre-annotated texts, the second with 20, and the third with all 40 of them. The reason for training multiple models for each set is that the order the texts are examined in is randomized for each model, in order to avoid generalizations made by the model, based on the order the texts are parsed in. This results in different models being produced by each training session, even if the texts remain the same. For samples of that size, 5 models cover a good area of potential results. SpaCy provides a loss value during training after each iteration of the samples by a model being trained. Before examining the results, it is worth setting down what loss is, and how it affects the models being trained.

Loss is a metric that represents the inaccuracy of a given model, with its current weight tables, when compared to the reference material (in this case, the pre-annotated training texts). A higher loss value means a higher measure of inaccuracy in that particular prediction, and the minimum loss is 0, if the prediction is completely accurate both in locating all the named entities, and correctly labeling them. When using loss to improve a model's weight tables, greater values of loss signify a model needing more significant recalibration, and thus lead to more drastic adjustments of the weight tables, while a lower loss means the model's predictions were closer to the annotations provided as input. A loss of 0 is theoretically possible, and it would mean the model needs no adjustments of weight, but of course this is impossible in this case, as the training set of texts is always changing and expanding.

Now that loss has been clarified, the results are as follows:

Examining the results above, while the percentage of loss seems to remain relatively stable between groups, there is higher disparity to be amongst models of the same group, which is more stark the fewer sample texts there are. This becomes apparent when looking at the highest and lowest ranked models in terms of loss percentage, with the difference being 13.10% in Table 1, while the same difference is only 1.70% in Table 2 and 1.80% in Table 3. The reason for this is that fewer sample texts make the order in which they are parsed more impactful, and thus creates volatility in the loss and the resulting model. In contrast, models trained with more texts are not necessarily the ones with the least percentage of loss (that would be “Model 2” in Table 1, but they are still very low in loss, and more importantly, are more consistent in their results. The model that would be integrated in the system for a week would not greatly matter if it was chosen from Table 3, but the random chance would impact the system greatly in the case of Table 1, with equal chance to give the worst or the best performance of the set.

**Table 1.** Models trained with 10 samples and their loss function results.

	Initial Loss	Finishing Loss	Percentage of Loss That Persisted
Model 1	563.46	76.19	13.52%
Model 2	580.71	2.49	0.42%
Model 3	593.89	27.58	4.64%
Model 4	595.70	15.77	2.64%
Model 5	589.98	8.16	1.38%
<b>Mean</b>	<b>584.74</b>	<b>26.03</b>	<b>4.52%</b>

**Table 2.** Models trained with 20 samples and their loss function results.

	Initial Loss	Finishing Loss	Percentage of Loss That Persisted
Model 1	946.01	23.58	2.49%
Model 2	805.59	11.84	1.46%
Model 3	842.47	19.18	2.27%
Model 4	862.17	6.87	0.79%
Model 5	766.42	18.60	2.42%
<b>Mean</b>	<b>844.53</b>	<b>16.01</b>	<b>1.88%</b>

**Table 3.** Models trained with 40 samples and their loss function results.

	Initial Loss	Finishing Loss	Percentage of Loss That Persisted
Model 1	1075.73	38.63	3.59%
Model 2	1083.87	41.26	3.80%
Model 3	1028.73	29.41	2.85%
Model 4	1070.23	28.62	2.67%
Model 5	1139.28	22.81	2.00%
<b>Mean</b>	<b>1079.56</b>	<b>32.04</b>	<b>2.98%</b>

However, loss alone is not a sufficient indication of capability. A model with a training pool of 40 texts should still be superior than a model with a training pool of 10 texts of the same or even slightly better percentage. In order to verify this claim, a model from each of the previous texts was selected to be tested with 3 sample texts, similar to the ones anticipated on the platform. In order to ensure a comparison from the same starting point, the model with the median loss percentage of each table was selected, and its findings were compared to a user-annotated text. Below are presented each text and its annotations (Tables 4–9).

**Text 1:**

*William Shakespeare (26 April 1564–23 April 1616) was an English playwright, poet and actor. He is widely regarded as the greatest writer in the English language and the world's greatest dramatist.*

**Table 4.** User annotations.

User
-William Shakespeare: PERSON -26 April 1564 : DATE -23 April 1616: DATE -English: NORP -English: NORP

**Table 5.** Model annotations.

Model of 10	Model of 20	Model of 40
-William Shakespeare: PERSON -26 April 1564 : DATE -23 April 1616: DATE -English: NORP -English: NORP	-William Shakespeare: PERSON -26 April 1564 : DATE -23 April 1616: DATE -English: NORP -English: NORP	-William Shakespeare: PERSON -26 April 1564 : DATE -23 April 1616: DATE -English: NORP -English: NORP

As is evident, all the models were able to correctly identify all the entities in the text. As explained in Section 4.3, the format of this text is fairly common among academic writing pieces similar to the ones expected in the Invisible Museum platform, and therefore all models had sufficient examples to be trained.

**Text 2:**

*The Phaistos Disc is a disk of fired clay from the Minoan palace of Phaistos on the island of Crete, possibly dating to the middle or late Minoan Bronze Age (second millennium BC).*

**Table 6.** User annotations.

User		
-Phaistos Disc: MONUMENT -Minoan palace of Phaistos: MONUMENT -Crete: GPE -middle or late Minoan Bronze Age: DATE -second millennium BC: DATE	-Phaistos Disc: MONUMENT -Minoan: NORP -Phaistos: GPE -Crete: GPE -middle or late Minoan Bronze Age: DATE -second millennium BC: DATE	-Phaistos Disc :MONUMENT -Minoan: NORP -Phaistos: GPE -Crete: GPE -Minoan Bronze Age: DATE -second millennium BC: DATE

**Table 7.** Model annotations.

Model of 10	Model of 20	Model of 40
-The Phaistos Disc: MONUMENT -the Minoan: MONUMENT -Phaistos: GPE -Crete: GPE -Bronze: NORP	-The Phaistos Disc: MONUMENT -Minoan: PERSON -Phaistos: PERSON -Crete: NORP -Minoan Bronze: PERSON -BC: DATE	-Phaistos Disc: MONUMENT -Minoan: NORP -Phaistos: PERSON -Crete: GPE -Minoan Bronze Age: PERSON

The results of the second testing text are more nuanced. First of all, the user annotations table includes three equally valid interpretations of the text and annotations fitting either example are considered correct. As for the results, all models correctly identified the “Phaistos Disc” monument, although the first two added the optional “The” to the entity (not considered a mistake). Significant annotating disparities start being evident from this point onward, as all models correctly recognised “Minoan” as an entity, but the two models with lesser training misidentified the entity type (which should be of type “NORP”), and the least trained model also added ‘the’ to the entity.

In the next entity, the word ‘Phaistos’ presents an even more interesting case. Most notably, ‘Phaistos’ was identified correctly as a location (“GPE” entity) by the least trained model, but misidentified by all other models. It is important to note that in the context of the sentence and with no prior knowledge of Phaistos, an interpretation of ‘Phaistos’ as a person would still be valid, changing the meaning from ‘the Minoan palace located in Phaistos’ to ‘the Minoan palace of king Phaistos’.

The entity ‘Crete’ was misidentified by the model trained with 20 texts, and the term ‘Minoan Bronze Age’ seems to have proven a challenge for all the models, with the most accurate annotation being that of the Model of 40 which annotated it correctly (according to the third user annotation), but incorrectly labeled it a ‘PERSON’. Finally, no model identified the entity ‘second millennium BC’.

### Text 3:

*World War I, often abbreviated as WWI or WW1, also known as the First World War and contemporaneously known as the Great War, was an international conflict that began on 28 July 1914 and ended on 11 November 1918.*

**Table 8.** User annotations.

User
-World War I: EVENT
-WWI: EVENT
-WW1: EVENT
-First World War: EVENT
-Great War: EVENT
-28 July 1914: DATE
-11 November 1918: DATE

**Table 9.** Model annotations.

Model of 10	Model of 20	Model of 40
-World War I: PERSON	-World War: PERSON -WWI: GPE	-World War: EVENT -WWI: PERSON -WW1: PERSON
-First World War: EVENT -Great War: MONUMENT -28 July 1914: DATE -11 November 1918: DATE	-First World War: MONUMENT -Great War: DATE -28 July 1914: DATE -11 November 1918: DATE	-First World War: EVENT -Great War: EVENT -28 July 1914: DATE -11 November 1918: DATE

The third text was chosen for its density of ‘EVENT’ entities to test how the models would perform on a text that is unusually loaded with entities of the same type. This set of results is better examined not by comparing each entity on a model-by-model basis, but by examining the findings of each model and comparing overall performance.

Starting off with the dates at the end of the text, they follow the basic format that was most commonly encountered in CH texts, and thus even the model with the least training was accurate in annotating and labeling them. As for the rest of the results, there is a significant improvement between the accuracy of the most trained model and the previous models. The Model of 10 correctly identified 3 of the 5 ‘EVENT’ entities, but labeled only one of them correctly, while the Model of 20 correctly identified 4 of the entities, but failed to correctly label any of them. On the other hand, the Model of 40 managed to identify all 5 ‘EVENT’ entities and correctly labeled 3 of them.

## 6. Lessons Learned

Looking at the results regarding the Loss statistics and cross-referencing it with the following practical test of the 3 texts, the first conclusion to be drawn is that as the training set increases in size, the practical significance of the Loss percentage diminishes significantly. Specifically, it is not indicative of the overall performance of a model when compared to another of superior training, as was proven by the 3 models used for the testing texts, while 11 of their loss percentages were between 2.27% and 2.85%, the Model of 40 was demonstrably



more accurate in annotating the entities of the text. In addition, Loss percentages appear to be more volatile and present greater range as the training sets are reduced in size, leading to the conclusion that greater training sets lead to less unpredictability in Loss percentages, and thus more consistency in the accuracy of separate sets of models.

The results from the 3 samples of text for the NLP4CH seem to point to the conclusion that even in a relatively small scale of tens of texts, the improvement of a model's accuracy is quite evident, while the Model of 40 still had some difficulty annotating texts that stray from the more formulaic syntax that CH texts usually follow, it shows significant improvement compared to the other models. Further improvement, and training with larger-scale text corpora will be needed to confirm these claims, but these first results encourage further development.

## 7. Conclusions and Future Work

At its current stage, the field of Natural Language Processing is capable of producing remarkable tools to aid in the compilation and study of cultural heritage texts. The NLP4CH has already shown promising results, and further development is planned in the foreseeable future in order to mitigate mistakes and inaccuracies, and expand the scope of its capabilities.

Firstly, future efforts are planned to be focused on effectively connecting information from different narrations. One vital step towards this is identifying different names referring to the same entity. Due to the sheer amount of such terms, as well as new ones being used as time goes on, this problem demands an adaptive, scalable solution that will be able to account for such linguistic evolution.

In addition, the next step of advancement for the NLP4CH is its expansion in order to cover more events and to more accurately determine the ones already covered, while the latter comes down to further training specifically aimed at more extreme cases that might be encountered, expanding the scope of the events the NLP4CH covers is more complicated. Each type of entity to be included has to be studied in order to identify the contexts it most often appears in, words that might signify it in different formats, and examined with respect to any syntactical or other rules that can be applied to their use.

**Author Contributions:** Conceptualization, E.N. and E.Z.; methodology, E.N. and E.Z.; software, E.N.; validation, E.N., E.Z. and N.P.; formal analysis, E.N. and E.Z.; investigation, E.N. and N.P.; resources, E.Z. and C.S.; data curation, E.N.; writing—original draft preparation, E.N.; writing—review and editing, E.N. and N.P.; visualization, E.N.; supervision, E.Z. and N.P.; project administration, E.Z.; funding acquisition, E.Z. and C.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has been conducted in the context of the Unveiling the Invisible Museum research project (<http://invisible-museum.gr>, accessed on 8 November 2022), and has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH—CREATE—INNOVATE (project code: T1EDK-02725).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank all the employees of the Historical Museum of Crete, who participated in defining the requirements of the Invisible Museum platform, upon which the idea of this research work emerged.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
CH	Cultural Heritage
VR	Virtual Reality
AR	Augmented Reality
NBVT	Narrative Building and Visualising Tool

## References

- Zidianakis, E.; Partarakis, N.; Ntoa, S.; Dimopoulos, A.; Kopidaki, S.; Ntagianta, A.; Ntafotis, E.; Xhako, A.; Pervolarakis, Z.; Kontaki, E.; et al. The Invisible Museum: A User-Centric Platform for Creating Virtual 3D Exhibitions with VR Support. *Electronics* **2021**, *10*, 363. [[CrossRef](#)]
- Partarakis, N.; Kaplanidi, D.; Doulgeraki, P.; Karuzaki, E.; Petraki, A.; Metilli, D.; Bartalesi, V.; Adami, I.; Meghini, C.; Zabulis, X. Representation and Presentation of Culinary Tradition as Cultural Heritage. *Heritage* **2021**, *4*, 612–640. [[CrossRef](#)]
- Partarakis, N.N.; Doulgeraki, P.P.; Karuzaki, E.E.; Adami, I.I.; Ntoa, S.S.; Metilli, D.D.; Bartalesi, V.V.; Meghini, C.C.; Marketakis, Y.Y.; Kaplanidi, D.D.; et al. Representation of socio-historical context to support the authoring and presentation of multimodal narratives: The Mingei Online Platform. *Acm J. Comput. Cult. Herit.* **2021**, *15*, 1–26. [[CrossRef](#)]
- Partarakis, N.; Doulgeraki, V.; Karuzaki, E.; Galanakis, G.; Zabulis, X.; Meghini, C.; Bartalesi, V.; Metilli, D. A Web-Based Platform for Traditional Craft Documentation. *Multimodal Technol. Interact.* **2022**, *6*, 37. [[CrossRef](#)]
- Liddy, E.D. Natural language processing. In *Encyclopedia of Library and Information Science*, 2nd ed.; Marcel Decker, Inc.: New York, NY, USA, 2001.
- Chowdhury, G.G. Natural language processing. *Annu. Rev. Inf. Sci. Technol.* **2003**, *37*, 51–89. [[CrossRef](#)]
- Stein, R.; Coburn, E. CDWA Lite and museumdat: New developments in metadata standards for cultural heritage information. In Proceedings of the 2008 Annual Conference of CIDOC, Athens, Greece, 15–18 September 2008; pp. 15–18.
- Boughida, K.B. CDWA lite for Cataloguing Cultural Objects (CCO): A new XML schema for the cultural heritage community. In *Humanities, Computers and Cultural Heritage, Proceedings of the XVI International Conference of the Association for History and Computing, Amsterdam, The Netherlands, 14–17 September 2005*; Royal Netherlands Academy of Arts and Sciences: Amsterdam, The Netherlands, 2005; p. 49.
- Stein, R. Museumsdaten in Portalen—Die Vernetzungsstandards museumdat und museumvok. In *Informationskonzepte für die Zukunft: ODOK'07*; Neugebauer: Graz, Austria, 2008; pp. 61–69.
- Doerr, M. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Mag.* **2003**, *24*, 75.
- Goerz, G.; Scholz, M. Adaptation of nlp techniques to cultural heritage research and documentation. *J. Comput. Inf. Technol.* **2010**, *18*, 317–324. [[CrossRef](#)]
- 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*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 325–340.
- Mäkelä, E.; Hyvönen, E.; Ruotsalo, T. How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. *Semant. Web* **2012**, *3*, 85–109. [[CrossRef](#)]
- Ross, S. *Position Paper: Towards a Semantic Web for Heritage Resources*; Number 3; DigiCULT: Warsaw, Poland, 2003.
- Giaretta, D. The CASPAR approach to digital preservation. *Int. J. Digit. Curation* **2007**, *2*. [[CrossRef](#)]
- Vlachidis, A.; Bikakis, A.; Kyriaki-Manessi, D.; Triantafyllou, I.; Antoniou, A. The CrossCult Knowledge Base: A co-inhabitant of cultural heritage ontology and vocabulary classification. In *Proceedings of the European Conference on Advances in Databases and Information Systems, Nicosia, Cyprus, 24–27 September 2017*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 353–362.
- Sporleder, C. Natural language processing for cultural heritage domains. *Lang. Linguist. Compass* **2010**, *4*, 750–768. [[CrossRef](#)]
- Scholz, M.; Goerz, G. WissKI: A virtual research environment for cultural heritage. In *Proceedings of the 20th biennial European Conference on Artificial Intelligence, ECAI 2012, Montpellier, France, 27–31 August 2012*; IOS Press: Amsterdam, The Netherlands, 2012; pp. 1017–1018.
- Metilli, D.; Bartalesi, V.; Meghini, C.; Aloia, N. Populating narratives using wikidata events: An initial experiment. In *Proceedings of the Italian Research Conference on Digital Libraries, Pisa, Italy, 31 January–1 February 2018*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 159–166.
- Metilli, D.; Simi, M.; Meghini, C.; Lenzi, V.B. A Wikidata-Based Tool for the Creation of Narratives. Ph.D. Thesis, Università di Pisa, Pisa, Italy, 2016.
- Meghini, C.; Tavoni, M.; Zaccarello, M. Mapping the Knowledge of Dante Commentaries in the Digital Context: A Web Ontology Approach. *Romanic Rev.* **2021**, *112*, 138–157. [[CrossRef](#)]
- Bartalesi, V.; Meghini, C.; Metilli, D.; Tavoni, M.; Andriani, P. A web application for exploring primary sources: The DanteSources case study. *Digit. Scholarsh. Humanit.* **2018**, *33*, 705–723. [[CrossRef](#)]

23. Meghini, C.; Bartalesi, V.; Metilli, D.; Benedetti, F. A software architecture for narratives. In *Proceedings of the Italian Research Conference on Digital Libraries, Pisa, Italy, 31 January–1 February 2018*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 23–29.
24. Zabulis, X.; Partarakis, N.; Meghini, C.; Dubois, A.; Manitsaris, S.; Hauser, H.; Magnenat Thalmann, N.; Ringas, C.; Panesse, L.; Cadi, N.; et al. A Representation Protocol for Traditional Crafts. *Heritage* **2022**, *5*, 716–741. [[CrossRef](#)]
25. Meghini, C.; Bartalesi, V.; Metilli, D.; Benedetti, F. Introducing narratives in Europeana: A case study. *Int. J. Appl. Math. Comput. Sci.* **2019**, *29*, 7–16. [[CrossRef](#)]
26. Metilli, D.; Bartalesi, V.; Meghini, C. Steps Towards a System to Extract Formal Narratives from Text. In *Proceedings of the Text2StoryIR'19 Workshop, Cologne, Germany, 14 April 2019*; pp. 53–61.
27. SpaCy. Available online: <https://spacy.io/> (accessed on 22 September 2022).
28. Srinivasa-Desikan, B. *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, spaCy, and Keras*; Packt Publishing Ltd.: Birmingham, UK, 2018.
29. Vasiliev, Y. *Natural Language Processing with Python and SpaCy: A Practical Introduction*; No Starch Press: San Francisco, CA, USA, 2020.
30. Altinok, D. *Mastering spaCy: An End-to-End Practical Guide to Implementing NLP Applications Using the Python Ecosystem*; Packt Publishing Ltd.: Birmingham, UK, 2021.
31. CIDOC CRM Version 7.1.2. Available online: [https://cidoc-crm.org/html/cidoc\\_crm\\_v7.1.2.html](https://cidoc-crm.org/html/cidoc_crm_v7.1.2.html) (accessed on 22 September 2022).
32. Gergatsoulis, M.; Bountouri, L.; Gaitanou, P.; Papatheodorou, C. Mapping cultural metadata schemas to CIDOC conceptual reference model. In *Proceedings of the Hellenic Conference on Artificial Intelligence, Athens, Greece, 4–7 May 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 321–326.
33. Doerr, M. *Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM*; Technical Report; ICS/FORTH: Heraklion, Greece, 2000.
34. Theodoridou, M.; Doerr, M. *Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM*; Technical Report; ICS/FORTH: Heraklion, Greece, 2001.
35. Theodoridou, M.; Bruseker, G.; Daskalaki, M.; Doerr, M. *Methodological Tips for Mappings to CIDOC CRM*; ICS/FORTH: Heraklion, Greece, 2016.
36. Liu, B. Supervised learning. In *Web Data Mining*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 63–132.