Combining LLMs and Hundreds of Knowledge Graphs for Data Enrichment, Validation and Integration Case Study: Cultural Heritage Domain

Michalis Mountantonakis[1,2,*], Manos Koumakis[2] and Yannis Tzitzikas[1,2]

[1]Institute of Computer Science, FORTH, Heraklion, Greece
[2]Department of Computer Science, University of Crete, Heraklion, Greece

Abstract

Recently, there is a high trend for exploiting Large Language Models (LLMs) and Knowledge Graphs (KGS) for providing better access services and analytics, and for aiding user experience for the Cultural Heritage (CH) domain. In this direction, we discuss the corresponding challenges and potential use cases for both simple users and data owners for the CH domain. Afterwards, we present the research prototype GPToLODS+ (and its services), which offers large scale knowledge services that exploit the capabilities of hundreds of KGs and LLMs for several domains (including the CH one). In particular, it combines ChatGPT, LODsyndesis KG (that aggregates hundreds of RDF KGs) and Entity Recognition tools, for offering the following functionality: a) Question Answering using ChatGPT and hundreds of RDF KGs, b) Entity Recognition, Linking and Enrichment over the ChatGPT responses (or any given plain text) using LODsyndesis, c) Fact Validation of ChatGPT responses (or web texts) with provenance using LODsyndesis, d) Connectivity Analytics and Integration with existing RDF KGs at real time through LODChain service, e.g., for aiding data publishing and discoverability, and others. Finally, we provide dedicated examples over the CH domain for the use cases by exploiting GPToLODS+ and its services using real entities and CH KGs, by mainly focusing on CIDOC-CRM based KGs.

Keywords

RDF Knowledge Graphs, LLMs, Cultural Heritage, Fact Checking, Discoverability, Reusability

1. Introduction

There is a high proliferation of using Large Language Models (LLMs) for aiding several tasks of any domain, including Cultural Heritage (CH), such as Question Answering [1], Digital Storytelling [2], Entity Recognition and Enrichment [3], and many others [4]. For the CH domain, the ultimate target is to enhance the user experience through access services like interactive Artificial Intelligence (AI) chatboxes and web applications [5, 2], i.e., for aiding users to find resources on the websites, for offering digital storytelling into museum visits and others. LLMs can be of primary importance for the CH domain, since they have been trained from web texts including CH data (see the left side of Fig. 1), they are very creative, they offer human-like expressiveness and are suitable for the mentioned tasks. However, LLMs do not provide justifications for the responses, and they are vulnerable to hallucinations, including erroneous and outdated facts [3].

On the contrary, there are available thousands of Knowledge Graphs (KGs) having structured data with high correctness and information about their provenance [6], including numerous KGs for the CH domain (right side of Fig. 1). For instance, see a list of such KGs in Linked Open Data (LOD) Cloud [7] and in a portal [8] for KGs using the ISO standard CIDOC-CRM model [9]. However, a disadvantage is that it is not trivial to ask questions over such KGs, i.e., it is required either to be familiar with Linked Data technologies and SPARQL query language [10], or to provide user-friendly interfaces and access services for aiding the browsing from simple users. As it is shown in Fig. 1 the ultimate challenge is how to combine the advantages of both LLMs and KGs, for providing better access services and analytics and improving user experience (lower side of Fig. 1).

MBD2024: International Conference On Museum Big Data, November 18–19, 2024, Athens, Greece
mountant@ics.forth.gr (M. Mountantonakis); csd4281@csd.uoc.gr (M. Koumakis); tzitzik@ics.forth.gr (Y. Tzitzikas)
0000-0002-1951-0241 (M. Mountantonakis); 0009-0008-1441-8914 (M. Koumakis); 0000-0001-8847-2130 (Y. Tzitzikas)

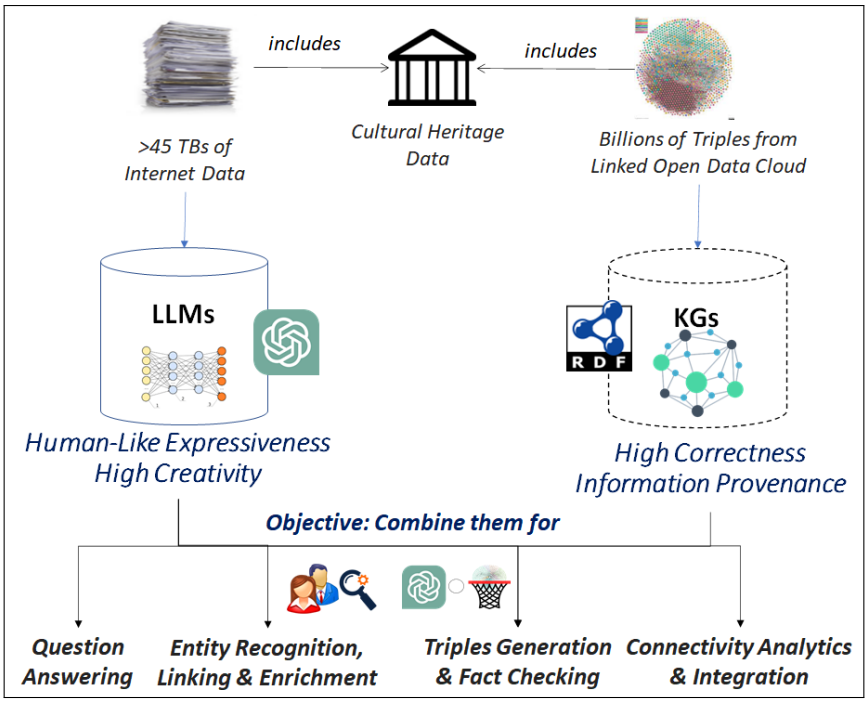

Figure 1: Vision: Combining LLMs and KGs over CH data

Below, we discuss the challenges concerning the CH domain for LLMs and KGs (see the right side of Fig. 2). First of all, we distinguish two types of users, a) simple users, i.e., visitors of museums, web users of digital libraries, etc., and b) data owners (of CH domain), i.e., responsible persons of a museum, a digital library, etc. Regarding the challenges, for any type of users i) it is not trivial to find more information (with provenance) about the entities of interest of an LLM response (e.g., of CH domain), since the LLMs are not directly connected with the several existing KGs (including KGs of the CH domain), and ii) it is difficult to validate (CH) facts from ChatGPT responses or from web texts, since evidence and provenance are not always provided. Moreover, iii) it is time consuming for a data publisher to generate a KG (from text) and it is difficult to discover relevant KGs (or datasets), given the high number of available KGs and the several data integration problems that should be faced [11], for providing a KG that will be connected with existing high quality KGs of the LOD Cloud. This problem exists even iv) with KGs of a specific domain that use the same model, such as CIDOC-CRM [9].

Towards this objective, we present the research prototype GPToLODS+ (and its underlying services), which provides both a user-friendly web application and a REST API, for enabling several services (which are related to the presented challenges): a) Entity Recognition, Linking and Enrichment for the entities found in the ChatGPT [12] responses through a dialogue-based user interface, b) Fact Validation over the ChatGPT responses with provenance from 400 RDF KGs from the LODsyndesis aggregated KG [13], c) KG generation from text (either from a ChatGPT response or plain text) and d) Connectivity Analytics and Integration with real RDF datasets for the generated KG. Concerning our contribution, we focus on providing use cases for the CH domain for both simple users and data owners, we present the services of GPToLODS+ and we provide real examples for the use cases by using GPToLODS+. Finally, we present services for CIDOC-CRM based KGs, by presenting a real example with connectivity analytics. As regards the novelty, to the best of our knowledge, there is not any other suite of services offering a dialogue-based user interface and a REST API that exploits hundreds of RDF KGs for enriching and validating ChatGPT responses.

The rest of this paper is organized as follows: §2 presents the desired use cases for the CH domain for both users and publishers. §3 discusses the related work, whereas §4 presents the steps and services of GPToLODs+. Finally, §5 shows how to perform the desired use cases using the presented services and §6 concludes the paper and discusses future directions.

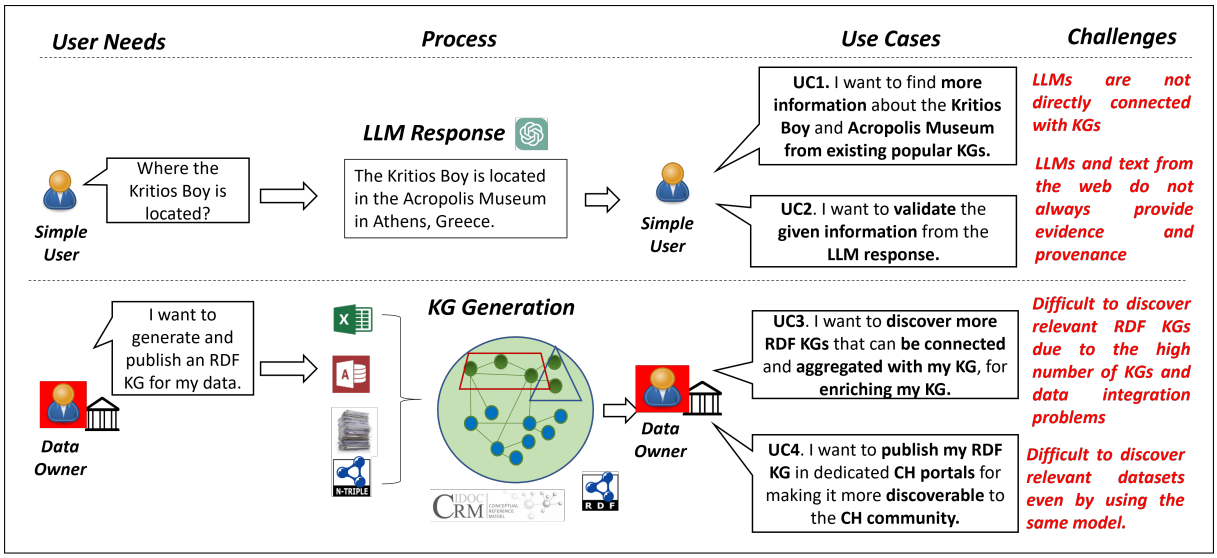

Figure 2: Use Cases related to CH domain for LLMs and KGs

2. Use Cases over Cultural Heritage (for LLM and KGs)

Based on the presented challenges, we provide four use cases, two for each type of users, which are covered by the GPToLODS+.

UC1. Browsing more Information about the Entities of Interest from KGs. As it is shown in the upper side of Fig. 2, the user asks an LLM about the Kritios Boy sculpture[1] and retrieves a response from the LLM. However, the user desires to find more facts, links (URIs) and KGs about it and the museum where it belongs to (see UC1 in the upper right side of Fig. 2).

UC2. Fact Validation (over the LLM response or a web text). Here, the same user (see UC2 in Fig. 2) desires to check the validity of the fact(s) returned by the LLM (since provenance usually is not given [3]), whereas the same user can also desire to check the validity of facts (e.g., related to CH domain) from texts of web pages.

UC3. KG Generation, Connectivity Analytics and Data Integration. Here, the data owner (e.g., of a museum or a digital library) first desires to create an RDF KG. In many cases, the publisher needs to combine more than one files in many formats to create the KG, e.g., CSV files, text files, existing RDF triples and others, and then to integrate all of them by using a given standard such as CIDOC-CRM [9]. Afterwards, the data publisher desires to discover more KGs (or datasets) containing complementary information about the same entities (see UC3 in Fig. 2), i.e., to create an enriched version of his/her KG connected with other high quality KGs. This will enable the execution of more complex queries and will make the dataset more discoverable and reusable for other users and publishers.

UC4. Data Publishing to CH portals. By having created the RDF KG, the data owner desires to publish the KG into dedicated CH portals, for improving its discoverability over the CH community, and for finding even more analytics with relevant KGs (see UC4 in Fig. 2).

3. Related Work

Here, we focus on approaches using CH data with KGs (see §3.1) and LLMs (see §3.2), whereas we provide a comparison with related approaches (see §3.3).

[1] https://www.theacropolismuseum.gr/en/youth-statue-kritios-boy

3.1. Cultural Heritage and KGs

CH is one of the most successful application domains of the Semantic Web technologies [14], i.e., there are numerous KGs about CH, and many of them can be listed in online catalogs and portals, such as the LOD Cloud [7] and the CIDOC-CRM portal [8] (containing KGs that have been modeled through the ISO standard CIDOC-CRM). The KGs of the CH domain include museums [15], digital libraries [16], historical archives [17], archaeological excavation [18] and others. Indicatively, the CIDOC-CRM portal [8] includes 30 KGs from the CH domain having more than 500 million RDF triples. The ultimate target of creating such KGs is to provide better access services to the users and analytics, including Browsing systems [19], Question Answering [20, 21], Virtual Exhibitions to museums [14] and Digital Storytelling [2], whereas the constructed KGs can be very useful for Digital Humanities research [22].

3.2. Exploiting LLMs for the CH domain

Similarly to the case of KGs, one of the ultimate challenges of using LLMs in the CH domain is to aid the user experience [23], e.g., museum visitors, users that browse digital libraries, and others. First, [24] surveys approaches that apply Machine Learning and Artificial Intelligence techniques for reducing the costs related to the compliance and interoperability of Cultural Heritage KGs, by focusing on CIDOC-CRM based KGs. Moreover, the authors in [25] studied the problem of connecting ArtGraph KG and Wikidata through the aid of LLMs (and specifically LLaMa), for improving Entity Alignment to enhance entity enrichment for ArtGraph KG. Furthermore, a context-aware visual QA system based on multi-modal LLMs is presented in [1], and its target is to offer accurate answers to questions of CH domain by also providing as input to the LLM the associated KG and corresponding images. In [26], the authors presented an LLM-based virtual art guide, for enabling users to express inquiries about the displayed artwork. Finally, [2] presents the system MAGICAL, which is a digital tour guide in museums and it exploits the capabilities of GPT-4 for composing texts and dialogues with the visitor.

3.3. Comparison with Related Work

Concerning the applications exploiting LLMs, KGs, or both of them, we mainly focus on combining hundreds of KGs and LLMs for offering Entity Enrichment, Fact Checking and Connectivity Analytics over LLM responses and web texts, and not on providing applications such as virtual exhibitions and QA systems [14, 1, 20, 21]. However, it is worth noting that the research prototype that we present can be possibly used by such systems, since it provides a REST API for enabling the reusing of services programmatically.

Compared to our previous work [3], here we give emphasis on how to combine LLMs and KGs for the CH domain, by discussing challenges and by providing use cases and dedicated examples for different types of users. Also, we provide a solution for these challenges by presenting the capabilities of the up-to-date version of the GPToLODS+ prototype, which i) offers a new dialogue-based user interface where all the services are accessible on the same page for both LLM responses and web texts, ii) can be connected to LODChain [27] for having access to connectivity analytics over the LOD Cloud, and iii) offers a REST API for aiding users to reuse the services in their (CH) applications. To the best of our knowledge, this is the first suite of tools offering such services by exploiting hundreds of RDF KGs over LLM responses.

4. The Services of GPToLODS+

We present the steps and services of GPToLODS+, which are shown in Figure 3. The user can either add as an input a question to ChatGPT, or a plain text. For the first case, the user can send a question through a dialogue interface, and then GPToLODS+ sends a request to ChatGPT API and retrieves the textual response. For the second case, the user can just add any text (e.g., found on the web). Afterwards, the user is able to exploit the following services over the desired text: either A) Entity Recognition

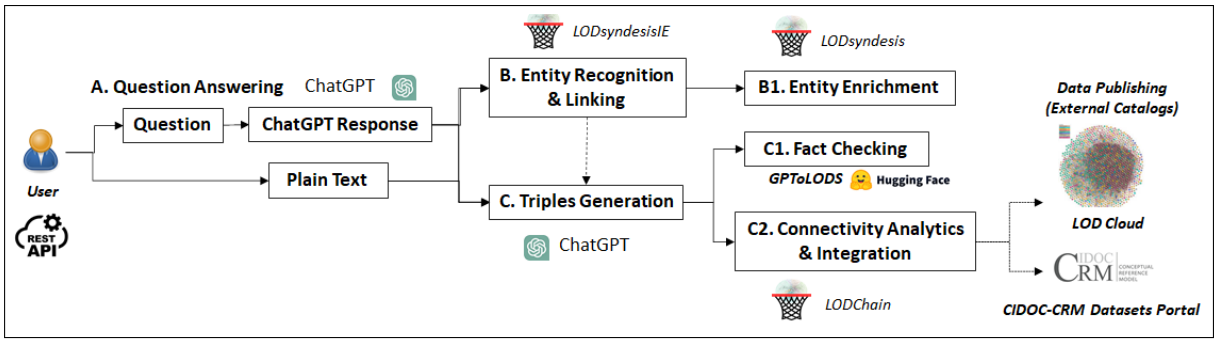

Figure 3: Services offered by GPToLODS+, and its underlying services

and Linking through LODsyndesisIE, for enabling the enrichment of all the entities of the text, or to perform B) Triples Generation, for retrieving the facts of the text in RDF format. The latter enables the service to check the validity of the facts found in the text, and the connectivity analytics and integration service through LODChain.

Service	Input	Output	Tools Used	Data Provenance	Data Volume
A. Question Answering (Dialogue)	User Questions in natural language	ChatGPT response	ChatGPT	Web Pages, Wikipedia, etc.	45TB raw data
B. Entity Recognition & Linking	ChatGPT response or plain text	Entities URIs from KGs, Enriched Version of Text	LODsyndesisIE	DBpedia, LODsyndesis	Billions of RDF triples
B1. Entity Enrichment	Entities URIs from KGs (entities from ChatGPT response or plain text)	More URIs, Datasets and Facts (with Provenance)	LODsyndesis	LODsyndesis KG (based on Semantics-aware indexes)	2 billion RDF triples
C. Triples (KG) Generation from text	ChatGPT response or plain text (+optionally enhanced with URIs)	RDF Triples (from ChatGPT)	ChatGPT (+optionally LODsyndesisIE)	ChatGPT	45TB raw data
C1. Fact Checking	RDF Triples (e.g., generated from ChatGPT)	Corresponding RDF triples with provenance from RDF KGs	GPToLODs Fact Checking Service	DBpedia (current version), LODsyndesis KG	>2 billion RDF triples
C2. Connectivity Analytics and Integration	RDF KG	Visualizations, Statistics, Enriched version of the input KG	LODChain	LODsyndesis KG	2 billion RDF triples

Table 1
An overview of the Services offered by GPToLODS+ and the underlying tools

4.1. The Core Services

Below, we provide more details for each of the services and the underlying tools, which are also listed in Table 1 and are depicted in Fig. 3. More detailed examples are given in §5.

A. Question Answering (through Dialogue). This functionality is offered through a dialogue interface. In particular the user can type a question in natural language, which is then sent to ChatGPT (current version used is v3.5) to retrieve the response. The key notion is that the resulted ChatGPT response can be used by the services described below.

B. Entity Recognition and Linking. Here, the input can be any plain text, e.g., the ChatGPT

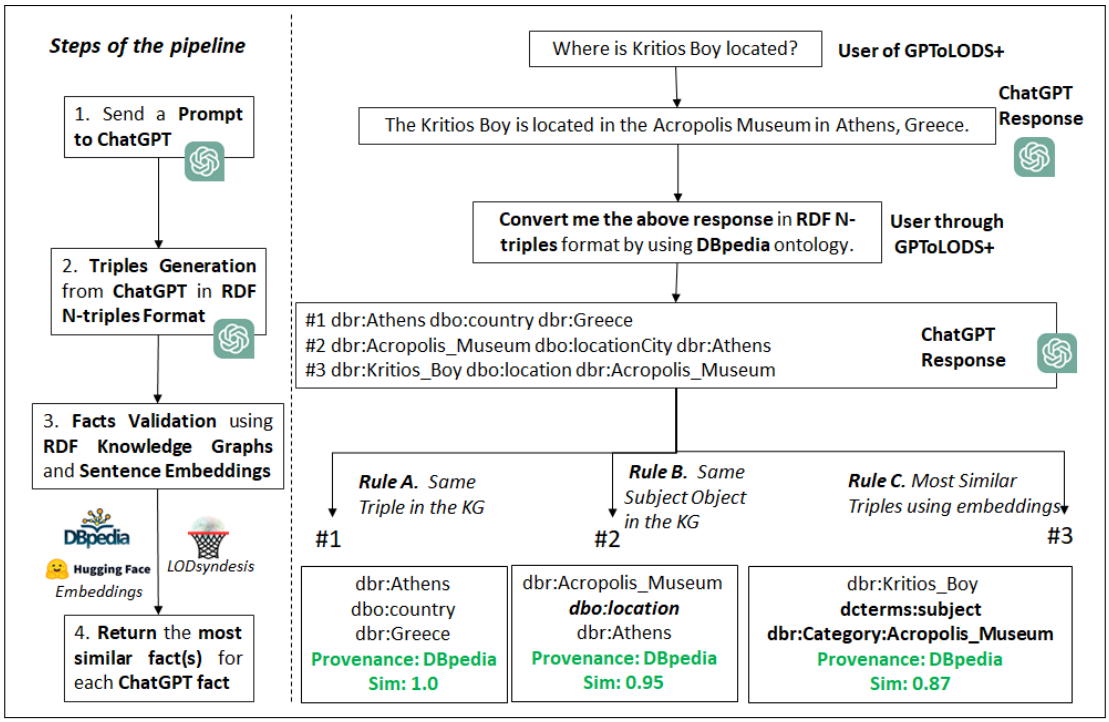

Figure 4: The pipeline of the fact validation service and the three different rules

response or any text from the web (in English), i.e., see Fig. 3. Afterwards, the user can exploit the Entity Recognition and Linking Service of LODSyndesisIE [28]. In particular, the mentioned tool can use any combination of three popular Entity Recognition tools, DBpedia Spotlight [29], WAT [30] and Stanford CoreNLP [31] for identifying the entities of a given text and for providing a unique URI (or link) for each entity to DBpedia [32] and LODsyndesis KG [13]. Afterwards, the entities of the text (e.g., of the ChatGPT response) are marked and the user can browse more information about each of them, as it is explained below.

B1. Entity Enrichment. The next step is to perform Entity Enrichment for the entities of the given text. In particular, the user can click on any recognized entity to retrieve more information about it from LODsyndesis [13], which is an Aggregated KG where the contents of 400 LOD datasets have been aggregated, after having computed the transitive and symmetric closure of 45 million equivalence relationships. For providing fast access to all the URIs, KGs and facts (with provenance) for each entity, it offers semantics-aware indexes and services for over 412 million entities and 2 billion triples. In this way, the user can have access to (and export) all this information for the selected entities of interest.

C. Triples (KG) Generation from text. The user can further process the text to create an RDF representation for the facts of the text, i.e., by generating RDF triples from text. This is an important step for further enabling the combination of an LLM response and KGs, i.e., for having the information for both of them in the same data format. In GPToLODS+, this can be done by sending a request to ChatGPT to convert the text into RDF triples, e.g., "give me the RDF N-triples using DBpedia format for the text T". For aiding ChatGPT to provide valid URIs for the entities of the desired text, we can optionally provide the list of URIs that were recognized through the Entity Recognition and Linking process. The result is a set of RDF triples (or facts). The ultimate target is the RDF facts to be exploited for performing real time fact validation with provenance and for offering advanced connectivity analytics for the generated KG, i.e., services C1 and C2.

C1. Fact Validation Service. The user can select any of the generated facts for validation by exploiting RDF KGs, either DBpedia [32] or LODsyndesis KG [13]. Afterwards, a specialized algorithm

(proposed in [3]) that uses both SPARQL queries and word embeddings from the “all-MiniLM-L6-v2" library[2], is exploited for detecting the K most similar facts to the desired fact. Finally, the top-K corresponding facts and their provenance are returned to the user along with a similarity score, for enabling the validation of the given fact.

The whole pipeline is depicted in Fig. 4. In particular, we can see that after generating the triples for the facts of ChatGPT response (middle part of Fig. 4), we have three different facts for validation. The algorithm uses three different rules. First, Rule A checks for finding exactly the same triple in the KG, e.g., for the fact #1, we found the same triple in DBpedia. Second, if Rule A fails, then Rule B checks for finding either the same subject-object or the same subject-predicate, e.g., for the fact #2, we have the same subject object (Acropolis Museum and Athens) with a different predicate (location instead of locationCity). Finally, if both rules fail, Rule C checks for the most similar RDF triples, e.g., see the example for the fact with ID #3, where the algorithm found the most similar triples by computing the cosine similarity of the embeddings.

C2. Connectivity Analytics and Integration. By having generated the RDF KG for a given text, the user can connect to LODChain [27]. This is a research prototype that enables the connection of a new or an existing RDF KG to the 400 RDF KGs of LODsyndesis for ensuring its connectivity, for fixing possible connectivity errors, and for enriching its contents by discovering related datasets. In particular, it computes at real time the transitive and symmetric closure of equivalence relationships between the given KG and the 400 RDF KGs of LODsyndesis, for discovering new connections for the KG. This functionality is offered through a user interface with connectivity analytics, visualizations and options to download the enriched data. As a result, the user can export an enriched version of the KG, with URIs to existing RDF KGs, or/and complementary facts. The enriched version can be published to the LOD Cloud, for aiding its discoverability and reusability from other publishers, and for offering advanced query services.

4.2. External CH Publishing Services (for a generated CIDOC-CRM based KG)

If the generated KG (or any KG) has been created using the CIDOC-CRM model, one option is to also publish it to the CIDOC-CRM portal [8], which offers several statistics and measurements for CIDOC-CRM based KGs. This can be of primary importance for several tasks including, i) Dataset Discovery and Selection, i.e., for enabling the discovery of the KG from interested users of the CH community, ii) Data Integration, i.e., for integrate the KG with existing datasets that use common CIDOC-CRM properties and classes for enriching their information, and iii) Ontology Evaluation, i.e, for detecting possible problems more easily (e.g., using the ontology in a wrong way).

4.3. How GPToLODS+ can be accessed

GPToLODS+ is available online[3] and offers the mentioned real time interactive services (see also a tutorial video[4]). It runs on a server with 8 GB main memory, 8 cores and 64 GB disc space. Moreover, a REST API is offered for most of the services (except for C2), for making it feasible to exploit the services programmatically, e.g., for integrating the offered services into external services. The REST API is available online[5], where one can find guidelines of how to use each of the offered services. Finally the code is available in GitHub[6].

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[3] https://demos.isl.ics.forth.gr/GPToLODS
[4] https://youtu.be/cE57RqHbDt8
[5] https://demos.isl.ics.forth.gr/GPToLODS/GPToLODSplusREST
[6] https://github.com/mountanton/GPT-LODs_plus

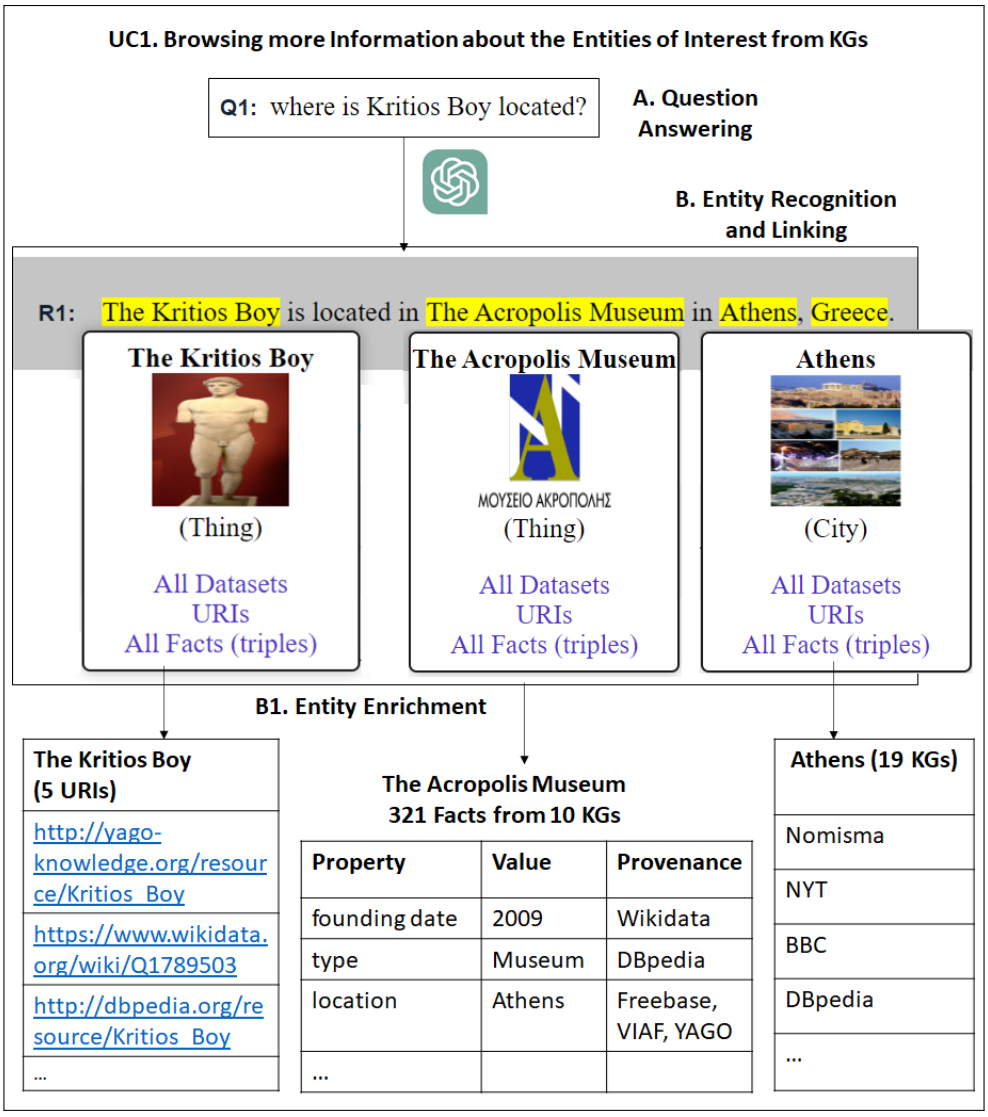

Figure 5: Use Case 1: Browsing more Information about the Entities of Interest from KGs, using GPToLODs+

5. The Use Cases using GPToLODS+ & Connectivity Analytics

Here, we present how to perform the use cases of §2 through the services of GPToLODS+ (i.e., §4). We provide scenarios with real examples and Kerameikos KG [33], which is a KG about Ceramics of Ancient Greece, including connectivity analytics and statistics for the mentioned KG.

5.1. UC1: Browsing more Information about the Entities of Interest from KGs

As we can see in Fig. 5, the user asks GPToLODS+ for the location of "Kritios Boy Sculpture". The answer is retrieved from ChatGPT (i.e., Acropolis museum in Athens) and then the entities of the text are recognized and marked from the LODsyndesisIE Entity Recognition and Linking Service. Then, the user selects to discover more information about the entities of the text. In that example (real data from LODsyndesis are presented), the user discovered all the URIs for "Kritios Boy" in LODsyndesis (5 URIs in total), all the facts for the Acropolis Museum (321 facts from 10 KGs in total), and all the KGs containing information about Athens (19 KGs in total).

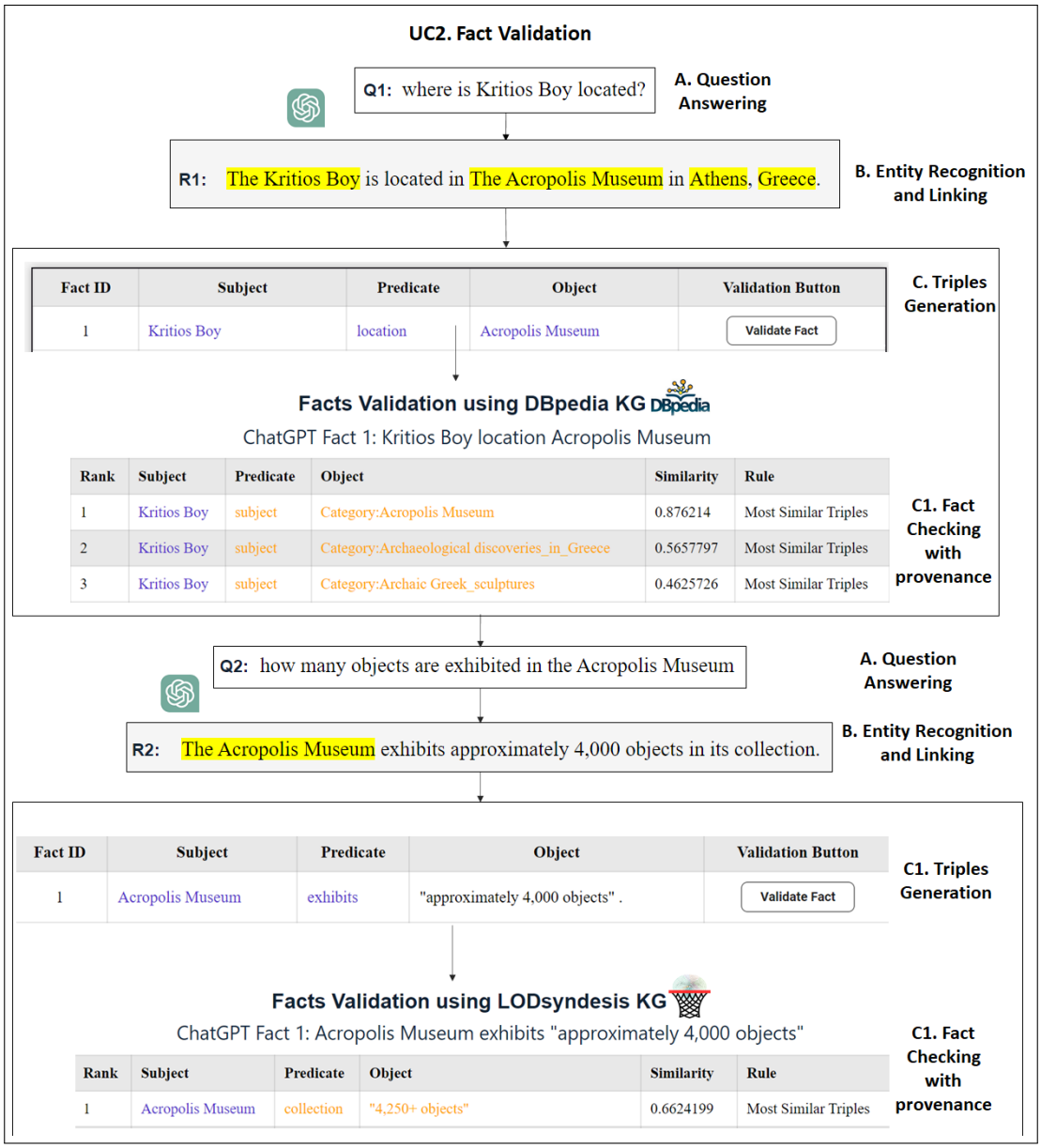

Figure 6: Use Case 2: Fact Validation over a dialogue between a user and ChatGPT

5.2. UC2: Fact Validation (over the LLM response or a web text)

Fig. 6 shows a dialogue between a user and GPToLODS+. After the Entity Recognition process, the facts of the text can be converted to triples and then can be validated by using DBpedia or LODsyndesis KG. As we can see, for the first question (location of the sculpture), we managed to confirm that it is located in the Acropolis Museum. Although we found a slightly different RDF triple in DBpedia KG (than the one produced by ChatGPT), they had a high similarity score. Afterwards, the user continued the dialogue and asked about the number of objects exhibited in the museum, and ChatGPT returned "approximately 4,000 objects". By performing the same steps, we found a more accurate answer in LODsyndesis KG, i.e., "4,250+ objects". The dialogue can be continued with as many as questions the user wants, and can return to any previous question for using all the mentioned services.

5.3. UC3: KG Generation, Connectivity Analytics and Data Integration

We suppose that the data owner has generated an RDF KG, by connecting different pieces of information, e.g., a part (or the whole KG) can be generated by converting texts to an RDF KG through the GPToLODS+.

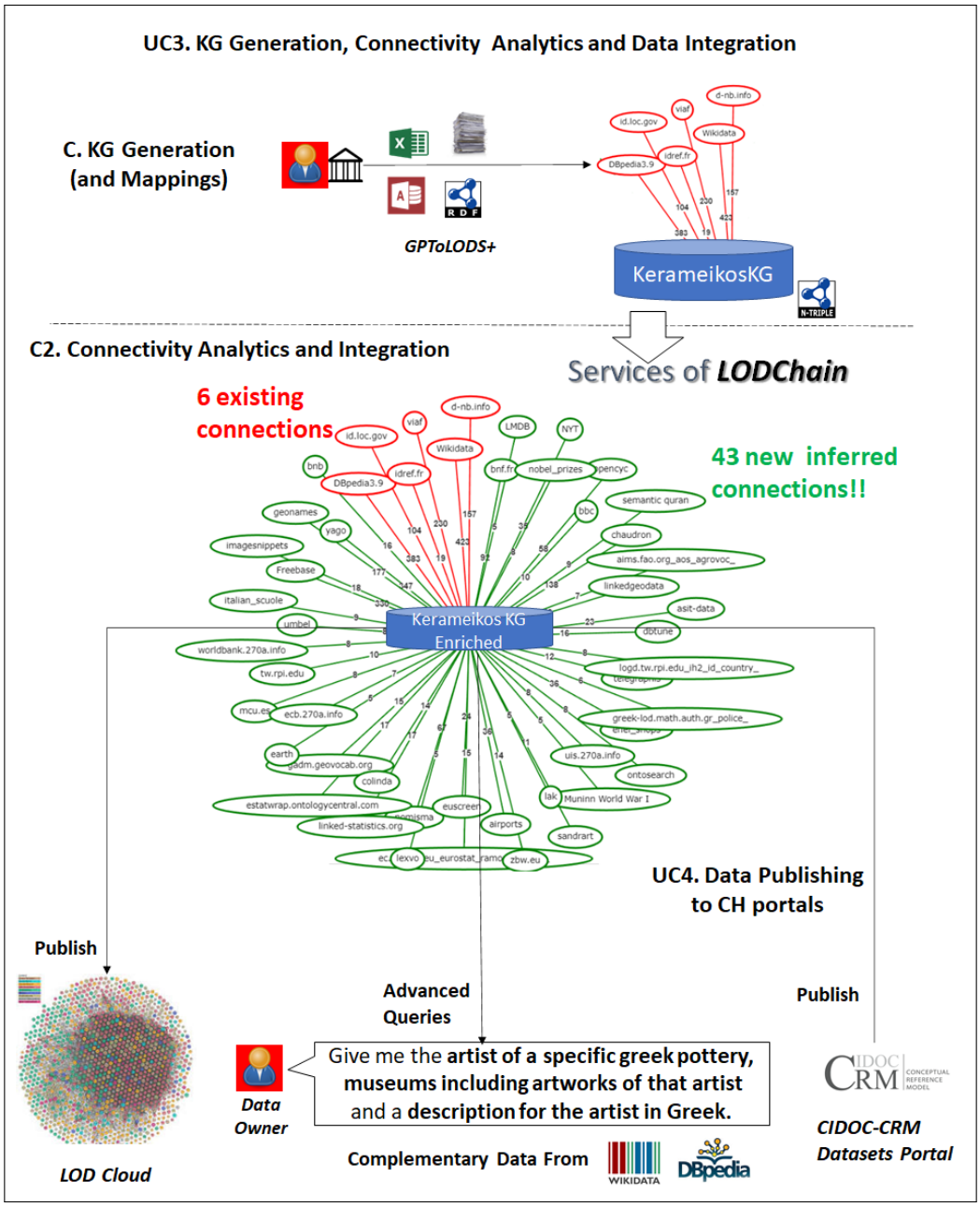

Figure 7: Use Cases 3 and 4 KG Generation, Connectivity Analytics, Data Integration and Publishing

For representing real measurements, here we use a real KG from the CH domain, called Kerameikos KG [33], which contains 289,596 triples about ceramic data of Ancient Greece. The data owner has created mappings (i.e., links) with 6 external LOD KGs (see the upper side of Fig. 7), whereas the KG has been modelled using the CIDOC-CRM standard.

Afterwards, the data owner uses the LODChain service for connecting and integrating the KG with more RDF KGs of the LOD Cloud, and as Fig. 7 shows, 43 more connections with RDF KGs were discovered, i.e., the nodes with green color represent the new connections (KGs) for Kerameikos KG. The data owner can export equivalent URIs, triples, complementary facts and others for publishing a more enriched KG in the LOD Cloud (lower left side of Fig. 7). In this way, more advanced queries could be expressed such as, “Give me the artist of a specific greek pottery, other museums including artworks of that artist and a description for the artist in Greek language" (lower side of Fig. 7). The first part (artist of the pottery) can be answered by Kerameikos KG, the second part (museums including artworks of the artist) from Wikidata KG [34] and the third one (description in Greek language) from DBpedia KG [32].

Measurement	Value
# of unique entities	21,872
# of common entities	645
KG with most common entities	Wikidata
# of connections before LODChain	6
# of inferred connections	43
# of connections after LODChain	49
# of connections with CH KGs before LODChain	4
# of connections with CH KGs after LODChain	12
# of complementary facts for Kerameikos entities	843,283
KG with most complementary facts	Wikidata

Table 2: Connectivity Analytics for Kerameikos KG using LODChain

Measurement	Value
Triples	289,596
Properties	85
CIDOC-CRM Properties	22
Classes	41
CIDOC-CRM Classes	15
Triples with CIDOC-CRM properties	162,847
Triples With CIDOC-CRM entities	269,640
KG sharing most CIDOC-CRM properties	Getty
KG sharing most CIDOC-CRM classes	LINCS

Table 3: Statistics for Kerameikos KG using CIDOC-CRM Portal

Table 2 shows some statistics for the connectivity of Kerameikos KG derived by the LODChain Service; indicatively we discovered 645 entities (of Kerameikos KG) that can be also found in at least one other RDF KG, whereas the data owner can enrich his/her dataset with 843,283 more facts for the entities that also exist in Kerameikos KG. Finally, we can see that the KG that can offer the most complementary information for the entities of Kerameikos KG is Wikidata.

5.4. UC4: Data Publishing to CH portals

By having generated the RDF KG and (optionally) published it to the LOD Cloud, one has also the option to connect it to the CIDOC-CRM portal [8] where ontology statistics/visualizations are offered based on VoID vocabulary. For instance, see the lower right side of Fig. 7, and also some statistics about Kerameikos KG derived from that portal in Table 3. Indicatively, we can see some dedicated statistics about the desired KG and CIDOC-CRM model, such as the number of CIDOC-CRM properties and classes, in how many RDF triples CIDOC-CRM properties and entities are used and which KGs share the most CIDOC-CRM properties and classes with Kerameikos KG. More analytics (for the Kerameikos and many other CH KGs) can be browsed in the website of the portal [8].

6. Concluding Remarks

Since there is a high need to exploit and combine LLMs and KGs for aiding the user experience over Cultural Heritage (CH) information, i.e., by providing advanced access services and analytics, we presented a research prototype, called GPToLODS+, that tries to combine the advantages of both LLMs and KGs for achieving that target. We presented the challenges and several use cases over the CH domain and all the services of GPToLODS+, including Entity Recognition, Linking, and Enrichment, Fact Validation, Connectivity Analytics and Data Integration and others. Afterwards, we provided real examples for each of the use cases by using the mentioned services, including connectivity analytics for a real KG from the CH domain, by also focusing on CIDOC-CRM standard. As regards the future work, the plan is to extend the services for covering more applications that can combine KGs and LLMs, such as a) converting natural questions to SPARQL queries over RDF KGs by exploiting LLMs, b) to provide and evaluate (e.g., through a task based evaluation with users) the offered services by using more LLMs, such as LLaMA or different versions of ChatGPT, and c) to create digital storytelling applications of CH data from RDF KGs by proposing and evaluating different LLM prompts.

Acknowledgments

This work has received funding from ECHOES, a project funded by the European Commission under Grant Agreement n.101157364. Views and opinions expressed in this paper are those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] P. K. Rachabatuni, F. Principi, P. Mazzanti, M. Bertini, Context-aware chatbot using MLLMs for cultural heritage, in: Proceedings of the 15th ACM Multimedia Systems Conference, 2024, pp. 459–463.
[2] G. Trichopoulos, Large language models for cultural heritage, in: Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter, 2023, pp. 1–5.
[3] M. Mountantonakis, Y. Tzitzikas, Real-time validation of ChatGPT facts using RDF knowledge graphs., in: ISWC (Posters/Demos/Industry), 2023.
[4] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of LLMs in practice: A survey on chatgpt and beyond, ACM Transactions on Knowledge Discovery from Data 18 (2024) 1–32.
[5] M. Casillo, F. Colace, B. B. Gupta, A. Lorusso, D. Santaniello, C. Valentino, The role of AI in improving interaction with cultural heritage: An overview, Handbook of Research on AI and ML for Intelligent Machines and Systems (2024) 107–136.
[6] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Computing Surveys (Csur) 54 (2021) 1–37.
[7] LOD cloud, https://lod-cloud.net/, 2024 (accessed July 7, 2024).
[8] M. Mountantonakis, I. Theocharakis, Y. Tzitzikas, Why we need ontology-specific data portals: A case study for CIDOC-CRM., in: SWODCH, 2023.
[9] M. Doerr, The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata, AI magazine 24 (2003) 75–75.
[10] W. W. W. Consortium, et al., Sparql 1.1 overview (2013).
[11] M. Mountantonakis, Y. Tzitzikas, Large-scale semantic integration of linked data: A survey, ACM Computing Surveys (CSUR) 52 (2019) 1–40.
[12] OpenAI, ChatGPT, https://openai.com/, 2021. Accessed: March 2024.
[13] M. Mountantonakis, Y. Tzitzikas, LODsyndesis: global scale knowledge services, Heritage 1 (2018) 23.
[14] D. Monaco, M. A. Pellegrino, V. Scarano, L. Vicidomini, Linked open data in authoring virtual exhibitions, Journal of Cultural Heritage 53 (2022) 127–142.
[15] P. Szekely, C. A. Knoblock, F. Yang, E. E. Fink, S. Gupta, R. Allen, G. Goodlander, Publishing the data of the smithsonian american art museum to the linked data cloud, International Journal of Humanities and Arts Computing 8 (2014) 152–166.
[16] D. Metilli, V. Bartalesi, C. Meghini, et al., Steps towards a system to extract formal narratives from text., in: Text2Story@ ECIR, 2019, pp. 53–61.
[17] I. Koch, C. Ribeiro, C. Teixeira Lopes, Archonto, a CIDOC-CRM-based linked data model for the portuguese archives, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2020, pp. 133–146.
[18] M. Katsianis, G. Bruseker, D. Nenova, O. Marlet, F. Hivert, G. Hiebel, C.-E. S. Ore, P. Derudas,

R. Opitz, E. Uleberg, Semantic modelling of archaeological excavation data. a review of the current state of the art and a roadmap of activities, Internet Archaeology (2023).
[19] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A full stack javascript framework for developing semantic portal user interfaces, Semantic Web 13 (2022) 69–84.
[20] N. Gounakis, M. Mountantonakis, Y. Tzitzikas, Evaluating a radius-based pipeline for question answering over cultural (CIDOC-CRM based) knowledge graphs, in: Proceedings of the 34th ACM Conference on Hypertext and Social Media, 2023, pp. 1–10.
[21] O. Suissa, M. Zhitomirsky-Geffet, A. Elmalech, Question answering with deep neural networks for semi-structured heterogeneous genealogical knowledge graphs, Semantic Web 14 (2023) 209–237.
[22] A. Ahola, E. Hyvönen, A. Kauppala, Publishing and studying historical opera and music theatre performances on the semantic web: case operasampo 1830–1960, in: International Workshop on Semantic Web and Ontology Design for Cultural Heritage, CEUR-WS. org, 2023, p. 12.
[23] G. Trichopoulos, G. Alexandridis, G. Caridakis, A survey on computational and emergent digital storytelling, Heritage 6 (2023) 1227–1263.
[24] Y. Tzitzikas, M. Mountantonakis, P. Fafalios, Y. Marketakis, CIDOC-CRM and machine learning: a survey and future research, Heritage 5 (2022) 1612–1636.
[25] A. S. Lippolis, A. Klironomos, D. F. Milon-Flores, H. Zheng, A. Jouglar, E. Norouzi, A. Hogan, et al., Enhancing entity alignment between wikidata and artgraph using LLMs., in: SWODCH, 2023.
[26] N. Constantinides, A. Constantinides, D. Koukopoulos, C. Fidas, M. Belk, CulturAI: Exploring mixed reality art exhibitions with large language models for personalized immersive experiences, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 102–105.
[27] M. Mountantonakis, Y. Tzitzikas, LODChain: Strengthen the connectivity of your RDF dataset to the rest LOD Cloud, in: ISWC, Springer, 2022, pp. 537–555.
[28] M. Mountantonakis, Y. Tzitzikas, LodsyndesisIE: Entity extraction from text and enrichment using hundreds of linked datasets, in: European Semantic Web Conference, Springer, 2020, pp. 168–174.
[29] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia spotlight: shedding light on the web of documents, in: Proceedings of the 7th international conference on semantic systems, 2011, pp. 1–8.
[30] F. Piccinno, P. Ferragina, From TagME to WAT: a new entity annotator, in: Proceedings of the first international workshop on Entity recognition & disambiguation, 2014, pp. 55–62.
[31] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky, The stanford coreNLP natural language processing toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.
[32] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., DBpedia–a large-scale, multilingual knowledge base extracted from wikipedia, Semantic web 6 (2015) 167–195.
[33] Kerameikos.org, http://kerameikos.org/, 2021. Accessed: March 2024.
[34] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.