

Why We Need Ontology-Specific Data Portals: A Case Study for CIDOC-CRM

Michalis Mountantonakis^{1,2,*}, Ioannis Theocharakis² and Yannis Tzitzikas^{1,2}

¹*Institute of Computer Science, FORTH, Heraklion, Greece*

²*Department of Computer Science, University of Crete, Heraklion, Greece*

Abstract

There are several ways to publish data on the web, from plain or GitHub web pages, to linked open data, and online catalogs. However, since each data owner can select a different way to publish his data, it is challenging to discover all the available datasets that are represented through a particular ontology *O*. The rationale for building a catalog for datasets expressed with respect to a particular ontology *O* (or specializations of *O*), is that the community which is interested in that ontology has various incentives to publish its datasets in such a catalog; for inspecting the use of the ontology (for spotting errors and/or for guiding the use and evolution of the ontology), for finding other datasets that could be easily integrated, and others. In this paper we focus on such catalogs, and in particular we showcase a catalog for datasets expressed with respect to CIDOC-CRM. Even if there are dozens of datasets that are represented using CIDOC-CRM, there is not any online resource that contains even a simple textual list of all these datasets. To fill this gap, we present an interactive portal that contains ontology-based descriptions for 30 CIDOC-CRM datasets. Through this portal, the user can browse all CIDOC-CRM datasets (and their statistics), can find all datasets that use a particular CIDOC-CRM property/class, can see the most frequent properties and classes, can check the commonalities between different datasets, and can enrich the catalog with new datasets. Finally, we provide indicative measurements over the 30 collected CIDOC-CRM datasets and for the properties/classes of the current version of CIDOC-CRM.

Keywords

CIDOC-CRM, Visualizations, Statistics, Data Discovery, Data Integration, Cultural Heritage


1. Introduction


There are numerous ways for publishing data on the web, such as in a web page, in GitHub, in Zenodo, as Linked Open Data (e.g., in a SPARQL endpoint or through a data dump) or/and by uploading them to an online catalog, which can offer more services comparing to the previous ones. In particular, *catalogs* can offer the following services (upper part of Fig 1): S1) hosting of datasets descriptions [1], i.e., metadata about these datasets (e.g., their URL, SPARQL endpoints and their availability), S2) services based on dataset's metadata, i.e., browsing, searching and analytics as well as testing if they are operational [2], and S3) services based on the actual contents (triples) of the datasets [3, 4], e.g., cross-dataset reasoning services for finding all the

SWODCH'23: *International Workshop on Semantic Web and Ontology Design for Cultural Heritage*, November 7, 2023, Athens, Greece

✉ mountant@ics.forth.gr (M. Mountantonakis); ioantheocharakis@gmail.com (I. Theocharakis); tzitzik@ics.forth.gr (Y. Tzitzikas)

ORCID 0000-0002-1951-0241 (M. Mountantonakis); 0000-0001-8847-2130 (Y. Tzitzikas)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

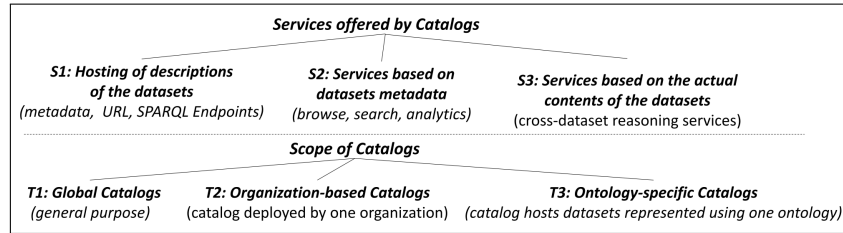


Figure 1: Services and scope of dataset catalogs

datasets of a URI. Concerning the scope of such catalogs (lower part of Fig 1), they can be divided into T1) Global, i.e., general purpose catalogs like lod-cloud.net [5], Datahub (<https://datahub.io/>) and Loupe [6], T2) Organization-based, e.g., a CKAN instance deployed by one organization (<https://ckan.org/>), a Zenodo channel, etc. and T3) Ontology-specific, where the hosted datasets are represented using one ontology (and its specializations). In this paper, we focus on T3, which is a special case of T1, since we restrict the datasets of the catalog by focusing only on datasets that use the ISO 21127 Standard CIDOC Conceptual Reference Model (CIDOC-CRM) [7]; an event-based ontology for the cultural domain that is used by dozens/hundreds of institutions and research projects for enabling semantic interoperability between cultural institutions [8].

The objective for T3 catalogs follows: i) the community that is interested in the focused ontology has incentive to publish their datasets in that catalog, and ii) it is more sustainable to achieve completeness for one ontology, than being complete for all the available ones. However, it is not trivial to create such a catalog; due to the numerous ways to publish a dataset, it is quite challenging even to discover all the datasets using a popular model, such as CIDOC-CRM. Indeed, the lack of even a simple list that includes information of all the CIDOC-CRM datasets (except for a table in a GitHub page [8] with 18 datasets), does not enable their discoverability and reusability, even from field experts. For tackling this limitation, we focus on how to i) discover all the CIDOC-CRM datasets, ii) compute ontology-based descriptions by using SPARQL queries, and iii) browse statistics and visualizations for the CIDOC-CRM datasets in an interactive way.

Concerning our contribution, we present an ontology-specific portal (or catalog), by focusing on CIDOC-CRM. The objective is to make it feasible the CIDOC-CRM users, dataset owners and experts, to discover the available CIDOC-CRM datasets and to browse statistics/visualizations about them, since it can be important for several use cases including data discovery, data integration, ontology evolution and others. Specifically, we first collect 30 CIDOC-CRM datasets and we compute ontology-based descriptions using VoID vocabulary [9]. Afterwards, we present an online portal (https://demos.isl.ics.forth.gr/CIDOC-CRM_Portal/) that offers: i) browsing of all the available CIDOC-CRM datasets by supporting ontology-based statistics and visualizations about each dataset, ii) searching for specific classes and properties, iii) the discovery of the most frequent (CIDOC-CRM) properties and classes, iv) measurements regarding the commonalities between pairs of datasets and v) a form for adding any new CIDOC-CRM dataset. Moreover, we offer an analysis for the 30 collected datasets, which reveals that there is power-law distribution concerning the usage of CIDOC-CRM properties and classes. To the best of our knowledge, this is the first work providing such an ontology-specific portal (i.e., for CIDOC-CRM model).

In the rest of this paper, §2 discusses the related work, §3 shows the use cases and §4 presents the steps for creating the ontology-based descriptions. §5 presents the functionality of the portal and §6 provides measurements over the collected datasets. Finally, §7 concludes the paper.

2. Related Work

We discuss dataset catalogs, CIDOC-CRM based services and a comparison with related work.

Dataset Catalogs. Concerning global catalogs, in Datahub and *lod-cloud.net* [5] publishers can upload a description of their datasets with some basic or enriched metadata, whereas Google Dataset Search [10] collects dataset metadata at web scale by using crawlers. Through such catalogs, the users can browse the datasets using keyword or/and faceted search mechanisms (S1 services). Moreover, there exists similar tools to the proposed portal, such as Aether [11], Loupe [6] and KartoGraphI [12], where VoID statistics for any RDF dataset are computed and ontology analytics are offered and visualized by using SPARQL queries (S2 services). Moreover, there are also global-scale approaches such as SPOTAL [1] and SPLENDID [13], which compute such statistics for aiding the selection of sources for federated SPARQL queries. Regarding catalogs offering S3 services, they analyze the contents of datasets (all their triples and entities), e.g., by collecting RDF data dumps and by constructing indexes. Such approaches include LODsyndesis [3], where the contents of 400 RDF datasets have been indexed (including 2 billion triples), LODVader [4] that offers analytics over 491 datasets, and LOD-a-LOT [14] where 28 billion RDF triples from thousands of documents have been collected. Their objective is to offer advanced data discovery mechanisms and content-based analytics over the LOD Cloud [2]. Finally, there exists popular organization-based catalogs such as Zenodo and CKAN, and domain specific catalogs, such as <https://bio2rdf.org/> and <http://linkedlifedata.com/> for the life science domain.

Services using CIDOC-CRM. There are many services that use the CIDOC-CRM model for various tasks, including Entity Recognition [15], Question Answering [16], Personalization and Recommendation [17] and others [8]. Concerning services that visualize CIDOC-CRM data, RDF visualizer [18] offers browsing mechanisms for CIDOC-CRM triples, whereas the CIDOC-CRM periodic table [19] is an interface for the documentation of the CIDOC-CRM model.

Placement and Novelty. The portal of this paper belongs to the scope T3, and mainly offers S2 services. Comparing to similar catalogs that compute VoID statistics [6, 11, 1, 12], the presented portal focuses on a single ontology and offers some more dedicated statistics and analytics for CIDOC-CRM. Concerning the CIDOC-CRM services, we focus on providing an interactive browsing system for visualizing statistics for the CIDOC-CRM model, and not for browsing all the triples [18] or for documentation purposes [19]. To the best of our knowledge, this is the first work providing an ontology-specific portal for a given ontology (i.e., CIDOC-CRM), and an analysis of the CIDOC-CRM model for multiple real datasets.

3. Use Cases of the Portal

We present the use cases where the portal can be exploited (and the corresponding users). First, the users are divided in 3 categories: a) simple (RDF) users, i.e., users familiar with Semantic Web technologies, b) CIDOC-CRM dataset owners, i.e., users that have published at least one dataset by using the CIDOC-CRM model, and c) CIDOC-CRM experts/researchers, i.e., the experts of the CIDOC-CRM community. Below, we provide 4 use cases (UC) and example user queries that we desire to support. The use cases (see Fig. 2) are the following: UC1) dataset discovery and selection, UC2) data publishing, UC3) data integration and UC4) ontology evaluation.

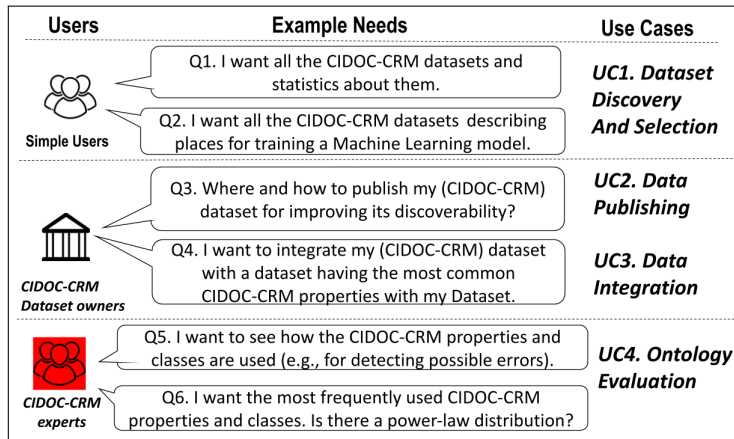


Figure 2: The use cases, example user queries and corresponding users

- **UC1. Dataset Discovery and Selection.** The objective is any user to be able to discover all the available CIDOC-CRM datasets (see Query Q1 in Fig. 2), i.e., each dataset is a kind of example of how to use CIDOC-CRM, thereby it can aid the adoption of the ontology. Moreover, one can find datasets having specific properties or/and classes (see Query Q2 in Fig. 2), such as datasets describing places, events, etc. In this way, the user can select the most appropriate dataset(s) for creating an application, such as for Question Answering [16], a recommendation system [17] or/and for training a Machine Learning model [8].

- **UC2. Data Publishing.** Since there are several ways to publish a dataset, and given the large number of published datasets, the key notion is a dataset to be easily discoverable and reusable by interested users. Therefore, by having a single portal including all the datasets of a specific model (such as CIDOC-CRM), we expect that it will be more discoverable from users that are interested in the given ontology or/and domain (i.e., in our case Cultural Heritage).

- **UC3. Data Integration.** In many cases, the data owners desire to integrate their data with existing datasets for enriching their information, i.e., for creating larger and more complete datasets. By offering services for all the datasets described through a specific model (e.g., CIDOC-CRM), the dataset owners will be able to discover the datasets having the most common properties and classes with their datasets for selecting them for semantic data integration [20].

- **UC4. Ontology Evaluation.** Since some ontologies are widely used, the experts of such models usually desire to evaluate which classes/properties are used and how, to detect problems more easily (e.g., using the ontology in a wrong way) and to think about possible extensions. We expect that queries like Q5 and Q6 of Fig. 2 will be quite useful for CIDOC-CRM experts, since there is a very active community through the CIDOC-CRM Special Interest Group (SIG), where many organizations and researchers participate (<https://www.cidoc-crm.org/sig-members-list>). This group is associated with several management activities: ontology versions, mappings, translations, compatible models, use cases, issues, best practices, meetings, and others.

4. Collecting CIDOC-CRM Datasets and Computing Statistics

Here, we present the steps (i.e., see Fig. 3) that are followed for collecting CIDOC-CRM datasets and for producing statistics using the VoID vocabulary [9].

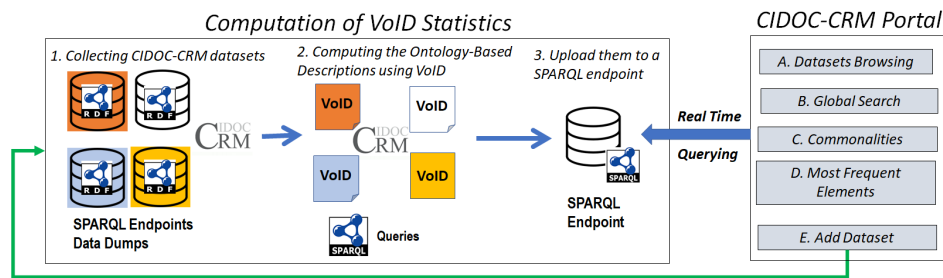


Figure 3: The steps of collecting and producing statistics of CIDOC-CRM datasets

Step 1. Collecting CIDOC-CRM datasets. We tried to collect all the available CIDOC-CRM datasets that offer either an online SPARQL endpoint or an RDF data dump. We used the list of 18 datasets provided in a GitHub page [8] and we further searched in google scholar and through catalogs like Zenodo and search engines, with the keywords “CIDOC-CRM dataset/endpoint/data dump”, for finding more datasets. At the time being, we managed to collect 30 real RDF datasets (having in total 560 million RDF triples), where 21 of them offer a public SPARQL endpoint and 9 of them only an RDF data dump (more statistics are presented in §6).

Step 2. Computing the Ontology-Based Descriptions using VoID. For the computation of the statistics we send queries to SPARQL endpoints, however there are datasets where a SPARQL endpoint is not provided. For these datasets, we downloaded the data dumps and we uploaded them to our SPARQL endpoint for performing the computations. The mentioned process was time consuming in some cases, due to i) the large size of some datasets and ii) syntax errors in some RDF files. Concerning the computation of statistics (Step 2 of Fig. 3), we use some basic SPARQL queries by exploiting the VoID vocabulary [9] and some extra SPARQL queries dedicated to CIDOC-CRM properties, classes and instances and at the end we produce a single file in “Turtle” format for each dataset including all the statistics. All the queries and produced files can be accessed in https://github.com/mountanton/CIDOC-CRM_Portal.

Example. Fig. 4 shows the file for the dataset ‘OpenArcheo’ [21], which is a semantic mediator for archaeological datasets. Indeed, lines 7-13 show the basic VoID statistics, and lines 14-21 show how we store the properties and classes (and the number of triples that they appear). Finally, lines 22-27 contain the dedicated CIDOC-CRM statistics, such as the number of unique CIDOC-CRM properties/classes and the number of triples (and their percentage) including a CIDOC-CRM property or instance (i.e., entities that are members of a CIDOC-CRM class).

Important Note. Concerning the CIDOC-CRM model, in this paper when we refer to “CIDOC-CRM properties and classes”, we refer to all the properties and classes of the RDF file of CIDOC-CRM version 7.1.2 (https://cidoc-crm.org/rdfs/7.1.2/CIDOC_CRM_v7.1.2.rdfs) which contains 309 CIDOC-CRM properties (including inverse properties) and 76 CIDOC-CRM classes, and not in properties and classes that extend the mentioned CIDOC-CRM properties and classes.

Step 3. Upload the Ontology-based Descriptions to a SPARQL Endpoint. The produced files (see Step 3 of Fig. 3) of all the datasets are uploaded in an online SPARQL Endpoint (i.e., 30 “Turtle” files). For describing all these (VoID) statistics for these datasets, 23,195 triples were created. The key notion is the endpoint to be used at real time from the portal for enabling a) the visualization of the already computed statistics, b) the computation of even more statistics through more SPARQL queries and c) the easy addition of any CIDOC-CRM dataset.

```

1@prefix void: <http://rdfs.org/ns/void#> .
2@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3@prefix dcterms: <http://purl.org/dc/terms/> .
4@prefix void-crm: <http://www.ics.forth.gr/isl/void-crm/>.
5@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/>.
6
7<http://openarchaeo.huma-num.fr/explorateur/home> rdf:type void:Dataset;
8  dcterms:title "Open Archaeo";
9  dcterms:description "A semantic mediator for archaeological datasets";
10 void:triples "1424168";
11 void:entities "266454";
12 void:properties "61";
13 void:classes "23";
14 void:propertyPartition [
15     void:property crm:P4_has_time-span;
16     void:triples "24344";
17 ]; ...
18 void:classPartition [
19     void:class crm:E53_Place;
20     void:triples "4368";
21 ]; ...
22 void-crm:propertiesCIDOC "30";
23 void-crm:classesCIDOC "14";
24 void-crm:triplesWithCIDOCproperty "652201";
25 void-crm:triplesWithCIDOCpropertyPercentage "45.80%";
26 void-crm:triplesWithCIDOCinstance "1195837";
27 void-crm:triplesWithCIDOCinstancePercentage "83.97%";

```

Figure 4: The produced Turtle file for the dataset OpenArchaeo [21]

5. The CIDOC-CRM Datasets Portal

Here, we provide some details about the architecture of the web portal and then we present the functionality of the portal and we explain how it corresponds to the users and use cases of §3.

5.1. The Architecture and the Code of the CIDOC-CRM Datasets Portal

The portal is available in https://demos.isl.ics.forth.gr/CIDOC-CRM_Portal/ and offers real time interactive browsing and visualizations. It runs on a server with 4 GB main memory, 8 cores and 60 GB disk space. Its architecture is shown in Fig. 5; the frontend has been designed by using the Angular framework (<https://angular.io/>) and the backend offers a REST API by using the Spring Boot Framework (<https://spring.io/>). When a user selects a mode, a request is sent to the REST API, which is connected to the public SPARQL endpoint that uses Virtuoso Openlink Software (<https://virtuoso.openlinksw.com/>) that contains the ontology-based descriptions. Afterwards the response is sent back to the frontend in JSON format, and it can be presented to the user through different types of visualizations: a) HTML tables, b) Bar and Rose charts by using the NGX-Echarts library (<https://xieziyu.github.io/ngx-echarts>) and c) Chord charts through the D3.js library (<https://d3js.org/d3-chord>).

Code and Queries. The code for all the components of the portal and all the SPARQL queries for all the modes are available in https://github.com/mountanton/CIDOC-CRM_Portal.

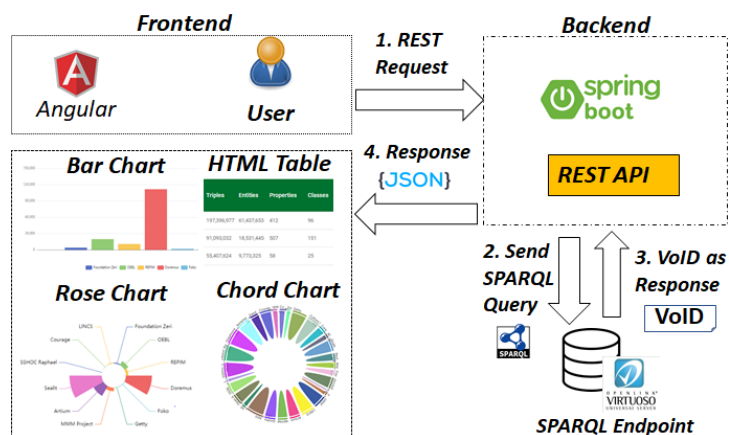


Figure 5: The architecture of the CIDOC-CRM Datasets Portal

5.2. The Modes of the CIDOC-CRM Datasets Portal

The webpage offers five interactive modes: i) Datasets Browsing, ii) Global Search, iii) Commonalities, iv) Most Frequent Elements, and v) Add Dataset. For each mode, Fig. 6 shows a screenshot with an example. Finally, a tutorial video can be accessed in https://youtu.be/ar8JEty94_w.

- **Mode A. Datasets Browsing.** This is the default mode and the user can browse statistics and visualizations for all the datasets (see the upper side of Fig. 6). In particular, one can browse ranking lists (using HTML tables) and visualizations through charts. By clicking on a single dataset, more information are shown for that dataset, including a description, a URL, its statistics and also the list of its properties and classes (including dedicated lists for CIDOC-CRM).

Use Cases. It mainly corresponds to the UC1 and can be useful for any user for discovering the most appropriate datasets for their needs. Secondly, it is connected to UC4, since the CIDOC-CRM experts can exploit all the statistics for evaluating how the CIDOC-CRM model is used, e.g., to check about distributions of CIDOC-CRM properties and classes.

- **Mode B. Global Search.** The user can search for any property/class, and the portal returns all the datasets containing the desired property/class and the number of triples that they appear. For aiding the user, we provide autocomplete services and a drop-down list including all the CIDOC-CRM properties and classes. For instance, Fig. 6 shows: i) the datasets describing places (class “crm:E53_Place”) and ii) the datasets including the property “crm:P52_has_current_owner”.

Use Cases. It corresponds to the UC1 and can be useful for discovering datasets containing entities of a desired class (e.g., places, births) or property (e.g., “took place at”, “carried out by”).

- **Mode C. Commonalities.** The user can discover all the common properties and classes between any pair of datasets, e.g., Fig. 6 shows the common CIDOC-CRM classes and properties between the datasets “Sealit” [22] and “WW1LOD” [23] (they have 8 classes and 11 properties in common). For aiding the user, we provide drop-down menus with the available datasets.

Use cases. It refers to the UC3, and can be useful for the dataset owners to discover which datasets have the most commonalities with their dataset, i.e., for creating an integration service.

- **Mode D. Most Frequent Elements.** Here, the objective is to find the most frequent properties and classes according to a) the number of datasets or b) the number of triples that they appear, e.g., see the example in Fig. 6 including the most popular CIDOC-CRM classes.

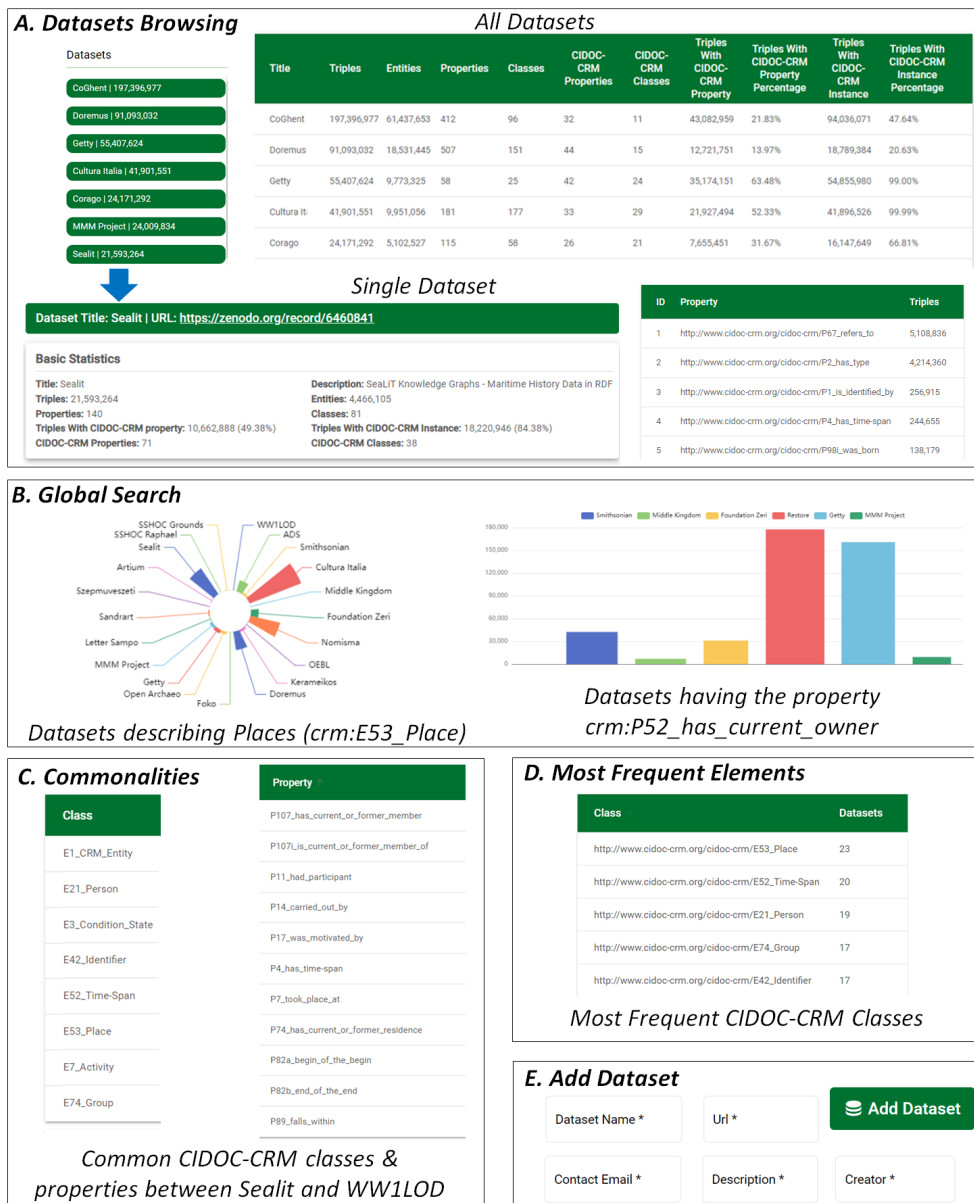


Figure 6: Screenshots of The CIDOC-CRM Datasets Portal

Use cases. It corresponds to the UC4, since it can be useful for the CIDOC-CRM experts for analyzing the distribution of the properties and classes.

• **Mode E. Add Dataset.** The objective is to enable the addition of new datasets for any dataset owner. Indeed, one can fill and submit a form (see the lower right part of Fig. 6) including some very basic details of the dataset. For avoiding spamming issues, the form is first evaluated by the administrators of the web portal and then the process of Fig. 3 is performed.

Use cases. It corresponds to the UC2 (Data Publishing), since any CIDOC-CRM dataset owner can fill the form for requesting to add their dataset to the portal.

Total Number of	Value
Collected (CIDOC-CRM) Datasets	30
Triples	560,452,817
Entities	129,931,741
Triples with a CIDOC-CRM property	168,158,485
Triples with a CIDOC-CRM instance	300,016,015

Table 1
Statistics about the CIDOC-CRM Datasets

Average Number of	Value
Properties per dataset	141.4
CIDOC-CRM properties per dataset	37.7
Classes per dataset	61.7
CIDOC-CRM classes per dataset	19.3

Table 2
Average Values for the CIDOC-CRM Datasets

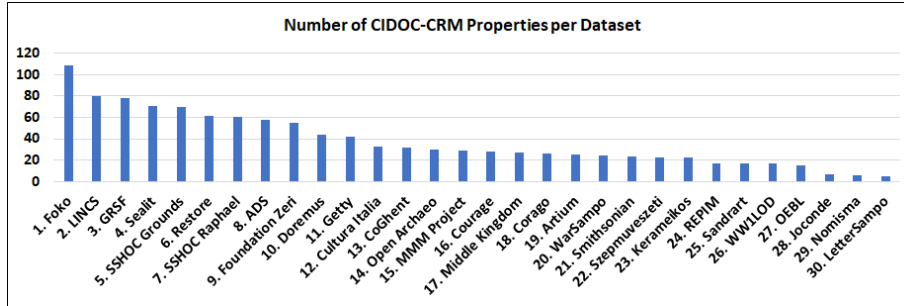


Figure 7: Number of CIDOC-CRM properties per dataset (in descending order)

6. Experimental Evaluation

Here, we provide indicative statistics and measurements concerning the 30 collected datasets, which can also be browsed and visualized in https://demos.isl.ics.forth.gr/CIDOC-CRM_Portal/.

6.1. Statistics for the Collected Datasets

First, Tables 1 and 2 provide some general statistics (total and average numbers) about the 30 collected datasets, based on the computed ontology-based descriptions. From the 560M triples of all the datasets, 168M triples contain a CIDOC-CRM property (approximately 30% of all triples) and 300M triples a CIDOC-CRM instance (approximately 53.5% of all triples).

Properties and Classes. Table 2 shows that each dataset uses on average 37.7 CIDOC-CRM properties and 19.3 CIDOC-CRM classes. Fig. 7 shows the exact number of CIDOC-CRM properties per dataset. Indicatively there are 7 datasets using ≥ 60 CIDOC-CRM properties, whereas only 3 datasets use ≤ 10 CIDOC-CRM properties. As regards the classes, Fig. 8 shows that 25 datasets use ≥ 10 CIDOC-CRM classes, while 12 datasets use ≥ 20 CIDOC-CRM classes.

Percentage of Triples (per dataset) using CIDOC-CRM properties and Instances. Fig. 9 shows for each dataset the percentage of triples containing a CIDOC-CRM property; indicatively 20 datasets use CIDOC-CRM properties in at least 30% of their triples. Concerning the instances, Fig. 10 depicts for each dataset the percentage of triples containing a CIDOC-CRM instance (i.e., an entity that is a member of a CIDOC-CRM class), and we can observe that half of the datasets (15 out of 30) include a CIDOC-CRM instance in at least 80% of their triples.

6.2. Frequency of CIDOC-CRM Properties and Classes

Here, we provide some indicative measurements about the CIDOC-CRM properties and classes.

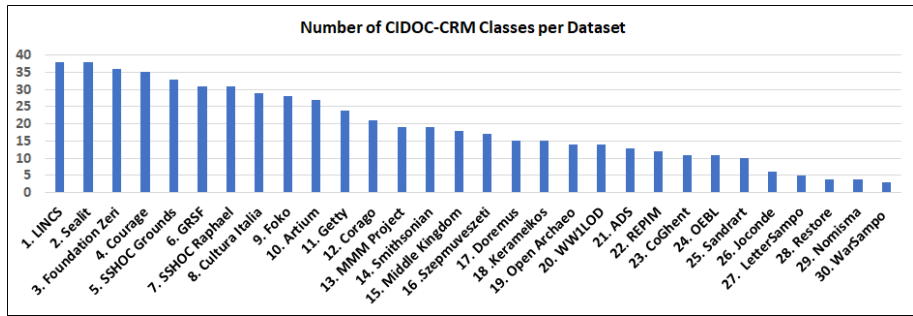


Figure 8: Number of CIDOC-CRM Classes per dataset (in descending order)

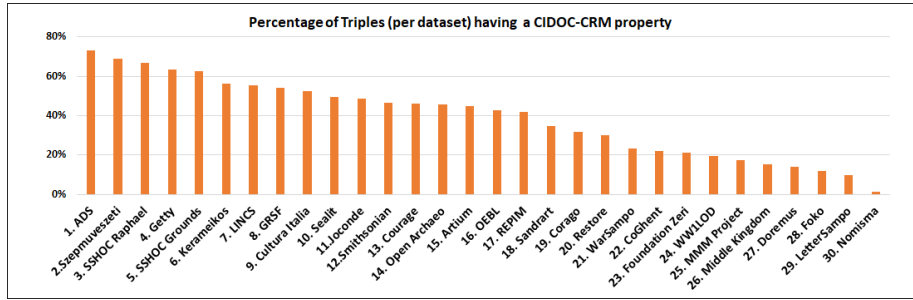


Figure 9: Percentage of triples per datasets having a CIDOC-CRM property (in descending order)

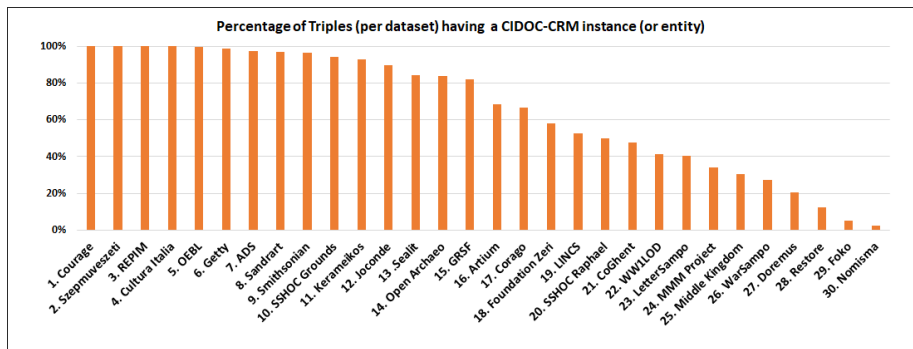


Figure 10: Percentage of triples per datasets having a CIDOC-CRM instance (in descending order)

Distribution measurements. The distribution of CIDOC-CRM properties and classes (according to the number of datasets that they appear) is shown in Figures 11 and 12, respectively. Concerning the properties, we observe a power-law distribution, i.e., some few CIDOC-CRM properties are used from many datasets, whereas most of them in a few datasets. Indicatively, only 29 CIDOC-CRM properties are used from ≥ 10 datasets, whereas 100 properties are used by two or a single dataset. Regarding the classes, most of them are also used by a low number of datasets, i.e., see Fig. 12. Finally, from the 309 properties and the 76 classes of the current CIDOC-CRM version, there are 100 properties (59 of them are inverse properties) and 12 classes that are not used by the collected datasets, i.e., the 32.3% of properties and the 15.7% of classes.

Most Popular CIDOC-CRM Properties and Classes. We show the most popular CIDOC-CRM properties and classes according to the number of a) datasets and b) triples, that they appear. Fig. 13 shows that the most popular properties are “crm:P14_carried_out_by” and

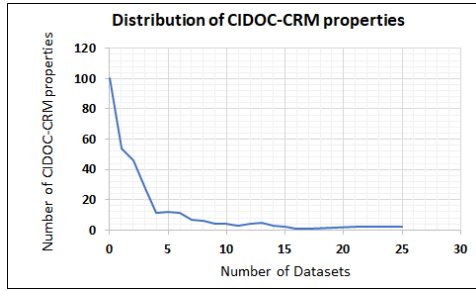


Figure 11: Distribution of CIDOC-CRM properties in the collected datasets

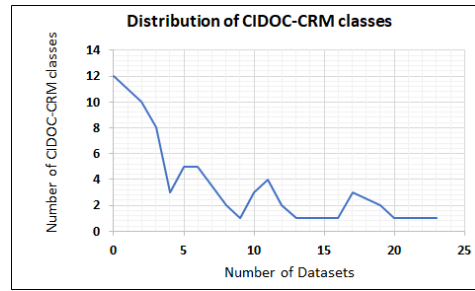


Figure 12: Distribution of CIDOC-CRM classes in the collected datasets

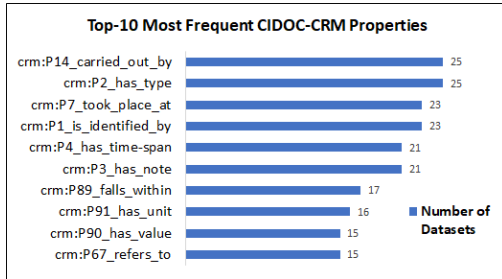


Figure 13: Top-10 most frequent CIDOC-CRM properties wrt the number of datasets

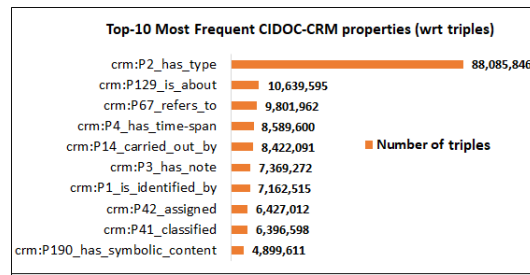


Figure 14: Top-10 most frequent CIDOC-CRM properties wrt the number of triples

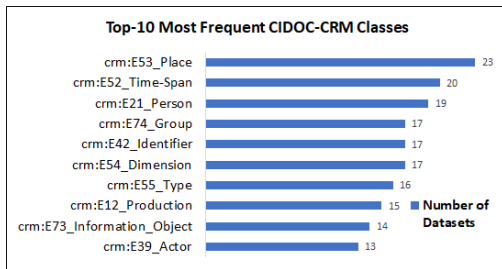


Figure 15: Top-10 most frequent CIDOC-CRM classes wrt the number of datasets

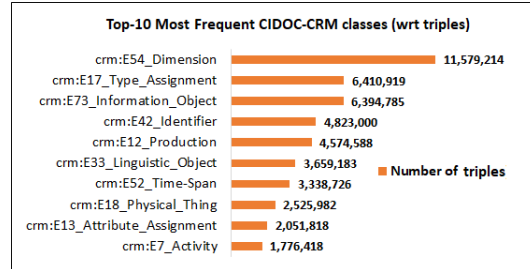


Figure 16: Top-10 most frequent CIDOC-CRM classes wrt the number of triples

“crm:P2_has_type” that appear in 25 datasets. Concerning the number of triples, i.e., see Fig. 14, again the property “crm:P2_has_type” is the top one, appearing in 88M triples. Regarding the CIDOC-CRM classes, the most frequent one is the “crm:E53_Place”, i.e., see Fig. 15, which appears in 23 datasets, whereas the class occurring in the highest number of triples (i.e., having the most instances) is crm:E54_Dimension with 11M triples (see Fig. 16).

7. Conclusion

We presented a portal that focuses on the ISO Standard CIDOC-CRM, for enabling the browsing and visualization of ontology-based descriptions of any CIDOC-CRM-based dataset. We described several use cases, all the details about how the statistics are computed, and the modes of the portal. We offered measurements about 30 real CIDOC-CRM datasets, which revealed a power-law distribution; some few CIDOC-CRM properties and classes are widely used, whereas

most of them are used by a few datasets. As a future work, we plan to a) compute/visualize more complex statistics (e.g., triple/path patterns since they can be exploited for Question Answering [16]), b) provide a more detailed analysis for the collected datasets through more measurements, and c) offer mechanisms for monitoring the changes in datasets and recomputing the statistics.

Acknowledgments. This work has received funding from the European Union’s Horizon 2020 coordination and support action 4CH (Grant agreement No 101004468).

References

- [1] A. Hasnain, Q. Mehmood, S. S. e Zainab, A. Hogan, SPORAL: profiling the content of public SPARQL endpoints, *International Journal on Semantic Web and Information Systems (IJSWIS)* 12 (2016) 134–163.
- [2] M.-E. Papadaki, Y. Tzitzikas, M. Mountantonakis, A brief survey of methods for analytics over RDF knowledge graphs, *Analytics* 2 (2023) 55–74.
- [3] M. Mountantonakis, Y. Tzitzikas, Content-based union and complement metrics for dataset search over RDF knowledge graphs, *Journal of Data and Information Quality (JDIQ)* 12 (2020) 1–31.
- [4] B. Neto, et al., Lodvader: An interface to LOD visualization, analytics and discovery in real-time, in: *Proceedings of the 25th International Conference Companion on World Wide Web, 2016*, pp. 163–166.
- [5] J. P. McCrae, A. Abele, P. Buitelaar, R. Cyganiak, A. Jentzsch, V. Andryushechkin, J. Debattista, The linked open data cloud, *Lod-cloud. net* (2019).
- [6] N. Mihindukulasooriya, M. Poveda-Villalón, R. García-Castro, A. Gómez-Pérez, Loupe - an online tool for inspecting datasets in the Linked Data Cloud., *ISWC (Posters & Demos)* (2015).
- [7] M. Doerr, The CIDOC CRM, an ontological approach to schema heterogeneity, in: *Dagstuhl Seminar Proceedings, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2005*.
- [8] Y. Tzitzikas, M. Mountantonakis, P. Fafalios, Y. Marketakis, CIDOC-CRM and machine learning: a survey and future research, *Heritage* 5 (2022) 1612–1636.
- [9] K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao, Describing linked datasets with the VoID vocabulary (2011).
- [10] D. Brickley, M. Burgess, N. Noy, Google dataset search: Building a search engine for datasets in an open web ecosystem, in: *The World Wide Web Conference, 2019*, pp. 1365–1375.
- [11] E. Mäkelä, Aether—generating and viewing extended VoID statistical descriptions of RDF datasets, in: *The Semantic Web: ESWC 2014 Satellite Events, Springer, 2014*, pp. 429–433.
- [12] P. Maillot, O. Corby, C. Faron, F. Gandon, F. Michel, Indegx: A model and a framework for indexing RDF knowledge graphs with SPARQL-based test suits, *Journal of Web Semantics* 76 (2023) 100775.
- [13] O. Görlitz, S. Staab, SPLENDID: SPARQL endpoint federation exploiting VOID descriptions., *COLD* 782 (2011).
- [14] J. D. Fernández, W. Beek, M. A. Martínez-Prieto, M. Arias, LOD-a-lot: A queryable dump of the LOD cloud, in: *ISWC, Springer, 2017*, pp. 75–83.
- [15] A. Felicetti, D. Williams, I. Galluccio, D. Tudhope, F. Niccolucci, NLP tools for knowledge extraction from italian archaeological free text, in: *2018 3rd digital heritage international congress, IEEE, 2018*, pp. 1–8.
- [16] 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 Hypertext, 2023*, pp. 1–10.
- [17] A. Dahroug, A. Vlachidis, A. Liapis, A. Bikakis, M. Lopez-Nores, O. Sacco, J. J. Pazos-Arias, Using dates as contextual information for personalised cultural heritage experiences, *JIS* 47 (2021) 82–100.
- [18] K. Petrakis, N. Minadakis, K. Doerr, M. Theodoridou, M. Doerr, RDF visualizer: A tool for displaying, browsing and exploring high density RDF data., in: *ISWC (Satellites), 2019*, pp. 197–200.
- [19] CIDOC-CRM periodic table, 2023. https://remogrillo.github.io/cidoc-crm_periodic_table.
- [20] M. Mountantonakis, Y. Tzitzikas, Large-scale semantic integration of linked data: A survey, *ACM Computing Surveys (CSUR)* 52 (2019) 1–40.
- [21] O. Marlet, T. Francart, B. Markhoff, X. Rodier, OpenArchaeo for usable semantic interoperability, in: *ODOCH 2019@ CAiSE 2019, 2019*.
- [22] P. Fafalios, A. Kritsotaki, M. Doerr, The SeaLiT ontology—an extension of CIDOC-CRM for the modeling and integration of maritime history information, *ACM Journal on Computing and Cultural Heritage* (2023).
- [23] E. Mäkelä, J. Törnroos, T. Lindquist, E. Hyvönen, WW1LOD: An application of CIDOC-CRM to World War 1 linked data, *International Journal on Digital Libraries* 18 (2017) 333–343.